Machine learning and domain adaptation for enhancing the measure of brain health with MEG and EEG signals

Apolline Mellot

université PARIS-SACLAY

Machine learning and domain adaptation for enhancing the measure of brain health with MEG and EEG signals

Apprentissage statistique et adaptation de domaine pour l'amélioration de la mesure de la santé cérébrale à partir de signaux MEG et EEG

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: Sciences du 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 et la co-direction de **Sylvain Chevallier**, Professeur

Thèse soutenue à Paris-Saclay, le 8 novembre 2024, par

Apolline MELLOT

Composition du jury

Membres du jury avec voix délibérative

Valérie EGO STENGEL Directrice de recherche, Institut des Neurosciences Paris-Saclay	Présidente
Mario CHAVEZ Directeur de recherche, CNRS, Institut du Cerveau (ICM), Sorbonne Université	Rapporteur & Examinateur
Michael W. TANGERMANN Professeur associé, Donders Centre for Cognition	Rapporteur & Examinateur
Mark WOOLRICH Professeur, Oxford University	Examinateur

THESE DE DOCTORAT
NNT : 2024UPASG068

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

Titre: Apprentissage statistique et adaptation de domaine pour l'amélioration de la mesure de la santé cérébrale à partir de signaux MEG et EEG

Mots clés: Neurosciences, Santé, Traitement du signal, Apprentissage statistique, Adaptation de domaine

Résumé:

Les études en neurosciences rencontrent des défis dans la collecte de grandes bases de données, limitant ainsi l'utilisation de l'apprentissage statistique. L'intégration de données publiques peut être une solution, mais les données recueillies dans différents contextes présentent souvent des différences systématiques, appelées décalages de données (dataset shifts). Ces décalages, causés par des variations dans les sites d'enregistrement, le dispositif d'enregistrement ou les protocoles expérimentaux, compliquent l'application des méthodes d'apprentissage, qui exigent généralement des données d'entraînement et de test similaires. Cette thèse examine ces décalages dans les données M/EEG pour en comprendre les causes, leurs effets sur les modèles d'apprentissage, et propose des solutions adaptées au type de décalage pour améliorer la généralisation des modèles prédictifs.

Dans un premier temps, nous avons analysé les décalages survenant dans les enregistrements M/EEG en lien avec l'activité cérébrale, l'anatomie ou la configuration des capteurs. Pour harmoniser la distribution des données, nous avons utilisé une approche riemannienne d'alignement des données et l'avons adaptée à la régression non supervisée. Pour évaluer l'efficacité de l'alignement, nous avons réalisé des expériences sur des données simulées et réelles. Nous avons montré que la performance des modèles d'apprentissage peut être affectée par ces décalages et qu'elle peut être améliorée en alignant les distributions de données.

Dans la deuxième partie, nous nous sommes concentrés sur les décalages survenant à la fois dans les données M/EEG et la distribution de la variable à prédire y. Cette situation est courante dans les études cliniques où les données sont recueillies dans différents sites et auprès de différentes populations. Dans ce contexte, l'alignement proposé précédemment n'est pas suffisant pour traiter les décalages. Nous avons proposé une nouvelle méthode pour aborder l'adaptation de domaine dans des situations où les domaines sources ont des distributions de y distinctes. Cette méthode exploite la structure géométrique de la variété riemannienne pour apprendre conjointement un opérateur de recentrage spécifique au domaine et le modèle de régression. Nous avons réalisé des comparaisons empiriques sur la généralisation inter-sites des modèles de prédiction de l'âge avec des données EEG provenant d'un grand ensemble de données multinationales. L'approche proposée a significativement amélioré la généralisation des modèles à travers les sites par rapport aux méthodes de référence.

Finalement, nous avons abordé le problème des différents dispositifs d'enregistrement EEG. Le nombre et les positions variables des capteurs rendent difficile la comparaison des données provenant de différents dispositifs, et rendent même impossible l'utilisation directe des méthodes d'apprentissage. Pour remédier à cela, nous avons proposé une approche non supervisée exploitant la physique des signaux EEG : nous avons interpolé les canaux EEG de diverses configurations sur des positions fixes avec l'interpolation basée sur la physique de la propagation électromagnétique. Une évaluation comparative avec d'autres méthodes a été effectuée sur six bases de données publiques pour la classification d'imagerie motrice main droite/gauche. L'interpolation s'est montrée similaire ou meilleure que les autres méthodes.

Les contributions de cette thèse visent à améliorer la généralisation des modèles d'apprentissage appliqués aux données M/EEG sous différents aspects et situations. Le but était de mieux comprendre les décalages de données M/EEG et de proposer des approches pour atténuer leurs effets dans des scénarios réalistes.

Title: Machine learning and domain adaptation for enhancing the measure of brain health with MEG and EEG signals

Keywords: Neurosciences, Health, Signal processing, Machine learning, Domain adaptation

Abstract:

Neuroscience studies face challenges in gathering large datasets, which limits the use of machine learning (ML) approaches. One possible solution is to incorporate additional data from large public datasets; however, data collected in different contexts often exhibit systematic differences called dataset shifts. Various factors, such as site, device type, or experimental protocol, can lead to substantial divergence of M/EEG signals that can hinder the success of ML across datasets. This variability can induce distribution shifts in the data and in the biomedical variables of interest. ML algorithms typically require similar feature distributions at train and test time. Thus, these shifts limit the application of supervised ML algorithms. This thesis investigates dataset shifts in M/EEG data to understand their causes, how they affect ML models, and proposes methods adapted to the context and shift type to improve the generalization of predictive models.

In the first part, we focused on dataset shifts occurring in M/EEG recordings, not considering shifts in the label distributions. We investigated how changes in brain activity, anatomy, or device configuration can lead to dataset shifts in M/EEG data. To harmonize the data distribution, we used a Riemannian data alignment approach and adapted it to an unsupervised regression context. To assess the effectiveness of the alignment methods, we performed a series of experiments on simulated and real data. We showed that the performance of ML models can be affected by dataset shifts and that it can be improved by aligning the data distributions.

In the second part, we focused on dataset shifts occurring jointly in M/EEG recordings and in the y distribution, the variable to predict. Such situations are common in clinical studies where data is collected from different sites and populations. In this context, the previously proposed alignment methods are not enough to handle these shifts. We proposed a novel method to address domain adaptation for situations in which source domains have distinct y distributions. This method exploits the geometric structure of the Riemannian manifold to jointly learn a domain-specific re-centering operator and the regression model. We performed empirical benchmarks on the cross-site generalization of age-prediction models with resting-state EEG data from a large multi-national dataset. The proposed approach significantly improved the generalization of the models across sites compared to state-of-the-art methods.

In the third part, we addressed the problem of different recording devices in EEG data. The varying number and positions of sensors make it difficult to compare data from different devices, and even make it impossible to directly train a ML model on data from one device and test it on data from another. To tackle this, we proposed an unsupervised approach leveraging EEG signal physics: we map EEG channels from various configurations to a template of fixed positions using field interpolation. Comparative analysis against other methods was conducted with leave-one-dataset-out validation on six public BCI datasets for a right-hand/left-hand motor imagery classification task. It demonstrated that field interpolation is similar or better than the other methods.

The contributions presented in this thesis aimed at improving the generalization of ML models across datasets in M/EEG data from different aspects and situations. The goal was to provide a better understanding of the dataset shifts in M/EEG data and to propose approaches to mitigate their effects in realistic scenarios.

Remerciements

Je veux tout d'abord adresser mes premiers mots de remerciements à mes directeurs de thèse, Alex, Denis, et Sylvain.

Alex, ta pédagogie et ton expertise ont été des piliers essentiels qui m'ont permis d'évoluer dans un environnement de recherche exigeant, où je craignais au départ de ne pas avoir ma place. Ton humour, toujours bien dosé, a apporté une précieuse légèreté à nos échanges, en particulier lors des moments de "mode commando, le couteau entre les dents". Je te remercie pour ta rigueur, mais aussi pour cette capacité à rendre le travail intense plus agréable.

Denis, ta bienveillance, ton écoute et tes encouragements constants m'ont toujours fait sentir soutenue. Ton enthousiasme communicatif pour la recherche a été une source d'inspiration et m'a aidée à maintenir ma motivation tout au long de ces trois années. Ton accompagnement chaleureux a véritablement fait une différence dans mon parcours.

Sylvain, je tiens à te remercier d'avoir accepté de rejoindre mon équipe de supervision en plein milieu de ma thèse. Ta disponibilité, ta bienveillance et ton ouverture d'esprit ont enrichi nos échanges et m'ont permis d'avancer avec confiance. Ton implication a été précieuse et j'ai beaucoup apprécié nos discussions toujours constructives.

Je souhaite également adresser mes remerciements à Antoine, avec qui j'ai eu le plaisir de collaborer pendant les deux dernières années de ma thèse. Antoine, ton rôle a été celui d'un véritable encadrant junior pour moi : toujours disponible pour répondre à mes questions, échanger des idées, et partager ton expertise. J'ai énormément apprécié ta rigueur scientifique, ta patience et ta capacité à rendre les discussions enrichissantes et stimulantes. Travailler à tes côtés a été non seulement formateur, mais aussi très agréable.

Je tiens également à exprimer ma gratitude aux membres du jury pour leur participation à ma soutenance et leurs précieuses remarques. Vos questions et observations ont conduit à des discussions enrichissantes, qui resteront pour moi des moments marquants de cette étape finale. Je remercie tout particulièrement les rapporteurs pour le temps qu'ils ont consacré à la lecture de mon manuscrit.

Vos retours détaillés et constructifs m'ont permis d'améliorer mon travail et de le présenter sous sa meilleure forme.

Je souhaite remercier chaleureusement mes collègues, avec qui j'ai eu le plaisir de partager ces trois années de thèse. J'ai toujours apprécié l'ambiance ouverte et inclusive qui régnait au sein de ces équipes, où il était possible de discuter de tout avec esprit d'ouverture.

Merci tout particulièrement à Théo et Thomas, pour être venus, comme moi, dans les bureaux désertés de l'Inria en pleine période de COVID. Votre présence et notre soutien mutuel au début de nos stages ont été précieux. Cédric et Benoît, mes acolytes de l'open space, merci pour votre bonne humeur, notre entraide et pour avoir participé au challenge Biomag avec moi. Maria, ta spontanéité rafraîchissante et nos nombreuses discussions, qu'elles soient sur nos projets ou nos vies personnelles, ont été une véritable bouffée d'air. Gabriela, merci cet été d'avoir créé et partagé un bureau de qualité avec moi, même si cela n'a pas duré très longtemps...

Je tiens également à remercier Alexandre Blain, Julia, Virginie, Florent, Nicolas, Pierre-Louis, Alexandre Perez, Léo, Félix, Jovan, Ambroise, Jade, Lilian, Alexis (et sans doute bien d'autres encore) pour avoir contribué à créer une atmosphère joyeuse et conviviale dans notre open space. Un grand merci aussi à Katia, Julie, Fernanda, Matthieu, Mathieu, Houssam, Pierre-Antoine, Alexandre Le Bris, Louis, Samuel, Raphaël, Jun, Riccardo, Marie, Sébastien, Célestin, ainsi qu'à tous les stagiaires, doctorants, post-doctorants, ingénieurs avec qui j'ai eu l'opportunité de collaborer, d'échanger, et de partager de bons moments, qu'ils soient de travail ou de détente.

Enfin, merci aux PIs des équipes MIND et SODA, pour avoir façonné un environnement de travail si agréable et envié par beaucoup. Votre vision et votre bienveillance ont permis de faire de ces équipes des lieux où il fait bon évoluer.

À mes amis, vous êtes un pilier essentiel de ma vie. Les moments que nous passons ensemble me sont très chers. J'ai littéralement partagé une partie de ma vie avec la plupart d'entre vous, nos colocations et tous ce que nous avons vécu ensemble sont des souvenirs que je chérirai toujours. Merci à mes premières colocs, Éline et Éloïse, qui ont rendu les années de prépa plus douces. Merci à Daniela, Lisa, Ana Maria, Justine et Clémentine de vous être joint à moi à Kley et d'avoir contribué à rendre les années SupOptique inoubliables et les plus festives de ma vie. Merci à Clément, Yann, Hector et Jérôme d'avoir partagé avec moi notre appartement du Moulon. Nos soirées, nos séances jeux de société ou jeux vidéo, nos moments à regarder des films, à faire la cuisine ensemble ou à planter des radis, et toutes les autres activités partagées avec vous au quotidien me manquent. Merci également à tous mes autres amis du lycée, de la prépa et de SupOptique, j'espère encore partager de nombreux moments avec vous.

Maman, Papa, Jules, vous m'avez vue grandir, vous m'avez soutenue, et aidée à devenir la personne que je suis aujourd'hui. C'est grâce à vous que j'ai pu entreprendre ce parcours, et je vous en suis profondément reconnaissante.

Enfin, merci Fred d'être là pour moi chaque jour, de m'épauler, de me faire rire, et d'apporter tant de légèreté et de bonheur à ma vie.

Contents

List of Figures

List of Tables

Notations

$\triangleq$	Definition of the left part of the equation as the right part
x	Scalar (lowercase character)
$\boldsymbol{x}$	Vector (bold lowercase character)
$\boldsymbol{X}$	Matrix (bold uppercase character)
$\boldsymbol{\Sigma}$	Spatial covariance matrix
$\boldsymbol{z}$	Features vector
$\boldsymbol{Z}$	Features matrix
$\boldsymbol{R}$	Autocovariance matrix
$\boldsymbol{S}$	Cross-spectral density matrix
$\boldsymbol{I}_P$	Identity matrix of size P
$\text{sign}(x)$	Sign function of x
$\text{diag}(\boldsymbol{x})$	Diagonal matrix with the elements of $\boldsymbol{x}$
$[\cdot]^\top$	Transpose operator of a vector or a matrix
$\mathbf{1}_N \triangleq [1, \ldots, 1]^\top$	N-dimensional vector of ones
$\left\|\cdot\right\|_F$	Frobenius norm
$\left\|\cdot\right\|_2$	2-norm
$\text{tr}(\boldsymbol{X})$	Trace of the matrix $\boldsymbol{X}$
$\text{diag}(\boldsymbol{X})$	Diagonal of the matrix $\boldsymbol{X}$
$\log(\boldsymbol{X})$	Matrix logarithm of the matrix $\boldsymbol{X}$
$\boldsymbol{X}^\alpha$	Power of the matrix $\boldsymbol{X}$
uvec	Vectorization operator of a matrix
δ_R	Riemannian distance associated with the affine invariant metric
$\mathbb{R}$	Set of real numbers
$\mathbb{R}^P$	Set of real vectors of size P
$\mathbb{S}_P$	Set of $P \times P$ symmetric matrices
$\mathbb{S}_P^{++}$	Set of $P \times P$ symmetric positive definite matrices
$T_{\boldsymbol{\Sigma}}\mathbb{S}_P^{++}$	Tangent space at $\boldsymbol{\Sigma} \in \mathbb{S}_P^{++}$
(N)	Group of $N \times N$ orthogonal matrices
$\mathbb{E}_t[\boldsymbol{x}]$	Expectation of the random variable $\boldsymbol{x}$ *w.r.t.* t
$\mathcal{N}(\mu, \sigma^2)$	Normal distribution with mean μ and variance σ
N	Number of data points in a dataset
P	Dimension of the covariance matrices

$\mathcal{S}$	Source dataset
$\mathcal{T}$	Target dataset
K	Number of source domains
$\mathcal{L}$	Loss function
$\nabla\mathcal{L}$	Gradient of the loss function
$\mathcal{L}'$	Derivative of the loss function

1 General background

Contents

Accurately assessing and measuring brain health is a key goal in neuroscience, especially in clinical contexts where early diagnosis and monitoring of neurological conditions can significantly improve patient outcomes. Techniques such as magnetoencephalography (MEG) and electroencephalography (EEG) provide non-invasive ways to capture the brain's electrical activity, making them invaluable tools for both research and clinical applications. However, despite the widespread use of M/EEG in brain health assessment, leveraging these signals for predictive modeling remains challenging due to their inherent variability. This variability arises from the complex nature of M/EEG signals and creates systematic differences, or shifts, between datasets. These shifts can significantly impair the performance of machine learning models, limiting their ability to generalize across different datasets and clinical settings. As a result, the development of robust methods capable of addressing these shifts is needed for the successful application of machine learning to M/EEG data in real-world clinical environments. In this thesis, we aim to investigate the problem of dataset shifts in M/EEG data and propose solutions to mitigate their effects.

This first chapter aims to gather all the necessary background and concepts important for to the rest of the manuscript. Even though the range of topics covered in this chapter is broad, they are all essential to understand the context of the work presented in the following chapters.

- In Section 1.1, we introduce how the brain is structured, how it operates, and how we capture its activity with M/EEG to collect a measurable signal. We briefly describe the biological basis of M/EEG signals and the recording techniques used to capture them. We conclude with the motivations and applications of M/EEG in clinical and research settings.
- In Section 1.2, we summarize how statistical learning has been applied to analyze M/EEG signals. We present the different types of learning algorithms used and the challenges associated with these methods. We also discuss the importance of preprocessing steps in the analysis of M/EEG signals and the different techniques used to clean the data.
- In Section 1.3, we focus on the theoretical framework behind the covariance-based approaches that will be the focus of the following chapters.
- Finally, in Section 1.4, we present the notion of domain adaptation and the motivations behind the use of these techniques in the context of biomedical data analysis. We give examples of existing types of domain adaptation related to the methods investigated in this thesis.

1.1 Capturing brain activity with M/EEG

Understanding the complexities of brain function requires sophisticated techniques to capture its dynamic activity. This section explores the origins and mechanisms of brain activity and introduce how magnetoencephalography (MEG) and electroencephalography (EEG) effectively capture these signals. These recording techniques revealed to be essential tools in the study of brain function and have a wide range of applications in clinical and research settings.

1.1.1 Origin and mechanisms of brain activity

In the following, we introduce the fondamental concepts of the brain's composition and functionning. We will first describe the structure of the brain and the neurons that compose it. Then, we will explain how neurons communicate with each other.

Structure of the brain

The brain is the central organ of our nervous system [Bear et al., 2020]. It is responsible for the control of most of the body's functions, from basic physiological processes such as breathing and heart rate to higher cognitive functions such as perception, memory, and decision-making [Kandel et al., 2000]. It is composed of two primary types of tissue: grey matter and white matter (Figure 1.1) [Bear et al., 2020]. Grey matter, which makes up the outer layers known as the cerebral cortex as well as deep brain structures like the basal ganglia and thalamus, consists mainly of neuronal cell bodies, dendrites, and unmyelinated axons. This tissue is crucial for processing information, as it contains the synapses where neuronal communication occurs (Figure 1.2). White matter, found beneath the grey matter and making up structures such as the corpus callosum and the internal capsule, consists predominantly of myelinated axons. These myelinated fibers facilitate the rapid transmission of electrical signals between different brain regions, allowing for efficient communication and integration of information.

Structure of neurons

The human brain is made up of billions of nerve cells called neurons that interact with each other through electrical and chemical signals [Bear et al., 2020, Hari and Puce, 2023]. Neurons are the basic units of the brain and are responsible for receiving, processing and transmitting information. They are highly specialized cells that can generate electrical impulses, known as action potentials, in response to stimuli [Purves et al., 2019]. These action potentials are the primary means by which neurons communicate with each other and are essential for the functioning of the nervous system. Each neuron consists of three main parts: the cell body (soma), dendrites, and an axon as shown in Figure 1.2. Dendrites are tree-like

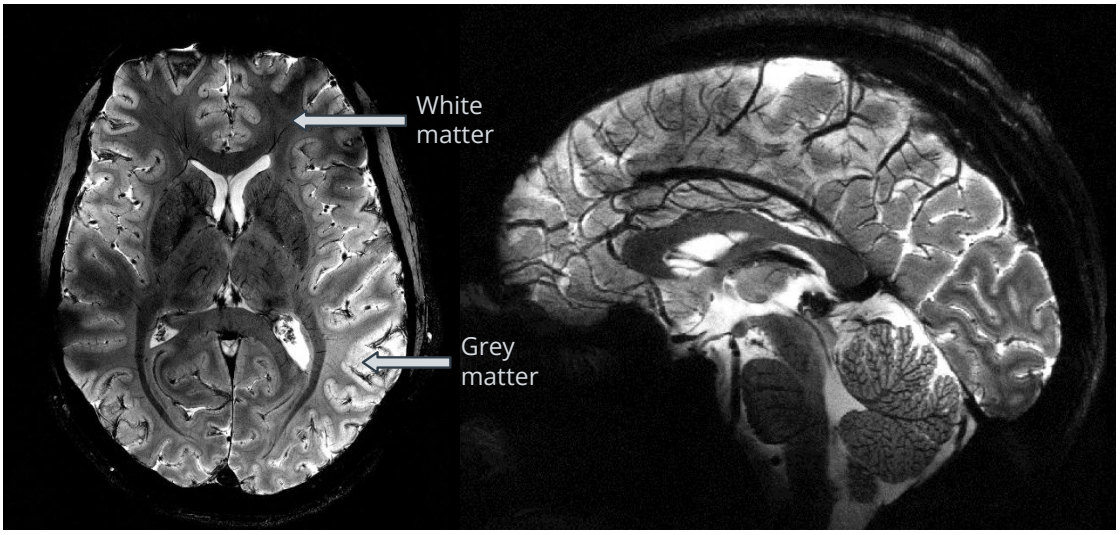

Figure 1.1: **First MRI images at 11.7T of the human brain [Boulant et al., 2024].** On the left, axial slice of the brain. On the right, sagittal slice of the brain. The images were acquired in the context of the Iseult project at the NeuroSpin center in Saclay, France. Iseult is a 11.7T MRI scanner dedicated to human brain imaging and is the strongest magnet with clearance for human participants.

extensions that receive signals from other neurons and convey this information to the soma. The soma contains the nucleus and is the metabolic center of the cell, which process the signal coming from the dendrites and can generate action potential depending on the received signal. The axon, a long, slender projection, transmits action potentials away from the soma to other neurons, muscles, or glands. At the end of the axon are terminal buttons that release neurotransmitters, chemical messengers that bridge the synaptic gap between neurons as represented in Figure 1.2.

Neuronal communication

The brain is a highly interconnected network of neurons that communicate with each other through synapses. This process of synaptic transmission is the basis of neural communication and underlies all brain activity [Bear et al., 2020, Purves et al., 2019]. The brain's electrical activity is generated by a flow of ions across the neuronal membrane. When a neuron generates an action potential, it travels down the axon to the synapse. Once the action potential reaches the terminal buttons at the end of the axon, the neuron releases neurotransmitters into the synaptic gap. These neurotransmitters bind to receptors on the dendrites of the next neuron, and affect the permeability of the receiving neuron's membrane to ions. Neurons maintain a resting membrane potential, a difference in electrical charge between

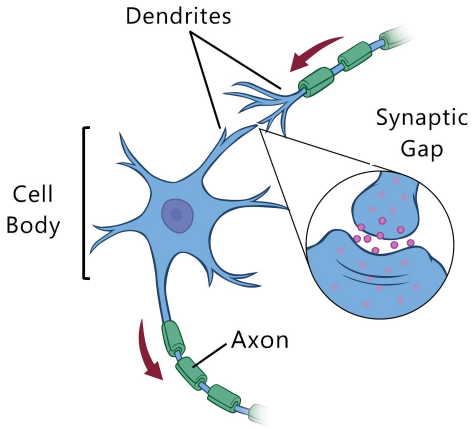

Figure 1.2: **Neuron and synapse structure.** Image adapted from ""Nervous System" by The Partnership in Education.

the inside and outside of the cell, by actively pumping ions in and out of the cell. When a neuron receives a signal, ion channels in the membrane open, allowing ions to flow into or out of the cell. This generates a change in the membrane potential, known as a post-synaptic potential, that can either depolarize (excite) or hyperpolarize (inhibit) the neuron. The electrical signals from the dendrites of a neuron are integrated by the soma and, if the sum of the post-synaptic potentials is strong enough, an action potential is generated and transmitted along the axon to other neurons. When generating an action potential, the neuron is said to fire.

1.1.2 Recording brain activity with M/EEG

Using the above knowledge of the brain's structure and function, we can now explore how MEG and EEG, or M/EEG for short, are used to capture the brain's electrical activity.

The neurophysiological basis of M/EEG signals

As mentionned previously, neurons communicate with each other through chemical signals. Both action potentials and post-synaptic potentials cause ion flows and thus produce electrical currents. These electrical currents create small electromagnetic field around the neuron. The electromagnetic field generated by a single neuron is so weak that it cannot be detected outside the brain. For measurable signals to be captured, a large number of neurons must produce electrical currents simultaneously. The temporal resolution of the post-synaptic potential ($\sim$ 10ms) is 10 times greater than the temporal resolution of the action potential ($\sim$ 1 ms) [Hämäläinen et al., 1993]. This makes post-synaptic potentials easier to synchronize and generate measurable electromagnetic fields. Additionally, the

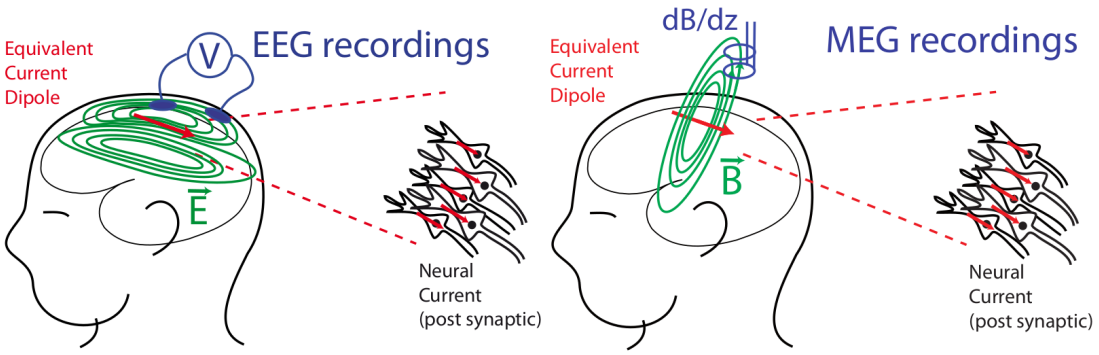

Figure 1.3: **Electric and magnetic field produced by a current dipole in the brain.** Adapted from [Gramfort, 2009]

electrical currents produced by these neurons must be oriented in the same direction to create a detectable field. In the cortex, pyramidal cells are the most common type of neuron and are organized in columns. When these neurons create post-synaptic potentials, they generate electrical currents that are oriented in the same direction, and perpendicular to the cortical surface. A group of pyramidal cells producing post-synaptic currents together acts as a current dipole in the cortex (see Figure 1.3). The synchronized activity of around 40,000 pyramidal cells creates an electromagnetic field strong enough to be detected on the scalp by techniques such as electroencephalography and magnetoencephalography [Hämäläinen et al., 1993, Buzsáki and Draguhn, 2004]. The signal measured by M/EEG is mostly emitted by groups of pyramidal cell in the cortex, due to their unique topology and placement relative to the scalp.

M/EEG recording techniques

The electrical activity of the brain can be recorded non-invasively using electroencephalography or magnetoencephalography. Both MEG and EEG provide high temporal resolution, on the order of milliseconds, with sampling rates between 250 and 2000 Hz. The recording setups for MEG and EEG are illustrated in Figure 1.4.

EEG is a technique that record the electrical activity of the brain that was first introduced by Hans Berger in 1924 during a surgical intervention [Berger, 1929]. Small electrodes are attached to the scalp using conductive gel. These electrodes detect the differences in electric potential generated by neurons in the brain and record them as voltage changes over time. The positions of the electrodes on the scalp are usually defined according to international standards, such as the 10-20 system, to ensure consistency across studies. Usually, a few dozen of sensors are used but some devices can have up to 256 electrodes. The signal recorded by EEG is typically of 50 to 100μV amplitude. Since the electrical signals

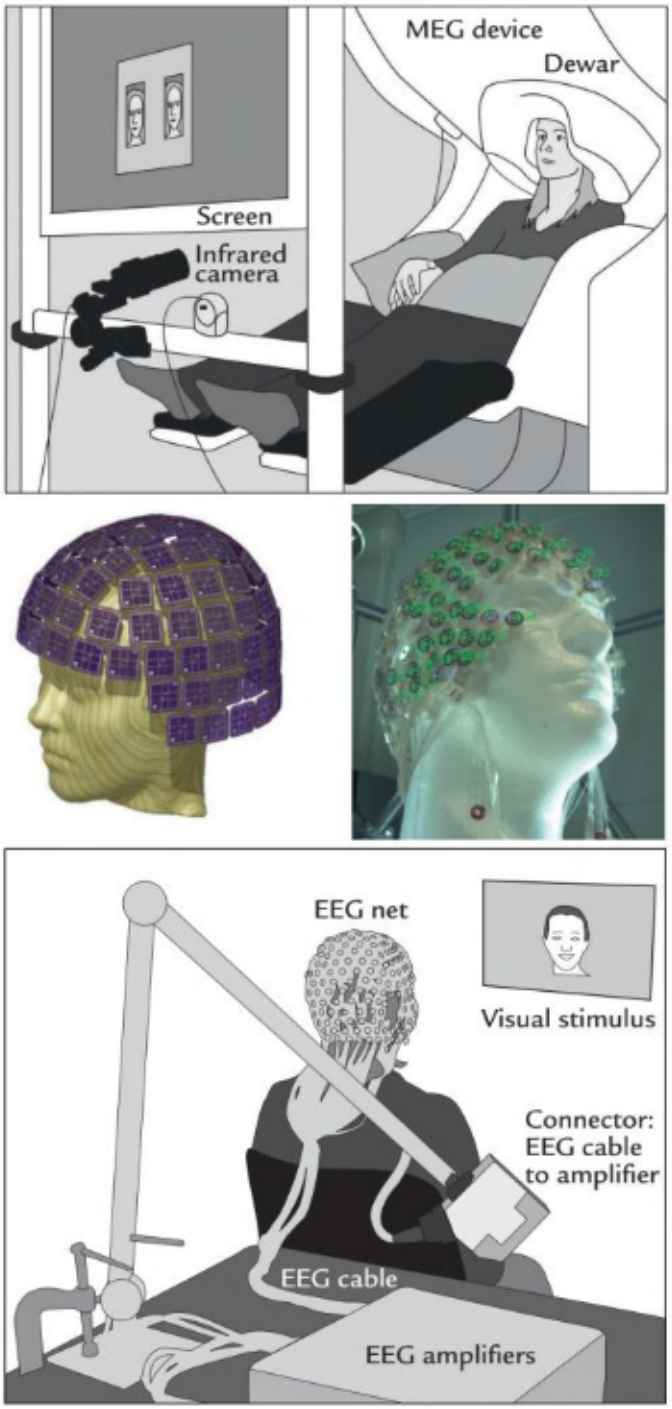

Figure 1.4: **M/EEG recording setups.** Top panel: schematic MEG setup with a screen for visual stimulation. Bottom panel: schematic EEG setup with a cap and electrodes, and a screen for visual stimulus presentation. Middle panel: MEG (left) and EEG (right) sensor arrays. Source: [Hari and Puce, 2023].

etected by the electrodes are weak, they are amplified to make signals that can be digitized and analyzed. The equipment necessary for EEG recording is generally less expensive to purchase and maintain than other brain imaging techniques. EEG is also more portable and flexible, allowing for a wider range of experimental setups and environments, including bedside monitoring in clinical settings.

MEG measures the magnetic fields generated by the electrical currents of the brain, first demonstrated by David Cohen in 1968 [Cohen, 1968]. These magnetic fields are recorded using an array of sensors placed at a few centimeters around the head. The typical number of sensors used in MEG is around 300. There are two types of MEG sensors: the magnetometers and the gradiometers. The magnetometers measure the strength and direction of magnetic fields, in Tesla, using one coil. The gradiometers are made up of two coils and are specifically designed to measure the gradient of the magnetic field, in Telsa per distance unit. Gradiometers are either positionned in a radial or tangential manner to record the gradient in different directions. Because the magnetic fields generated by the brain are weak, in the order of 10^{-13} T, MEG requires a specialized shielded room to reduce interference from the Earth's magneticfield and other sources of noise. The superconducting sensors and the magnetically shielded rooms make MEG a very expensive brain imaging device because of the liquid helium requirements. Yet, magnetic fields are not distorted by the tissues of the head than electric fields. Thus, MEG has a better spatial resolution than EEG.

The resulting recorded M/EEG signals provide information about the brain's electrical activity and can be used to study brain function, diagnose neurological disorders, and monitor the effects of treatments. An example of M/EEG signals from the MNE sample dataset [Gramfort et al., 2014] is displayed in Figure 1.5.

1.1.3 Motivations and applications

M/EEG techniques have the advantage to provide a direct measure of the neuronal activity. Thanks to their high temporal resolution, these techniques have been very useful for studying the temporal dynamics of the brain in clinical and research setting, leading to a wide range of applications.

Clinical applications of M/EEG

In clinical settings, M/EEG are primarily used to diagnose, monitor, and study the effect of treatments on brain disorders. MEG's ability to precisely localize the source of abnormal activity makes it a valuable tool for pre-surgical evaluation, guiding surgeons in planning the resection of epileptic foci. This feature of MEG also allows to map normally functioning brain areas before surgery to remove a tumor or treat epilepsy [Stufflebeam et al., 2009, Bagić, 2016]. Despite its clinical potential, MEG is not widely used in clinical settings due to the high cost of

the equipment. Consequently, MEG is predominantly used for research purposes. One of their main clinical applications of EEG is the diagnosis and monitoring of epilepsy. It is used to detect abnormal brain activity, such as spikes or seizures, and to localize the source of epileptic activity [Chowdhury et al., 2015]. EEG is also extensively used in sleep studies to diagnose sleep disorders such as sleep apnea, narcolepsy, and insomnia. By recording brain activity over several hours of sleep, EEG provides insights into the different stages of sleep and helps identify disruptions that may indicate a disorder [Moldofsky, 2001, Behzad and Behzad, 2021]. Furthermore, EEG is utilized to assess brain function in patients with brain injuries, stroke, or coma [Rohaut et al., 2024]. It provides information about the extent and location of brain damage, which is essential for prognosis and rehabilitation planning. During surgery, EEG monitoring is used to track brain activity and ensure the patient's state of anesthesia or unconsciousness, helping to prevent awareness during surgery and optimize anesthesia administration [Gibbs et al., 1937, Schneider et al., 2003].

M/EEG techniques are often combined with other neuroimaging methods, such as functional magnetic resonance imaging (fMRI), positron emission tomography (PET), or transcranial magnetic stimulation (TMS), to provide complementary spatial information about the brain. These multimodal approaches offer a more comprehensive understanding of brain function by integrating data from different perspectives.

M/EEG for research and development

In research settings, M/EEG techniques are employed to study the dynamics and connectivity of brain activity [Gross, 2019, da Silva Lopes, 2013]. In cognitive neuroscience, these methods are used during various cognitive tasks, including attention, memory, language, perception, and decision-making. Studying M/EEG oscillations along with functional and effective connectivity, helps elucidate the neural mechanisms underlying cognitive processes and how they are affected by different conditions or interventions.

Brain-computer interfaces (BCIs) represent another significant research application of M/EEG. These interfaces allow individuals to control prosthetic limbs, computer cursors, or other assistive technologies with their brain activity. The idea is to interpret the user's intentions and translate them into actionable commands. The signals recorded with M/EEG are interpreted to enable direct communication link between the brain activity and external devices. M/EEG-based BCIs typically involve the detection of specific brain signals, such as event-related potentials (ERPs), which are then processed and decoded using machine learning algorithms to infer the user's desired action [Abiri et al., 2019, Lotte et al., 2018]. The high temporal resolution of M/EEG allows for a nearly real-time feedback, allowing to develop closed-loop BCI systems.

Biomarker development is another area of research where M/EEG is widely used. Biomarkers are biological indicators that can be objectively measured and

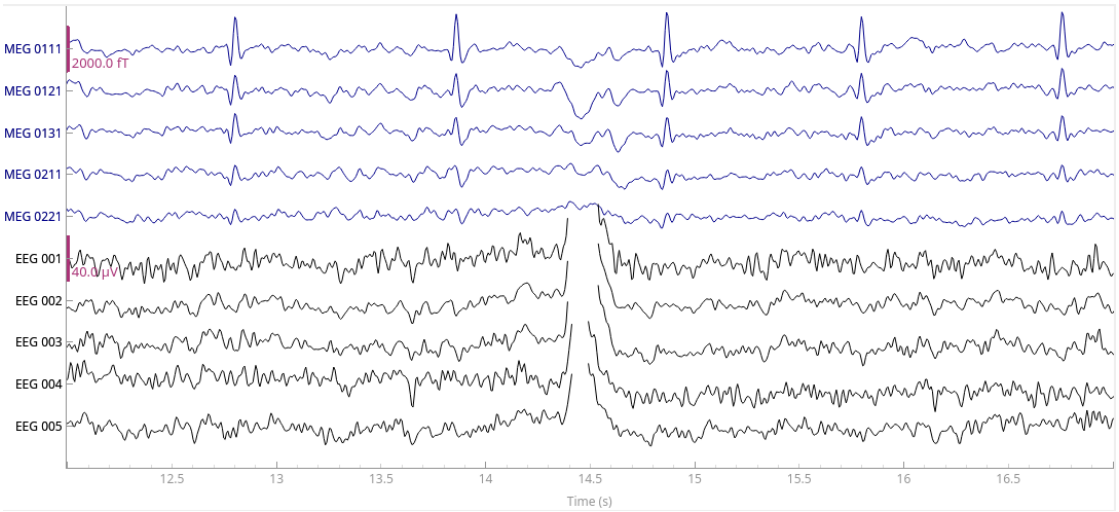

Figure 1.5: **M/EEG signals from the MNE sample dataset [Gramfort et al., 2014].** Raw signals from 5 EEG sensors and 5 MEG sensors during 5 seconds are shown. The EEG signals are in microvolts and the MEG signals are in femtoteslas. The signals are filtered between 0.1 and 40 Hz.

evaluated as indicators of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention. In the context of brain disorders, M/EEG biomarkers can help for identifying early signs of neurological diseases such as Alzheimer's disease [Horvath et al., 2018], attention-deficit and hyperactivity disorder (ADHD) [Lenartowicz and Loo, 2014] or depression [de Aguiar Neto and Rosa, 2019]. Furthermore, the International Conference on Biomagnetism (BIOMAG) has organized a data analysis competition in 2022 to develop biomarkers for the detection of dementia and mild cognitive impairment using MEG data. The solution my team proposed for this competition was based on the Riemannian approach presented in Section 1.2.2 was selected as the winning solution.

EEG is also increasingly employed in drug development, particularly to assess the efficacy and safety of new pharmacological compounds [Itil and Itil, 1995]. It plays an important role in pharmacokinetic-pharmacodynamic (PK-PD) modeling, which explores the relationship between the drug's concentration in the body and its physiological effects on brain activity [Groenendaal et al., 2008, Ebert et al., 2001].

1.2 Statistical learning with M/EEG

In this section, we provide an overview of the statistical learning methods used in M/EEG data analysis. We discuss the importance of pre-processing in

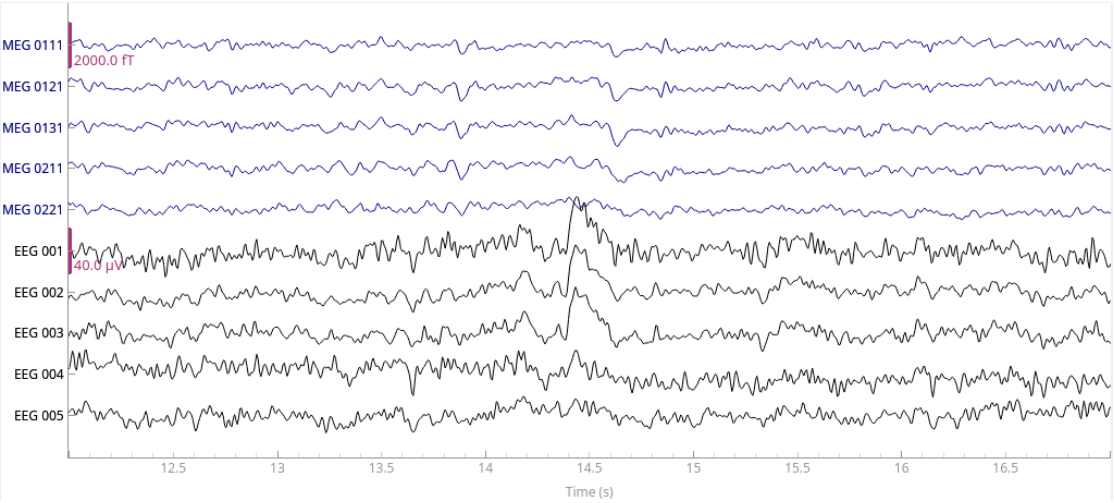

Figure 1.6: **Pre-processed M/EEG signals from the MNE sample dataset [Gramfort et al., 2014].** The signals, corresponding to the ones displayed in Figure 1.5, have been cleaned using Independent Component Analysis (ICA) with 20 components. In the MEG signals cardiac artifacts were removed, and in the EEG signals occular artifacts are attenuated.

removing noise and artifacts from the acquired signals and introduce the classical pre-processing steps in M/EEG analysis. Then, we detail the statistical methods commonly used in the litterature, including a-priori defined features extraction, supervised spatial filtering, and deep learning.

1.2.1 Pre-processing

The acquired M/EEG signals are often contaminated by noise and artifacts rising from other sources than the neuronal activity in the brain in which we are interested. These artifacts can be caused by the environment such as the frequency of the AC power line or vibrations and electromagnetic field noise coming from activities in the surrounding building. The instrumention used for the recording also induce noise due to sensors malfunction or to the Analog-to-Digital Conversion. Even biological phenomenon like heart beats and eye movements contaminate M/EEG signals.

Pre-processing these data is thus a consequential step to access the signal of interest, linked to the underlying brain activity. Several studies have shown that the choice of pre-processing steps can have a significant impact on the performance of predictive models [Bomatter et al., 2024, Delorme, 2023]. While Delorme [2023] demonstrated that simpler pre-processing steps are often sufficient in non-ML contexts, Bomatter et al. [2024] highlighted that minimal pre-processing in EEG data can lead to suboptimal results in machine learning applications, where thorough artifact cleaning is essential not only for performance but also for model inter-

pretability.

Figure 1.6 shows the effect of pre-processing on the signals with only one artefact removal method, compared to the signals displayed in Figure 1.5. In the following we introduce the pre-processing steps in M/EEG analysis that were used in this thesis.

Handling bad channels

Sensors, or channels, sometimes malfunction for a part or for the entire recording session, leading to excessively noisy or flat signals. These channels are referred to as bad channels. Detection of bad channels is often done manually, either during data acquisition or by visually inspecting the signals afterward. Once identified, researchers can either ignore these channels during the downstream analysis or attempt to repair them. Ignoring bad channels can be a straightforward solution; however, in cross-subject analysis, it is important to maintain the same data dimension across all subjects. Therefore, repairing bad channels is often preferred. Interpolation techniques are commonly used for this purpose, with Spherical Spline Interpolation (SSI) being the standard for EEG data and Field Interpolation (FI) for MEG data. These methods allow for the reconstruction of the missing or noisy data by estimating the signal from the surrounding channels, thereby preserving the data integrity and dimensionality for subsequent analyses. More details about these interpolation techniques are presented in Chapter 4.

Spectral filtering and resampling

Filtering is an essential step in the pre-processing pipeline of M/EEG data to reduce noise, though it can have unwanted effects if not used carefully [Widmann et al., 2015].

The M/EEG signals relevant for neuroscience are primarily in low frequencies, typically between 1 Hz to 50 Hz. To isolate these frequencies, a bandpass filter is applied. The high-pass filter removes low-frequency noise, such as that caused by head movements, while the low-pass filter eliminates high-frequency noise, like muscle activity. However, this bandpass filtering alone may not be sufficient to remove noise that overlaps with the frequencies of interest. A notch filter is often applied to the raw data to remove power-line noise at 50 or 60 Hz and its harmonics. This zero-phase filter targets narrow frequency bands around the selected frequencies, but it must be used cautiously to avoid distorting the signal.

M/EEG data are typically recorded at high sampling rates, around 1000 Hz or more, providing high temporal precision essential for analyzing the timing of events. However, this high sampling rate can be memory and computationally expensive. In cases where high-frequency information is not necessary, such as in the analyses performed in this thesis, downsampling the data can save time and resources. The Nyquist-Shannon sampling theorem must be respected during

resampling, ensuring that the sampling frequency is at least twice the highest frequency present in the signal, to prevent aliasing.

Artifact detection and removal

Even though more advanced tools have been developed for artifact detection [Allain et al., 2022] and removal, Independent Component Analysis (ICA) and Signal-Space Projection (SSP) [Uusitalo and Ilmoniemi, 1997] remain the most commonly used methods for cleaning M/EEG data.

ICA is a blind source separation technique that decomposes the recorded signal into independent components. It allows for the separation of the brain signal of interest from various noise artifacts. The noise components, often related to ocular and muscle artifacts, are identified manually and removed from the data. Thus, ICA is primarily used to eliminate biological artifacts.

On the other hand, SSP projects the data into a lower-dimensional subspace that is orthogonal to the artifact subspace. The artifact subspace is estimated either from the covariance matrices computed over artifact segments or from empty-room recordings where the subject is not present. SSP is mainly employed to remove stationary environmental artifacts, such as power-line noise, but can also be used to address biological artifacts.

Additionally, the autoreject method automatically detects and repairs bad data segments [Jas et al., 2017], known as epochs, in M/EEG recordings. It identifies these segments using peak-to-peak amplitude thresholding, optimized through cross-validation. For repairing bad epochs, autoreject employs interpolation methods as mentioned previously in the **handling bad channels** paragraph.

Setting the EEG reference

EEG signals represent the difference in electric potential between the recording electrodes and a reference electrode. This step is specific to EEG data. Because the signals from the reference electrode will be subtracted from all other electrodes, it is essential that the reference electrode captures only the common noise across all electrodes and not the brain activity of interest. Typically, the reference electrode is placed in a location close to the head but far from areas generating significant brain activity. Common locations include the earlobe, the nose, or the mastoid. However, in setups where no dedicated reference electrode is present, one of the recording electrodes may be chosen as the reference, or the Common Average Reference (CAR) technique can be employed. It is also systematically used to perform inverse modeling and source imaging. In CAR, the average of all electrode signals is used as the reference, helping to minimize the influence of any one electrode's specific activity on the overall recording.

Signal-Space Separation

Signal-Space Separation (SSS) is a preprocessing step specific to MEG data. SSS is a method that separates the MEG signals into components originating from inside the head and components coming from external sources outside the head [Taulu and Kajola, 2005]. This technique exploits the physics of electromagnetic fields behind the MEG signals. Similar to SSP, SSS projects the data into a lower-dimensional subspace, but unlike SSP, which empirically estimates the subspace from the data, SSS derives the subspace theoretically based on the physics of MEG signals. A derived version of SSS, known as the Maxwell filter, was developed for Neuromag data. The Maxwell filter is used on the MEG data in this thesis, effectively improving the signal-to-noise ratio by removing external noise and ensuring the focus remains on brain-generated signals.

Creating epochs

M/EEG signals are generally recording continuously during several minutes to several hours. To analyze these data, they are divided into smaller segments called epochs. The length of these epochs can vary depending on the type of analysis performed. For example, in event-related analysis, epochs are usually centered around the event of interest: for example 200 ms before the event and 500 ms after the event for auditory stimuli. In resting-state analysis, epochs are usually defined to be longer, around 10 seconds, to capture the spontaneous brain activity. To have a good signal-to-noise ratio, it is often necessary to average many epochs corresponding to a same type of event.

1.2.2 Common approaches to predict from M/EEG data

The goal is to predict an outcome variable $y \in \mathcal{Y}$ related to brain activity recorded with M/EEG. We denote the M/EEG signal at time t as $\boldsymbol{x}(t) \in \mathbb{R}^P$. Examples of y include the age of the subject or the presence of a disease (further applications are discussed in Section 1.1.3). Due to the complexity of the data and the high dimensionality of the feature space, statistical learning, or machine learning (ML), is employed to predict y from $\boldsymbol{x}(t)$. ML methods approximate the relationship between $\boldsymbol{x}(t)$ and y by learning a function $f : \mathbb{R}^P \to \mathcal{Y}$ that minimizes a loss function $\mathcal{L}(y, f(\boldsymbol{x}(t)))$ between the predicted outcome $\hat{y} = f(\boldsymbol{x}(t))$ and the true (known) outcome y. Predictions can be made at the event level, such as sleep staging, or at the participant level, such as brain age prediction. If y is continuous, regression models are used, whereas classification models are used if y is categorical. The aim is to train ML models on a dataset $\mathcal{D} = \{(\boldsymbol{x}_i(t), y_i)\}_{i=1}^N$ and ensure they generalize well to unseen data, including new recordings or participants. In the following sections, we introduce popular machine learning approaches to predict from M/EEG data [Engemann et al., 2022, Gemein et al., 2020].

A-priori defined features

A-priori defined features were one of the first approaches developed to decode M/EEG data and remain popular due to their interpretability. These features are extracted from signals based on findings from neuroscience studies, using prior knowledge of the brain and the outcome to predict. After pre-processing, features are extracted from the clean epochs and then used in the machine learning (ML) model of choice. It is also possible to extract many features and use an ML model, such as a random forest [Breiman, 2001], which can handle non-informative features effectively.

Commonly extracted features include common statistical features like mean, variance, skewness, quantiles, peak-to-peak amplitude, kurtosis, spectral entropy and power, ... [Estrada et al., 2004, Hosseinifard et al., 2013, Engemann et al., 2022]. 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 [Buzsáki and Llinás, 2017]. This includes metrics like spectral entropy and power ratios in decibels among defined frequency bands, which reflect the different bands and their importance in brain activity.

End-to-end approaches

End-to-end approaches, popular in the fields of computer vision [Krizhevsky et al., 2012] and natural language processing [Deng and Liu, 2018], are increasingly being applied to M/EEG data analysis [Roy et al., 2019]. These methods use raw or minimally pre-processed signals as input to the model, eliminating the need for manual feature extraction. Instead, features are learned directly by the model through joint optimization of the features and the prediction function. While this approach offers significant advantages, including the potential for improved performance by learning highly relevant features, it is less interpretable due to the complexity and the learned nature of the features.

End-to-end models are data and computation intensive, requiring large amounts of data and significant computational resources. This is feasible in fields where large datasets are readily available, but it poses a challenge for M/EEG applications where data are more scarce. Despite these challenges, various deep learning architectures have been investigated for their applicability to M/EEG data [Schirrmeister et al., 2017, Supratak et al., 2017]. These architectures can be designed to incorporate prior knowledge about the data, aiding in the extraction of relevant features [Banville et al., 2021].

Covariance-based approaches

Covariance-based approaches sit between feature-based approaches and end-to-end approaches. They use spatial covariance matrices of pre-processed signals as

representations and input data.

These methods include supervised spatial filtering approaches like Source Power Comodulation (SPoC) [Dähne et al., 2014], which extracts spatial filters that maximize the correlation between the signal and the outcome variable, and Common Spatial Filtering (CSP) [Koles et al., 1990], which extracts spatial filters that maximize the variance of the signal for one class and minimize it for the other class.

Other covariance-based approaches benefit from the invariant features to field spread when utilizing Riemannian geometry. This has been popularized in BCI applications and research [Barachant et al., 2012]. Instead of designing and optimizing filters based on covariance matrices, these methods use mathematical tools from the Riemannian geometry of Symmetric Positive Definite (SPD) matrices. The covariance matrices are either used with minimum distance to mean (MDM) algorithms on the manifold or transformed into Euclidean vectors that can be used in any linear ML model. These approaches have been investigated and compared in [Sabbagh et al., 2019, 2020] on both simulated and real M/EEG data. Covariance matrices can be computed from multiple frequency bands to capture the spectral content of the data, enhancing the representation.

A comparison of different ML approaches to predict from M/EEG data was conducted on four datasets in a brain age benchmark study [Engemann et al., 2022]. The results from this benchmark showed that the covaraince-based approach achieved comparable results to deep learning approaches. In addition, the former approach is particularly beneficial when working with smaller datasets: deep learning models, which tend to have more parameters, are more likely to overfit small datasets. A detailed presentation of the covariance-based approach is provided in Section 1.3.

1.2.3 Evaluation metrics

To evaluate the performance of the predictive model on M/EEG data, a range of metrics are commonly used in the literature. The Mean Absolute Error (MAE), the R^2 score and the Spearman correlation are used for regression tasks, while the Accuracy and the Balanced Accuracy are used for classification tasks.

Mean Absolute Error

Mean Absolute Error (MAE) is a metric used to measure the average magnitude of errors in a set of predictions. It is calculated as the average of the absolute differences between predicted values and actual values:

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i| \tag{1.1}$$

where y_i is the true value, $\hat{y}_i$ is the prediction, and N is the number of observations.

MAE is easy to interpret as it provides a direct measure of the average error. The lower the MAE, the better the model's performance in predicting the outcome variable. Unlike Mean Squared Error (MSE), MAE is not overly sensitive to outliers. However, it does not provide information on the direction of errors, meaning it does not indicate whether predictions are generally overestimating or underestimating the actual values.

R^2 Score

The R^2 score, also known as the coefficient of determination, is a regression score function that indicates how well the model fits the data. It measures the proportion of the variance in the dependent variable that is predictable from the independent variables:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{1.2}$$

where $\bar{y}$ is the mean of the true values.

The R^2 score provides a measure of how well the model explains the variability of the response data. The best possible score is 1, indicating that the model led to perfect predictions. A score of 0 means the model always predicts the mean $\bar{y}$. The R^2 score can also be negative, meaning that the model is worse than predicting the mean.

Spearman Correlation

Spearman correlation is a non-parametric measure that assesses how well the relationship between two variables can be described using a monotonic function. It is calculated as:

$$\rho = \frac{\mathrm{cov}(\mathrm{rk}_y, \mathrm{rk}_{\hat{y}})}{\sigma_{\mathrm{rk}_y}\sigma_{\mathrm{rk}_{\hat{y}}}} \tag{1.3}$$

where rk_y and $\mathrm{rk}_{\hat{y}}$ are respectively the rank variables of the true values and the predicted values, σ_{rk_y} and $\sigma_{\mathrm{rk}_{\hat{y}}}$ are their standard deviation.

A positive sign indicates that the variables move in the same direction, while a negative sign indicates that they move in opposite directions. Scores of 1 and -1 signify perfect monotonicity between the variables, either increasing or decreasing respectively. A score of 0 means there is no correlation, indicating no clear direction of $\hat{y}$ when y increases.

Accuracy and Balanced Accuracy

Accuracy is a metric that measures the proportion of correct predictions among the total number of cases examined:

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \tag{1.4}$$

Accuracy is simple to understand and compute. However, it can be misleading in the presence of imbalanced data, as it does not account for the distribution of classes.

Balanced Accuracy is used to address the limitations of accuracy in imbalanced datasets. It is the average of the recall obtained on each class. For a binary classification problem it is defined as:

$$\text{Balanced Accuracy} = \frac{1}{2}\left(\frac{\text{TP}}{\text{TP} + \text{FN}} + \frac{\text{TN}}{\text{TN} + \text{FP}}\right) \tag{1.5}$$

where TP represents true positives, TN represents true negatives, FP represents false positives, and FN represents false negatives.

1.3 Predictive modeling with M/EEG covariance matrices

In this section, we describe the covariance-based approach for predictive regression modeling that is used throughout this thesis. First, the concepts of Riemannian geometry on covariance matrices are presented, as well as the related predictive pipeline. Then, we introduce the covariance matrix representation of M/EEG signals.

1.3.1 Riemannian geometry of SPD matrices

Covariance matrices belong to the set $\mathbb{S}_P^{++}$ of $P \times P$ symmetric positive definite matrices [Skovgaard, 1984, Pennec et al., 2006]. $\mathbb{S}_P^{++}$ is open in the set $\mathbb{S}_P$ of $P \times P$ symmetric matrices, and thus $\mathbb{S}_P^{++}$ is a smooth manifold [Boumal, 2023].

A manifold is a set equipped with a smooth structure that allows defining Euclidean vector spaces at each point of the manifold, called tangent spaces, which locally approximate the manifold [Boumal, 2023]. We denote the tangent space at $\boldsymbol{\Sigma} \in \mathbb{S}_P^{++}$ as $T_{\boldsymbol{\Sigma}}\mathbb{S}_P^{++}$, which is equal to the set $\mathbb{S}_P$. The tangent space and the manifold have the same dimension.

When a smooth inner product is defined at every tangent space, the smooth manifold becomes a Riemannian manifold. A common choice is the affine invariant Riemannian metric [Skovgaard, 1984, Förstner and Moonen, 2003, Pennec et al., 2006]. This metric adds useful theoretical properties to the $\mathbb{S}_P^{++}$ manifold, such as being geodesically complete and putting null eigenvalues at an infinite distance. Given $\boldsymbol{\Gamma}, \boldsymbol{\Gamma}' \in T_{\boldsymbol{\Sigma}}\mathbb{S}_P^{++}$, this metric is:

$$\langle \boldsymbol{\Gamma}, \boldsymbol{\Gamma}' \rangle_{\boldsymbol{\Sigma}} = \text{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Gamma}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Gamma}'\right) \tag{1.6}$$

The associated norm on the tangent space of $\boldsymbol{\Gamma} \in T_{\boldsymbol{\Sigma}}\mathbb{S}_P^{++}$ is defined as:

$$\|\boldsymbol{\Gamma}\|_{\boldsymbol{\Sigma}}^2 = \langle \boldsymbol{\Gamma}, \boldsymbol{\Gamma} \rangle_{\boldsymbol{\Sigma}} \tag{1.7}$$

The norm at the tangent space at the identity $T_{\boldsymbol{I}}\mathbb{S}_P^{++}$ is the Frobenius norm.

Riemannian distance

The Riemannian distance, also called geodesic distance, between two points on the manifold associated with the affine invariant metric defined in (1.6) is:

$$\delta_R(\boldsymbol{\Sigma}, \boldsymbol{\Sigma}') \triangleq \left\| \log\left(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}'\boldsymbol{\Sigma}^{-1/2}\right) \right\|_F, \tag{1.8}$$

where $\log : \mathbb{S}_P^{++} \to \mathbb{S}_P$ is the matrix logarithm defined as:

$$\log(\boldsymbol{\Sigma}) = \boldsymbol{U} \operatorname{diag}(\log(\lambda_1), \ldots, \log(\lambda_P))\boldsymbol{U}^\top, \tag{1.9}$$

with $\boldsymbol{\Sigma} = \boldsymbol{U} \operatorname{diag}(\lambda_1, \ldots, \lambda_P)\boldsymbol{U}^\top$ representing the singular value decomposition (SVD) of $\boldsymbol{\Sigma}$.

It should be noted that, since the Riemannian metric is affine invariant, δ_R is also affine invariant, i.e., for every $\boldsymbol{A} \in \mathbb{R}^{P \times P}$ invertible, we have:

$$\delta_R(\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top, \boldsymbol{A}\boldsymbol{\Sigma}'\boldsymbol{A}^\top) = \delta_R(\boldsymbol{\Sigma}, \boldsymbol{\Sigma}'). \tag{1.10}$$

Affine invariance is an essential property for predictive modeling with M/EEG data because it ensures that the distance between M/EEG signals associated with the covariance matrices remains unaffected by linear transformations. This is particularly powerful since many effects on M/EEG data can be approximated by linear transformations. Examples of such effects include slightly moving the sensor positions, having different measurement scales or units, and encountering different mixtures of brain activities in a person's brain [Yger et al., 2017, Congedo et al., 2017].

Riemannian mean

With the Riemannian distance defined, we can now identify the matrix that is equidistant from a set of N matrices. This point of interest is known by various names in the literature, such as the center of mass, Fréchet mean, and Karcher mean. In this work, we refer to it as the Riemannian mean.

This Riemannian mean, denoted as $\overline{\boldsymbol{\Sigma}}$, is defined for a set $\{\boldsymbol{\Sigma}_i\}_{i=1}^N \subset \mathbb{S}_P^{++}$ as

$$\overline{\boldsymbol{\Sigma}} \triangleq \underset{\boldsymbol{\Sigma} \in \mathbb{S}_P^{++}}{\arg\min} \sum_{i=1}^{N} \delta_R^2(\boldsymbol{\Sigma}, \boldsymbol{\Sigma}_i) \tag{1.11}$$

There is no closed-form solution for (1.11) when $N \geq 3$. However, a solution for its optimization problem always exists [Karcher, 1977] In practive, the Riemannian mean is computed with an iterative algorithm, such as the Riemannian gradient descent [Pennec et al., 2006, Zhang and Sra, 2016].

1.3.2 Learning on the Riemannian manifold

Some approaches exist to directly learn a model on the manifold, such as the Minimum Distance to Mean (MDM) classifier [Barachant et al., 2012] or the geodesic regression [Fletcher, 2011]. Alternatively, the tangent space approaches have the advantage of defining a linear vector space that can be operated using simple algebra, thus allowing the use of classical machine learning models. In the following we introduce the concept to perform the transformation of covariance matrices into vectors using the Riemannian logarithmic mapping and the vectorization operator.

Riemannian logarithmic mapping

As introduced previously, an Euclidean space is defined at each point of the manifold. A mapping that transforms a point of the manifold into a point of the tangent space at this point while preserving the local properties of the manifold. This mapping is called the logarithmic mapping. When using the Riemannian metric, this logarithmic mapping of $\boldsymbol{\Sigma}'$ at $\boldsymbol{\Sigma}$ is defined as

$$\log_{\boldsymbol{\Sigma}}(\boldsymbol{\Sigma}') \triangleq \boldsymbol{\Sigma}^{1/2} \log\left(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}'\boldsymbol{\Sigma}^{-1/2}\right)\boldsymbol{\Sigma}^{1/2} \in T_{\boldsymbol{\Sigma}}\mathbb{S}_P^{++} \,. \tag{1.12}$$

with the matrix power defined as:

$$\boldsymbol{\Sigma}^{\alpha} = \boldsymbol{U}\,\mathrm{diag}(\lambda_1^{\alpha},\ldots,\lambda_P^{\alpha})\boldsymbol{U}^{\top}. \tag{1.13}$$

The logarithmic mapping (1.12) at the identity is simply the matrix logarithm.

The inverse mapping, to go from the tangent space to the manifold, is called the exponential mapping, and is defined as

$$\exp_{\boldsymbol{\Sigma}}(\boldsymbol{\Gamma}) \triangleq \boldsymbol{\Sigma}^{1/2} \exp\left(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Gamma}\boldsymbol{\Sigma}^{-1/2}\right)\boldsymbol{\Sigma}^{1/2} \in \mathbb{S}_P^{++} \,. \tag{1.14}$$

Vectorization operator

After using the log mapping, we obtain an element of the tangent space. To convert this matrix representation into a Euclidean vector suitable for machine learning algorithms, we want to use a vectorization operator. Since the tangent space is isomorphic to the set of symmetric matrices, we can use a vectorization operator which flattens the upper triangle part of the matrix.

Following this procedure, we have a non-linear feature extraction method of a dataset $\{\boldsymbol{\Sigma}_i\}_{i=1}^N$ by computing the Riemannian logarithmic mapping of each covariance matrix at their Riemannian mean $\overline{\boldsymbol{\Sigma}}$ and then vectorizing the resulting matrices:

$$\phi(\boldsymbol{\Sigma}_i, \overline{\boldsymbol{\Sigma}}) \triangleq \mathrm{uvec}\left(\log\left(\overline{\boldsymbol{\Sigma}}^{-1/2}\boldsymbol{\Sigma}\overline{\boldsymbol{\Sigma}}^{-1/2}\right)\right) \in \mathbb{R}^{P(P+1)/2} \tag{1.15}$$

where uvec : $\mathbb{S}_P \to \mathbb{R}^{P(P+1)/2}$ vectorizes the upper triangle part with off-diagonal elements multiplied by $\sqrt{2}$ to preserve the norm.

This method has been widely applied in M/EEG data for various applications, such as Brain Computer Interfaces (BCI) [Barachant et al., 2012], or biomarker exploration [Sabbagh et al., 2020].

1.3.3 Covariance matrix of M/EEG signals

The spatial covariance matrix serves as a useful representation of M/EEG signals, offering a framework for analyzing the statistical properties of multivariate time series. Under certain statistical assumptions, it is possible to estimate a set of parameters describing the statistical behavior of multivariate time series, including the mean vector and the cross-spectral density matrices, as well as band-specific covariance matrices. Predictive models can then be developed based on these covariance matrices and the associated mathematical tools discussed previously. However, M/EEG spatial covariance matrices often suffer from rank deficiency, which prevents them from being symmetric positive definite (SPD). This issue is typically addressed through the application of regularization methods or dimensionality reduction techniques.

Statistical assumptions

To analyze M/EEG signals, the classical approach is to consider the signal at time t as a random vector $\boldsymbol{x}(t) \in \mathbb{R}^P$. Assumptions of wide-sense (up to the second order) stationarity and ergodicity enable the representation of M/EEG time series using cross-spectral covariance matrices. Wide-sense stationarity implies that the mean of the multivariate time serie is constant over time:

$$\mathbb{E}[\boldsymbol{x}(t)] = \mu \quad \forall t \tag{1.16}$$

and the autocovariance matrix between times t and $t+\tau$ is only a function of τ, and not of t:

$$\boldsymbol{R}(t, t+\tau) = \boldsymbol{R}(\tau) = \mathbb{E}[(\boldsymbol{x}(t+\tau) - \mu)(\boldsymbol{x}(t) - \mu)^\top] \tag{1.17}$$

Under this assumption and following the Wiener-Khinchin theorem, we define the cross-power spectral density of a multivariate time serie as the discrete-time Fourier transform (DTFT) of the autocovariance matrices [Priestley, 1981]:

$$\boldsymbol{S}(f) = \sum_{k=-\infty}^{+\infty} \boldsymbol{R}(k) e^{-j2\pi fk}, \tag{1.18}$$

In practice, since we have only a finite number samples, the sum has a finite number of terms and is convergent. The cross-spectral density matrices $\boldsymbol{S}(f)$ are positive definite matrices.

However, assuming wide-sense stationarity alone is often inadequate for M/EEG signals. Instead, we assume that the mean and autocovariance of the multivariate time series are constant over small time windows around t.

Wide-sense ergodicity implies that if a sufficiently large set of random samples is taken from a process, it can represent the average statistical properties of the entire process. Combining wide-sense ergodicity and stationarity, we can estimate:

$$\hat{\boldsymbol{\mu}} = \frac{1}{T}\sum_{t=0}^{T-1}\boldsymbol{x}(t) \tag{1.19}$$

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

$$\hat{\boldsymbol{S}}(f) = \mathrm{DTFT}(\hat{\boldsymbol{R}}(\tau)) \tag{1.21}$$

These estimations provide a framework to capture the statistical properties of M/EEG signals over time.

Covariance matrix

A $\mathcal{F}$-bandpass filtered M/EEG signal $\boldsymbol{x}(t)$ is zero-mean due to the highpass filtering part of the bandpass, thus $\mu = 0$. Therefore the inverse DTFT of the cross-spectral density matrix is the covariance matrix of $\boldsymbol{x}(t)$:

$$\int_{\mathcal{F}} \boldsymbol{S}(f)df = \boldsymbol{R}(0) = \mathbb{E}[\boldsymbol{x}(t)\boldsymbol{x}(t)^\top] = \boldsymbol{\Sigma} \tag{1.22}$$

The M/EEG signals, denoted as $\boldsymbol{X} \in \mathbb{R}^{P\times T}$, are multivariate time series recorded from d sensors over T time points. We now consider the covariance matrix of M/EEG signals, denoted $\boldsymbol{\Sigma} \in \mathbb{S}_P^{++}$. This matrix is defined as:

$$\boldsymbol{\Sigma} = \frac{1}{T}\boldsymbol{X}\boldsymbol{X}^\top \ , \tag{1.23}$$

where T is the number of samples and $\boldsymbol{X} \in \mathbb{R}^{P\times T}$ the M/EEG data matrix. The covariance of M/EEG signals holds the sensors' variance, or power, on its diagonal. The off-diagonal terms correspond to the covariance between sensors.

These covariance matrices are usually computed in several frequency bands by bandpass filtering the M/EEG signals. In this situation, one covariance matrix is computed per frequency band. As each matrix is a symmetric positive definite matrice, the concepts introduced in Section 1.3.1 and Section 1.3.2 can be applied to them independently. By repeating this process for each observation, we obtain a set of covariance matrices $\{\boldsymbol{\Sigma}_i\}_{i=1}^N$ per frequency bands.

Rank deficiency of M/EEG covariance matrices

M/EEG preprocessing steps often lead to rank deficiency. For instance, re-referencing EEG signals to a common average reference, applying ICA to remove artifacts, using SSS to remove environmental noise, or applying SSP (Signal Subspace Projection) to remove EOG and ECG artifacts can result in rank deficiency.

To address this rank deficiency, one approach is to regularize the covariance matrices using shrinkage methods to obtain a regularized full-rank matrix:

$$\boldsymbol{\Sigma}_{\text{shrunk}} = (1 - s)\boldsymbol{\Sigma} + s\mu\boldsymbol{I}_P \ , \tag{1.24}$$

with $\mu = \text{tr}(\boldsymbol{\Sigma})/P$ and s the shrinkage coefficient that can be estimated with methods such as Ledoit-Wolf [Ledoit and Wolf, 2004] or the Oracle Approximating Shrinkage (OAS) [Chen et al., 2010].

Another effective method to ensure full-rank matrices is to apply a Principal Component Analysis (PCA) to reduce the dimensionality of the covariance matrices, rendering them full rank [Sabbagh et al., 2019]. The final dimension of the covariance matrices after PCA is denoted as R. We use as filters the PCA eigenvectors $\boldsymbol{W} \in \mathbb{R}^{R \times P}$ corresponding to the R largest eigenvalues of the mean covariance matrices $\overline{\boldsymbol{\Sigma}}$ as defined in (1.11). The resulting filtered matrices are:

$$\boldsymbol{\Sigma}_{\text{filtered}} = \boldsymbol{W}\boldsymbol{\Sigma}\boldsymbol{W}^\top \ \in \mathbb{R}^{R \times R} \ . \tag{1.25}$$

These methods ensure that the covariance matrices used for subsequent analysis are full rank and suitable for the Riemannian geometry framework.

1.4 Domain adaptation for biological signals

In machine learning applied to biological signals, achieving the most accurate predictive models often necessitates training on large datasets to embrace the variability of these data. However, gathering substantial data in neuroscience poses significant challenges due to the high cost and complexity of data collection. Recently, several large databases for health applications have emerged [Andreu-Perez et al., 2015], facilitating the development and training of machine learning models. Nevertheless, a critical issue arises: models trained on specific datasets often fail to generalize to new datasets due to high variabilities between them [Dockès et al., 2021, Quiñonero-Candela et al., 2022]. These variabilities in biological data stem from statistical discrepancies across different populations, recording devices, experimental conditions, and more. Addressing this challenge requires the development of models that can effectively generalize to new datasets, known as domain adaptation. In the following sections, we introduce domain adaptation and its mathematical framework. We discuss dataset shifts examples that can be observed in biological signals and discuss existing methods aimed at reducing these shifts.

1.4.1 The challenge of domain adaptation

Machine learning aims at learning a model on training data to predict on new unseen data. However, most models assume that the training and testing data are drawn from the same distribution as well as the same feature space and the same labels distribution. However, traditional models assume that the training and testing data are drawn from the same distribution, feature space, and label distribution. In practice, especially with biological signals, these assumptions are often not met. Domain adaptation addresses this challenge [Ben-David et al., 2010, Redko et al., 2019] and is is closely related to transfer learning. While transfer learning generally refers to the process of applying knowledge gained from one task or domain to another, domain adaptation specifically addresses the scenario where the source and target domains exhibit different distributions.

In computer vision, domain adaptation is applied for tasks such as object detection and image classification [Csurka et al., 2017]. For example, a model trained to detect objects in daytime images may struggle with images taken at night or in diverse weather conditions. Domain adaptation methods adjust the model to perform well across varied conditions without extensive training on nighttime or weather-specific datasets. Similarly, in speech recognition, models trained on one accent or language may not generalize to others [Sun et al., 2018]. Domain adaptation techniques improve model performance across various accents and dialects, improving the usability of speech recognition systems.

In biomedical signals like M/EEG data, various factors contribute to different data distributions between datasets. Differences in recording devices, experimental conditions, subjects, and other variables can all impact the distribution of data. Domain adaptation methods are essential in this context to align and adapt models effectively, ensuring robust performance across different datasets and improving the reliability of biomedical signal analysis [Dockès et al., 2021].

1.4.2 Mathematical framework

We denote the input feature space $\mathcal{X}$ and the output label space $\mathcal{Y}$. In classical statistical learning we assume that the training and the testing data to be drawn from the joint feature/label distribution $\mathcal{P}$. We denote $\mathcal{P}_{\mathcal{X}}$ and $\mathcal{P}_{\mathcal{Y}}$ respectively the feature and the label marginals of $\mathcal{P}$. In addition, the risk function of a model h with respect to a distribution $\mathcal{P}$ is defined as the expected loss between the prediction of the model and the true label:

$$R(h) = \mathbb{E}_{\mathcal{P}}[\mathcal{L}(y, h(\boldsymbol{x}))] \tag{1.26}$$

where $\mathcal{L}$ is the loss function.

However, in practical applications such as biological signals, training and testing data are often drawn from different marginal distributions $\mathcal{P}_{\mathcal{S}} \neq \mathcal{P}_{\mathcal{T}}$ leading to what we call dataset shift. Here, $\mathcal{P}_{\mathcal{S}}$ is the training distribution, or source domain, and $\mathcal{P}_{\mathcal{T}}$ is the testing distribution, or target domain. In this context, a ML models trained on the source domain $\mathcal{P}_{\mathcal{S}}$ might fail to generalize to the target domain $\mathcal{P}_{\mathcal{T}}$.

Domain adaptation aims at learning a function h that performs well on $\mathcal{P}_\mathcal{T}$ using the knowledge acquired on $\mathcal{P}_\mathcal{S}$. Specifically, we want the model trained on the source domain to minimize the expected risk of the source labeled data:

$$R_\mathcal{S}(h) = \mathbb{E}_{\mathcal{P}_\mathcal{S}}[L(y, h(\boldsymbol{x}))], \tag{1.27}$$

while also minimizing the expected risk on unlabeled target data:

$$R_\mathcal{T}(h) = \mathbb{E}_{\mathcal{P}_\mathcal{T}}[L(y, h(\boldsymbol{x}))] \tag{1.28}$$

Domain adaptation approaches can be categorized into unsupervised and supervised methods based on the availability of labeled data in the target domain. Here, we focus on unsupervised domain adaptation.

The source dataset is denoted $\mathcal{D}_\mathcal{S} = \{(\boldsymbol{x}_{\mathcal{S},i}, y_{\mathcal{S},i})\}_{i=1}^{N_\mathcal{S}}$ with $N_\mathcal{S}$ the number of samples in the source dataset, and $\boldsymbol{x}_{\mathcal{S},i}, y_{\mathcal{S},i} \sim \mathcal{P}_\mathcal{S}$, and the target dataset is denoted $\mathcal{D}_\mathcal{T} = \{(\boldsymbol{x}_{\mathcal{T},i})\}_{i=1}^{N_\mathcal{T}}$ with $N_\mathcal{T}$ the number of samples in the target dataset, and $\boldsymbol{x}_{\mathcal{T},i} \sim \mathcal{P}_{\mathcal{X},\mathcal{T}}$. The shift between the source and target domains can occur in many ways. In the following we discuss some common types of shifts: covariate shift, target shift and concept shift, illustrated in Figure 1.7.

The covariate shift occurs when the conditionals probabilities $P_\mathcal{S}(y|\boldsymbol{x}) = P_\mathcal{T}(y|\boldsymbol{x})$ are equal between the source and target domains, but the feature marginal probabilities are different, i.e., $P_{\mathcal{X},\mathcal{S}}(\boldsymbol{x}) \neq P_{\mathcal{X},\mathcal{T}}(\boldsymbol{x})$.

The target shift also called prior shift or label shift, occurs when the conditional probabilities $P_\mathcal{S}(\boldsymbol{x}|y) = P_\mathcal{T}(\boldsymbol{x}|y)$ are equal between the source and target domains, but the label marginal probabilities are different, i.e., $P_{\mathcal{Y},\mathcal{S}}(y) \neq P_{\mathcal{Y},\mathcal{T}}(y)$.

The concept shift occurs when the conditional probabilities $P_\mathcal{S}(y|\boldsymbol{x}) \neq P_\mathcal{T}(y|\boldsymbol{x})$ or $P_\mathcal{S}(\boldsymbol{x}|y) \neq P_\mathcal{T}(\boldsymbol{x}|y)$ are different between the source and target domains.

In some applications, like BCI or sleep staging, models are trained on several source domains, referred to as multi-source domain adaptation. In this situation, we denote the source datasets with the index k as $\mathcal{D}_k = \{(\boldsymbol{x}_{k,i}, y_{k,i})\}_{i=1}^{N_k}$, while we still denote the target dataset as $\mathcal{D}_\mathcal{T} = \{(\boldsymbol{x}_{\mathcal{T},i})\}_{i=1}^{N_\mathcal{T}}$.

1.4.3 Common domain adapatation methods

Several strategies exist to tackle dataset shift in machine learning models. Domain adaptation methods compatible with classical ML models can be broadly categorized into two main categories [Farahani et al., 2021]. Additionally, there are deep domain adaptation techniques that leverage deep network properties; however, these are not detailed here as they do not fall within our current scope.

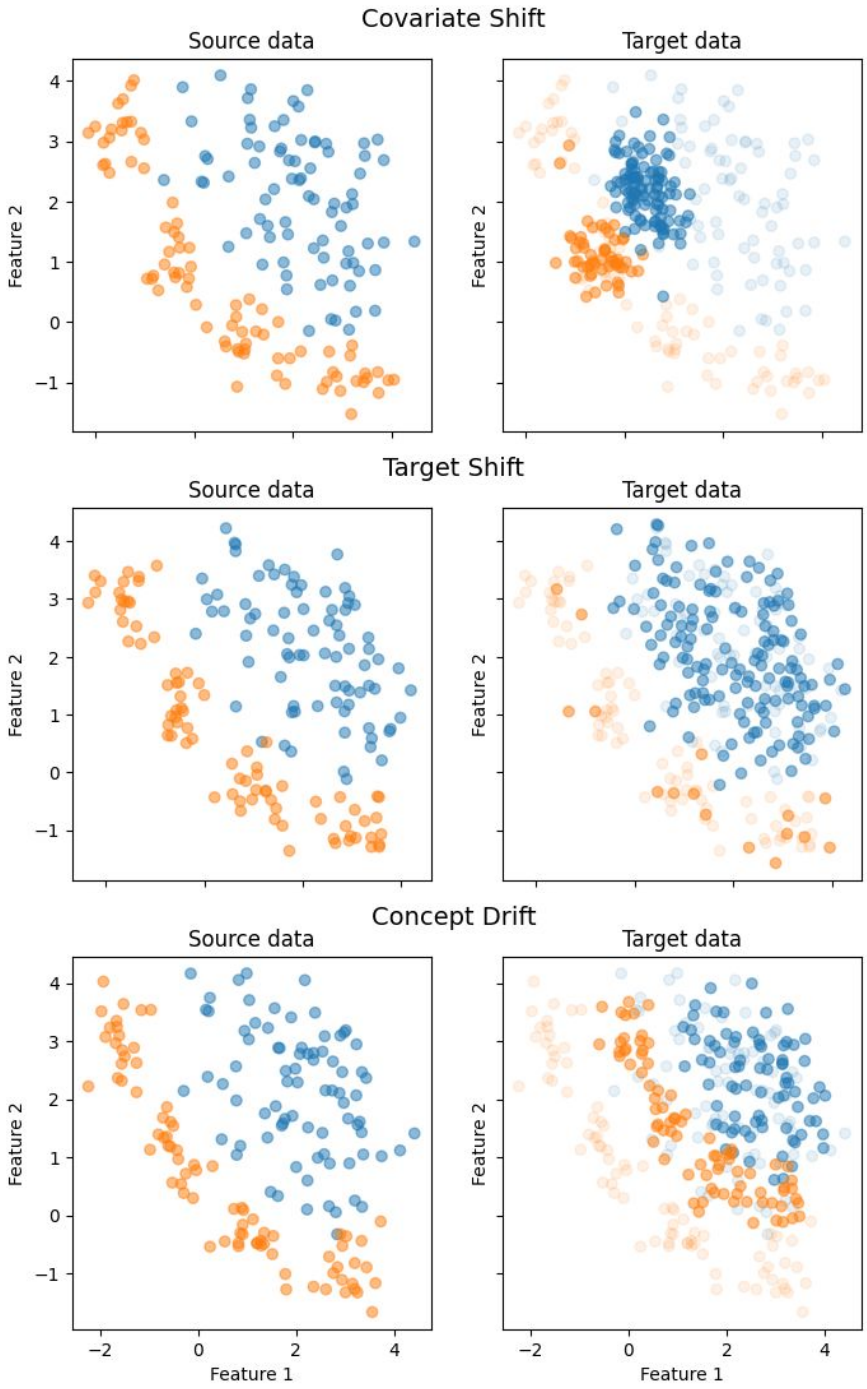

Figure 1.7: **Illustration of three types of dataset shift.** Adapted from the documentation of SKADA python library for domain adaptation [Gnassounou et al., 2024]

Instance-based or sample reweighting

The idea behind sample reweighting is to give more importance to source points that are likely to ressemble the target distribution. This approach aims to make the source distribution more similar to the target distribution by focusing on minimizing the target risk using the labeled source data.

Given that we typically do not have access to target labels, we need to assume that only the marginal distributions are different between the source and target domains. Hence, in the case of covariate shift (1.28) becomes:

$$R_{\mathcal{T}}(h) = \mathbb{E}_{\mathcal{P}_{\mathcal{S}}}\left[\frac{P_{\mathcal{T}}(\boldsymbol{x}, y)}{P_{\mathcal{S}}(\boldsymbol{x}, y)} L(y, h(\boldsymbol{x}))\right] = \mathbb{E}_{\mathcal{P}_{\mathcal{S}}}\left[\frac{P_{\mathcal{X},\mathcal{T}}(\boldsymbol{x})}{P_{\mathcal{X},\mathcal{S}}(\boldsymbol{x})} L(y, h(\boldsymbol{x}))\right] \tag{1.29}$$

Here, the density ratio $w(\boldsymbol{x}) = P_{\mathcal{X},\mathcal{T}}(\boldsymbol{x})/P_{\mathcal{X},\mathcal{S}}(\boldsymbol{x})$ represents the importance weight for the source data. Several methods have been developped to estimate this density ratio. For instance, Gaussian approximation [Shimodaira, 2000], kernel mean matching [Gretton et al., 2008] or Kullback-Leibler Importance Estimation Procedure (KLIEP) [Sugiyama et al., 2007] are commonly used approaches.

Feature based adaptation

Feature-based adaptation methods aim to extract invariant feature representations across domains. Typically, they create new feature representations by transforming original features to minimize the gap between the source and target features.

Subspace Approaches: These methods assume the existence of a common subspace where all domains are similar and preserve label information. They seek to estimate a projection to a common low-dimensional representation that minimizes the discrepancy between domains. Methods often involve projecting data from each domain into domain-specific lower-dimensional subspaces and then reducing the distance between these latent spaces to find a common representation. Techniques like PCA are used for dimensionality reduction, as seen in methods such as Transfer Component Analysis (TCA) [Pan et al., 2011], which aims to find a kernel subspace mapping that minimizes the maximum mean discrepancy (MMD) between source and target domains.

Mapping Approaches: These techniques directly transform original features into new representations that minimize the shift between their distributions. They operate under the assumption that there exists a mapping of source data such that $P_{\mathcal{S}}(m(\boldsymbol{x}), y) = P_{\mathcal{T}}(\boldsymbol{x}, y)$. Correlation Alignment (CORAL)[Sun et al., 2016] is a notable method that minimizes the difference in second-order statistics between source and target features. Optimal transport, a subset of mapping approaches, assumes the existence of a transport mapping T in feature space that preserves joint distributions, as demonstrated in methods like the optimal transport approach

of Courty et al. [2017], which learns a transport plan while keeping source points with similar labels close to each other during transport. Another example is the method proposed in [Gnassounou et al., 2023], which adapts signal power spectrum densities to a Wasserstein barycenter estimated from source data.

1.5 Objective of the thesis and contributions

1.5.1 Dataset shifts in M/EEG data

M/EEG data exhibit significant variability, which origniate from many factors including differences in recording devices, populations, and experimental tasks. These variations create systematic differences across datasets, that we earlier introduced as dataset shifts, which ultimately affect the robustness and generalizability of models trained on M/EEG data.

One major source of variability in M/EEG data arises from the use of different recording devices. For instance, MEG systems from different manufacturers, such as CTF and MEGIN, use distinct types of sensors. For EEG, helmets often have different electrode configurations. These differences are not limited to the types and positions of the sensors but extend to the number of electrodes used and the specifications of the amplifiers involved. Each of these factors introduces variations in how the brain's electrical activity is captured, creating challenges in integrating data from different devices. The dimensionality of the recorded data can also vary based on the number of sensors employed, further complicating harmonization across datasets. Another layer of variability stems from the populations involved in M/EEG studies. The selection criteria for study participants can differ significantly based on age, medical conditions, lifestyle ... Furthermore, differences in populations arise due to the study's geographical location, the country in which data collection took place, the particular hospital involved, and variations in recording conditions (e.g., noise levels in the recording environment). These population-based differences can lead to different brain activity patterns and create variability in the data that may not be easily accounted for by standard analysis techniques. Experimental tasks used during M/EEG recordings represent yet another source of variability. The nature of the task can influence the brain areas activated, the cognitive processes involved, and the specific brain rhythms elicited. Different tasks can lead to the engagement of distinct neural networks, resulting in varying patterns of brain activity that are difficult to compare across studies.

This inherent variability in M/EEG data can be illustrated by the HarMNqEEG dataset [Li et al., 2022], where differences in mean log powers across frequency bands are observed between recording sites and age groups as shown in Figure 1.8. These differences highlight the challenge of integrating data across studies with varying recording sites and populations, demonstrating the need for

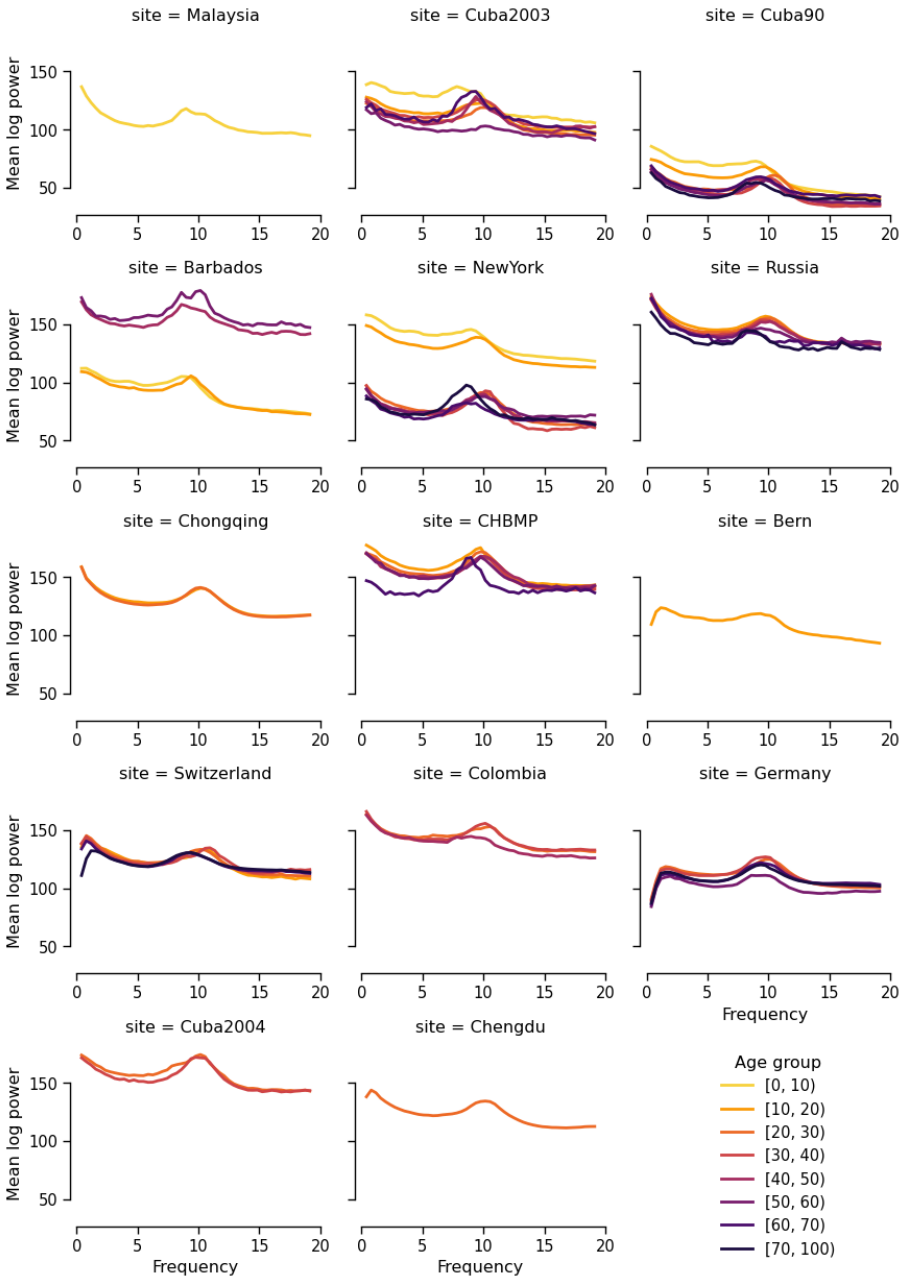

Figure 1.8: **Mean log powers per frequency of EEG recordings from the HarMNqEEG dataset [Li et al., 2022].** The log power spectrum are organized by recording sites in seperate panels, and by age groups (colors). The log powers of each participant were computed from all channels and then averaged across a same age group.

harmonization to mitigate the impact of these variations.

Due to these systematic differences, dataset shifts frequently occur in M/EEG research, making it difficult to build models that generalize across studies. To address this challenge, harmonization and domain adaptation techniques are required to make datasets more comparable and ensure that models trained on specific data can perform robustly across different domains.

Despite the clear need for harmonization and domain adaptation in M/EEG research, existing methods fall short in several key areas. First, many domain adaptation methods are designed for classification problems and are not applicable to regression tasks, which are commonly used in M/EEG studies. Second, these methods typically do not address joint shifts in both data and labels. In real-world scenarios, the relationship between brain activity (data) and the cognitive or behavioral outputs (labels) can vary across domains, and standard domain adaptation techniques do not account for these shifts. Finally, current approaches often fail to address the dimensionality variations introduced by different recording devices.

Overall, existing domain adaptation techniques are often tailored to specific situations and lack the generalizability required to handle the complex, real-life variability seen in M/EEG data. Therefore, more comprehensive approaches are needed to address the multiple sources of variability that occur in M/EEG datasets, allowing for the development of more robust and generalizable models.

1.5.2 Objectives

Dataset shift leads to poor generalization of models and is particularly prevalent in biological signals such as M/EEG data. This thesis addresses the issue of dataset shift within the framework of covariance-based predictive approaches, which have proven effective in predicting outcomes from M/EEG data, initially in the context of brain-computer interfaces [Barachant, Bonnet, Congedo, and Jutten, 2012, Congedo, Barachant, and Andreev, 2013] and more recently for brain-age prediction [Sabbagh et al., 2019, 2020, Engemann et al., 2022].

The primary objective of this thesis is to investigate dataset shift in M/EEG data. This includes understanding when and why dataset shift occurs by identifying different types of shifts and their causes. We examine how changes in environment, population, and recording devices can lead to shifts in data distribution. Specifically, we look both at covariate shifts, which involve changes in the input data distribution, and label shifts, which involve changes in the distribution of output labels, due to the above mentioned changes. The second part of this investigation is understanding the impact of these shifts on the performance of machine learning models. Through experimental studies based on simulated and real data, we assess how the presence of shifts degrades the model performances. In this part we demonstrate the necessity of mitigating these shifts to ensure robust predictive modeling and generalizability. Finally, to address the challenges posed

by dataset shifts in M/EEG data, we focus on implementing and evaluating various domain adaptation methods. To begin with, we evaluate existing domain adaptation methods that have been used in classification tasks but not in regression. We then attempt to improve these methods by developing novel approaches. By comparing model performance with and without domain adaptation techniques, we aim to highlight the improvements in robustness and generalizability.

Interpretability is a key aspect of machine learning models, particularly in medical applications where understanding the decision-making process of the model helps to understand the underlying mechanisms. In this thesis, in addition to assessing the impact of domain adaptation, we explore what was the impact of the models on data to compensate for shifts.

1.5.3 Organization of the thesis

The remaining chapters of the manuscript are organized as follows:

Chapter 2: Shift in data distribution

We consider and investigate dataset shifts in M/EEG recordings in which the label distributions are assumed to be the same across the different domains considered. We describe M/EEG data with a generative model and link possible causes of shifts in the data distribution to the parameters of this model. We evaluate existing domain adaptation methods, previously applied in classification contexts and often in supervised settings, to an unsupervised regression task. We evaluated these methods through controlled simulations by creating several shift scenarios. We also applied these methods to real M/EEG data, by first evaluating their performance when only the task performed by the subjects changes, and then when the population changes.

Chapter 3: Joint shift in data and label distributions

We extend the work presented in the previous chapter by considering joint shifts in M/EEG data and the labels we aim to predict. We investigate the impact of these shifts on model performance and evaluate the limitations of the methods proposed in Chapter 2. To address this issue, we develop a new method inspired by mixed-effects models, which involves a domain-specific re-centering operator and the joint learning of this operator alongside the regression model. We compare the performance of this new method with other approaches on a multi-national EEG dataset with recordings from different hospitals and different populations.

Chapter 4: Shift due to different recording devices

We investigate the challenges related to dimensionality variations caused by different recording devices. Different recording devices can mean different number

of electrodes with different positions on the scalp. Few studies have addressed this issue, and we propose to employ an interpolation method that leverages the inherent physics of the data to match the EEG electrodes configurations across datasets. We perform a comparative study with other methods on multiple BCI datasets in the context of motor imagery classification.

1.5.4 Publications and data analysis challenge

Publications linked to the manuscript

Apolline Mellot, Antoine Collas, Pedro L. C. Rodrigues, Denis Engemann, Alexandre Gramfort. "Harmonizing and aligning M/EEG datasets with covariance-based techniques to enhance predictive regression modeling." Imaging Neuroscience 2023; 1 1-23.
https://doi.org/10.1162/imag_a_00040.

Apolline Mellot, Antoine Collas, Sylvain Chevallier, Alexandre Gramfort, and Denis A. Engemann. "Geodesic Optimization for Predictive Shift Adaptation on EEG data." under review,
https://doi.org/10.48550/arXiv.2407.03878.

Apolline Mellot, Antoine Collas, Sylvain Chevallier, Denis Engemann, and Alexandre Gramfort. "Physics-informed and Unsupervised Riemannian Domain Adaptation for Machine Learning on Heterogeneous EEG Datasets." 32nd European Signal Processing Conference (EUSIPCO), Lyon, France, 2024,
https://doi.org/10.48550/arXiv.2403.15415.

Other publications

Denis A. Engemann, Apolline Mellot, Richard Höchenberger, Hubert Banville, David Sabbagh, Lukas Gemein, Tonio Ball, Alexandre Gramfort, "A reusable benchmark of brain-age prediction from M/EEG resting-state signals", NeuroImage, Volume 262, 2022, 119521, ISSN 1053-8119,
https://doi.org/10.1016/j.neuroimage.2022.119521.

Maria Sayu Yamamoto, Apolline Mellot, Sylvain Chevallier and Fabien Lotte, "Novel SPD Matrix Representations Considering Cross-Frequency Coupling for EEG Classification Using Riemannian Geometry", 31st European Signal Processing Conference (EUSIPCO), Helsinki, Finland, 2023, pp. 960-964,
10.23919/EUSIPCO58844.2023.10290043.

Data analysis challenge

Dementia screening challenge, Biomag 2022, Birmingham, England, 2022, Winning solution in MEG signal processing in order to screen dementia and mild cognitive impairment. Team: Apolline Mellot, Benoit Malezieux and Cedric Allain.

2 Shift in data distribution

Contents

As introduced in the first chapter, magneto- and electro- encephalography (M/EEG) are brain recording methods with a high temporal resolution on the order of milliseconds, offering a unique and non-invasive neuroscience method enabling basic research and clinical applications [Hari and Puce, 2023]. While quantitative approaches to analyzing M/EEG signals have historically focused on detecting statistical effects, the field has progressively embraced machine learning (ML) approaches whose success is evaluated through predictive modeling. In the context of brain health, classification models are widely used for various applications, e.g., for epileptic seizure detection [Tzallas et al., 2009], Brain Computer Interface (BCI) [Lotte et al., 2018], or automatic sleep staging [Chambon et al., 2018, Perslev et al., 2021]. Even though the regression context has been less explored in the literature, it has been shown to be successful for biomarker learning, e.g., focusing on brain age as an application [Al Zoubi et al., 2018, Sun et al., 2019, Engemann et al., 2020].

In this chapter, we focus on methods for regression modeling in the particular case of statistical discrepancies between datasets, for example, due to different populations, acquisition devices, or tasks performed during the recording. In other words, we aim to fit a regression model on one dataset and apply it on another. We explore domain adaptation techniques that aim to align the statistical distributions of the source and target datasets by leveraging the mathematical properties of the space of covariance matrices. In addition, we adapted these techniques to an unsupervised setting, where no labels are available in the target domain.

2.1 Covariance-based alignment applied to M/EEG regression modeling

Different approaches have been explored to predict cognitive-behavioral or biomedical outcomes from M/EEG data. Methods like Common Spatial Filtering (CSP) [Koles, 1991] or Source Power Comodulation (SPoC) [Dähne et al., 2014] build on top of supervised spatial filtering for dimensionality reduction and unmixing of overlapping, yet physiologically distinct, signal generators. Lately, deep learning based techniques have been the focus of interest as they can learn good feature representation directly from the raw signal, hence potentially simplifying processing pipelines [Roy et al., 2019, Schirrmeister et al., 2017]. Independently, an alternative approach has emerged from the BCI community which, like spatial filter methods, summarizes M/EEG data by covariance matrices. But instead of decomposing covariance matrices into filters, this approach uses mathematical tools motivated by the Riemannian geometry of the space of symmetric positive definite (SPD) matrices [Barachant et al., 2012, 2013] to define non-linear feature transformations that facilitate statistical learning with linear models. These techniques perform remarkably well given their simplicity [Congedo et al., 2013, Nguyen et al., 2017, Sabbagh et al., 2020] and are competitive with methods exploiting anatom-

ical information or end-to-end deep learning approaches [Engemann et al., 2022]. As the field of Riemannian geometry applied to M/EEG expand and consolidate, many opportunities remain unexplored. In this work, we focus on investigating the utility of the Riemannian framework for defining dataset-harmonizing transformations.

The recent emergence of large public databases and advances in ML have led to promising prediction models. Yet, these models can be sensitive to shifts in the data distribution and may perform poorly when applied to datasets from other clinical or research contexts [Quiñonero-Candela et al., 2022, Dockès et al., 2021]. This issue has also been referred to as batch effects in Li et al. [2022]. In this work, we focused on domain adaptation techniques that attempt to deal with these shifts. We aim for a predictive model to not only perform well on the data it has been trained on, the source domain, but also when applied to data from a distinct statistical distribution, the target domain. Many domain adaptation methods exist, ranging from simple approaches minimizing the difference between the second-order statistics of source and target domains [Sun et al., 2017], to more sophisticated models measuring the distance between deep representations of the source and target domains based on optimal transport [Damodaran et al., 2018]. In the context of brain data analysis, Canonical Correlation Analysis (CCA) and multiway CCA (MCCA) have been largely applied to find common brain activity and combine data across subjects when the same stimulus is presented to them [Lankinen et al., 2014, 2018, Dmochowski et al., 2018, de Cheveigné et al., 2019]. The experimental setups we are interested in do not meet these assumptions, for example, we consider recordings at rest and from different datasets with different subjects for which CCA or MCCA are not adapted. As we wish to work with M/EEG covariance matrices as basic signal representations for machine learning, we focus on techniques that explicitly use the geometry of SPD matrices to model the statistical distributions of distinct source and target datasets. One first approach proposed in the BCI context is to re-center the distributions to a common point of the SPD space [Zanini et al., 2018, Yair et al., 2019, Li et al., 2021]. To get a better alignment of the distributions, [Rodrigues et al., 2019, Maman et al., 2019, Bleuzé et al., 2021] propose to complement this re-centering step by adding a re-scale and a rotation correction. These covariance-based methods were initially developed to solve classification problems and are not necessarily applicable to regression without modification. In addition, most of them require labels in the target domain [Rodrigues et al., 2019, Bleuzé et al., 2021] for alignment. We focus on unsupervised alignment methods that can be readily used for regression modeling.

In this work, we develop a model-based approach for tackling dataset shifts in M/EEG data in which we consider re-centering, re-scaling, and rotation techniques from previous research on classification [Rodrigues et al., 2019, Maman et al., 2019, Bleuzé et al., 2021] to regression contexts, while assuming that no labeled data is available in the target domain. We build on top of the conceptual

framework from [Sabbagh et al., 2020], linking M/EEG-based regression to neural signal models, to investigate how dataset shifts can be expressed and handled with an appropriate generative model linking brain activity to both M/EEG measurements and biomedical outcomes. We elucidate how observed dataset shifts can be conceptually decomposed into differences in brain activity and differences in the relationship between the location and orientation of M/EEG signal generators relative to the recording device that reflects the device type, body posture, and individual brain anatomy. With this approach, we establish the connection between particular alignment steps and the parameters of the generative model as well as the physiological and physical shifts they are meant to compensate for. Using statistical simulations, based on the generative model, we then explore different dataset-shift scenarios and investigate the effectiveness of data alignment techniques — combined and in isolation. Through empirical benchmarks on the Cam-CAN MEG dataset (N=646) and two EEG datasets (TUAB-normal, N=1385; LEMON, N=213), we evaluate the practical impact of these alignment techniques for boosting the generalization capacity of regression models across acquisition protocols (resting state vs. audiovisual & motor tasks) and cohorts (clinical EEG versus research & laboratory-grade EEG). We focus on brain age as it is a label easy to collect and valuable as a surrogate biomarker.

The rest of the chapter is organized as follows. In Section 2.3, we extend the generative model from [Sabbagh et al., 2020] to express and decompose dataset shifts into distinct factors, which motivates the three steps we use to compensate for dataset shifts: re-centering, re-scaling, and rotation correction. In Section 2.4 and Section 2.5, we assess the robustness of these alignment steps using simulations and real-world M/EEG data.

2.2 Contribution

The content of this chapter is based on the work published in the following article:

Apolline Mellot, Antoine Collas, Pedro L. C. Rodrigues, Denis Engemann, Alexandre Gramfort. “Harmonizing and aligning M/EEG datasets with covariance-based techniques to enhance predictive regression modeling.” Imaging Neuroscience 2023; 1 1-23.

In this work, our first contribution is to study and relate physiological and physical changes in the brain and the recording setup to the statistical shifts observed in the M/EEG data. The harmonization procedure of Rodrigues et al. [2019] was developed for supervised classification tasks and required labeled data in the target domain, but has not been evaluated on unsupervised regression tasks. Our second

contribution was to adapt and benchmark this approach for unsupervised regression tasks and, in this way, measure the dataset shift when changing of recording task or population.

The scripts and code for the alignment methods and the results presented in this chapter are publicly available on GitHub:

https://github.com/apmellot/harmonizing_aligning_meeg.

2.3 Regression modeling M/EEG covariance matrices

To describe dataset shifts that can occur with M/EEG signals, we extend the generative model of M/EEG regression tasks from Sabbagh et al. [2020] where the prediction outcome is continuous. A canonical example that we will use in this work is brain age prediction [Xifra-Porxas et al., 2021, Engemann et al., 2022]. This model has also been applied to event-level regression of muscular activity with electromyogram and MEG recordings [Sabbagh et al., 2020]. We describe and discuss the parameters of the generative model to understand which mechanisms can explain dataset shifts. Finally, we present various alignment strategies aiming to draw a geometrical analysis of the possible shifts and compensate for them.

2.3.1 Statistical generative model of M/EEG signals

Generative model M/EEG signals $\boldsymbol{x}(t) \in \mathbb{R}^P$ are multivariate time series recorded with P sensors at (or above) the surface of the scalp, and that capture the electrical or magnetic activity generated by large-scale neural synchrony. These neurophysiological generators are not directly observable, and here we focus on the situation in which we do not have access to information about the individual brain anatomy, e.g., when MRI scans are not available. Thus, we use a statistical approach inspired by blind source separation to approximate the signal's generative mechanism. We model the M/EEG signals as a linear combination of statistical brain generators corrupted by some additive noise. In this work, we consider a dataset with N observations $\boldsymbol{X} = \{\boldsymbol{X}_i, i = 1 \dots N\}$ for which one observation corresponds to one participant. One observation $\boldsymbol{x}_i(t) \in \mathbb{R}^P$ is written as:

$$\boldsymbol{x}_i(t) = \boldsymbol{A}'\boldsymbol{s}_i(t) + \boldsymbol{A}''\boldsymbol{n}_i(t) \ , \tag{2.1}$$

where $\boldsymbol{s}_i(t) \in \mathbb{R}^Q$ is the underlying activity generating this observation with $Q \leq P$, and $\boldsymbol{n}_i(t) \in \mathbb{R}^{(P-Q)} \sim \mathcal{N}(0, \sigma_n^2 \mathbf{I}_{P-Q})$ causes a contamination due to noise. We denote $\boldsymbol{A}' = [\boldsymbol{a}_1, \dots, \boldsymbol{a}_Q] \in \mathbb{R}^{P \times Q}$ the mixing matrix whose columns are the spatial patterns of the neural generators, and $\boldsymbol{A}'' = [\boldsymbol{a}_{Q+1}, \dots, \boldsymbol{a}_P] \in \mathbb{R}^{P \times (P-Q)}$ the matrix of the spatial noise patterns. Note that in this model, the noise is not considered independent across sensors but spatially correlated, as is typically the case with environmental or physiological artifacts present in M/EEG data.

This model can be rewritten by combining the generator patterns and the noise in a single invertible matrix $\boldsymbol{A} = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_Q, \boldsymbol{a}_{Q+1}, \ldots, \boldsymbol{a}_P] \in \mathbb{R}^{P \times P}$ which is the concatenation of $\boldsymbol{A}'$ and $\boldsymbol{A}''$. The model is thus given by

$$\boldsymbol{x}_i(t) = \boldsymbol{A}\boldsymbol{\eta}_i(t) \ , \tag{2.2}$$

where $\boldsymbol{\eta}_i(t) \in \mathbb{R}^P$ denotes the concatenation of $\boldsymbol{s}_i(t)$ and $\boldsymbol{n}_i(t)$. In this model, we assume $\boldsymbol{A}$ to not depend on i nor t. It is also assumed that the statistical generators $\boldsymbol{s}_i(t) = \{\boldsymbol{s}_{i,j}(t), j = 1 \ldots Q\}$ are zero-mean, uncorrelated, and independent from the noise. In other words, we assume that the noise generated by artifacts is completely independent of brain activity. In the following, j will denote the generator's index.

We now consider the covariances $\boldsymbol{\Sigma}_i$ of M/EEG signals $\boldsymbol{X}_i \in \mathbb{R}^{P \times T}$ with T the number of time samples as defined in (1.23):

$$\boldsymbol{\Sigma}_i = \frac{1}{T}\boldsymbol{X}_i\boldsymbol{X}_i^\top \in \mathbb{R}^{P \times P} \ . \tag{2.3}$$

The covariance of M/EEG signals holds the sensors' variance on its diagonal. In our statistical model and with the previous assumptions, the covariance of the statistical generators is a diagonal matrix whose elements are the variances of each generator $\mathbb{E}_t\left[\boldsymbol{s}_i(t)\boldsymbol{s}_i(t)^\top\right] = \mathrm{diag}(\boldsymbol{p}_i)$ with $\boldsymbol{p}_i \in \mathbb{R}^Q$, also referred to below as "powers". Thus, we can conveniently summarize the M/EEG covariances as follows:

$$\boldsymbol{\Sigma}_i = \boldsymbol{A}\boldsymbol{H}_i\boldsymbol{A}^\top \tag{2.4}$$

where $\boldsymbol{H}_i = \mathbb{E}_t\left[\boldsymbol{\eta}_i(t)\boldsymbol{\eta}_i(t)^\top\right] \in \mathbb{R}^{P \times P}$ is a block matrix of $\mathrm{diag}(\boldsymbol{p}_i)$ on the upper $Q \times Q$ part, and the noise covariance is in the lower $(P-Q) \times (P-Q)$ block. We here assume that $\mathbb{E}_t\left[\boldsymbol{s}_i(t)\boldsymbol{n}_i(t)^\top\right] = 0$, meaning that the matrix $\boldsymbol{H}_i$ is block diagonal.

For regression modeling from M/EEG, it is natural to model the outcome y_i as a linear combination of a function of the generators' power $p_{i,j} = \mathbb{E}_t[s_{i,j}^2(t)] \in \mathbb{R}$:

$$y_i = \beta_0 + \sum_{j=1}^{Q} \beta_j f(p_{i,j}) + \epsilon_i \ , \tag{2.5}$$

where β_j are regression coefficients, f is a known function, and $\epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2)$ is an additive random perturbation. For example, ageing (y) could impact brain activity in distinct brain networks $(\boldsymbol{s})$ to different extents $(\beta_1, \ldots, \beta_Q)$. This could lead for example to a log-decay or log-increase of brain activity per year, hence, motivating a logarithmic function $f = \log$, which is a wide-spread function describing the scaling of various facets of brain structure and function [Buzsáki and Mizuseki, 2014] including neural firing rates, axonal diameters, synaptic weights, and, importantly power and frequency scaling. Replacing the generator power

with the empirical average of the squared generators, the model is given by:

$$y_i = \beta_0 + \sum_{j=1}^{Q} \beta_j \log\left(\frac{1}{T}\sum_{t=1}^{T} s_{i,j}^2(t)\right) + \epsilon_i \ . \tag{2.6}$$

Model violations The assumption that $\boldsymbol{A}$ does not depend on the observation (subject) is not valid when working with actual M/EEG data. Each subject has a different head morphology, which results in slight variations in their respective mixing matrices: $\boldsymbol{A}_i = \boldsymbol{A} + \boldsymbol{E}_i$ with $\boldsymbol{E}_i \in \mathbb{R}^{P\times P}$. When subscript i is omitted below, $\boldsymbol{A}$ represents the average head morphology of the subjects. In our simulations below, we will assume that each element of $\boldsymbol{E}_i$ is drawn from $\mathcal{N}(0, \sigma_A^2)$.

2.3.2 Regression model

The approach we focus on in this work involves learning linear models from covariance matrices [Barachant et al., 2012, 2013], which is introduced in Section 1.3. Sabbagh et al. [2019, 2020] show that this Riemann-based model is robust to different preprocessing choices and to model violation. This model also stands out in terms of performance when applied for regression tasks to M/EEG data in various settings.

In this framework, we remind that the covariances $\boldsymbol{\Sigma}_i$ are used as input of the model. In Section 1.3.2 we defined an operator to transform the covariance matrices into feature vectors with (1.15). Thus, we follow this approach and the covariance matrices are vectorized:

$$\boldsymbol{z}_i = \phi(\boldsymbol{\Sigma}_i, \overline{\boldsymbol{\Sigma}}) = \text{uvec}(\log(\overline{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\boldsymbol{\Sigma}_i\overline{\boldsymbol{\Sigma}}^{-\frac{1}{2}})) \in \mathbb{R}^{P(P+1)/2} \ , \tag{2.7}$$

We denote the variables $\phi(\boldsymbol{\Sigma}_i, \overline{\boldsymbol{\Sigma}})$ by $\boldsymbol{z}_i$ in the following for simpler notations. $\boldsymbol{z}_i$ are called tangent vectors. Since these tangent vectors are elements of an Euclidean space, we can use them as input on classical machine learning models.

2.4 Dataset shift in M/EEG data and how to deal with it

2.4.1 Possible data shifts

Each parameter of the model described in equations (2.2) and (2.6) can vary for different reasons. We are interested in fitting a regression algorithm to a source dataset $\boldsymbol{X}_{\mathcal{S}} = \{\boldsymbol{x}_{\mathcal{S},i}, i = 1, \ldots, N_{\mathcal{S}}\}$ to later predict outcomes on a distinct target dataset $\boldsymbol{X}_{\mathcal{T}} = \{\boldsymbol{x}_{\mathcal{T},i}, i = 1, \ldots, N_{\mathcal{T}}\}$ both recorded with P sensors at the same locations. The datasets are not necessarily composed of the same number of subjects.

Applying our previous notations, we can describe the source dataset as follows:

$$\begin{cases} \boldsymbol{x}_{\mathcal{S},i}(t) = \boldsymbol{A'}_{\mathcal{S},i}\ \boldsymbol{s}_{\mathcal{S},i}(t) + \boldsymbol{A''}_{\mathcal{S},i}\boldsymbol{n}_{\mathcal{S},i}(t) = \boldsymbol{A}_{\mathcal{S},i}\boldsymbol{\eta}_{\mathcal{S},i}(t) \\ y_{\mathcal{S},i} = \beta_0 + \sum_{j=1}^{Q} \beta_j \log\left(\frac{1}{T}\sum_{t=1}^{T}(s_{\mathcal{S},i,j}(t))^2\right) + \epsilon_{\mathcal{S},i} \end{cases} \tag{2.8}$$

where $\boldsymbol{A}_{\mathcal{S},i} = \boldsymbol{A}_{\mathcal{S}} + \boldsymbol{E}_{\mathcal{S},i}$ wtih entries of $\boldsymbol{E}_{\mathcal{S},i}$ are i.i.d $\mathcal{N}\left(0, (\sigma_{\mathcal{S},A})^2\right)$, $\boldsymbol{n}_{\mathcal{S},i} \sim \mathcal{N}\left(0, (\sigma_{\mathcal{S},n})^2\, \boldsymbol{I}_{P-Q}\right)$ and $\epsilon_{\mathcal{S},i} \sim \mathcal{N}\left(0, (\sigma_{\mathcal{S},\epsilon})^2\right)$. We remind that the statistical generator powers are defined as $\mathrm{diag}(\boldsymbol{p}_{\mathcal{S},i}) = \mathbb{E}_t\left[\boldsymbol{s}_{\mathcal{S},i}(t)\boldsymbol{s}_{\mathcal{S},i}^{\top}(t)\right]$. The same equations can be written for the target dataset by replacing the exponent $\mathcal{S}$ by $\mathcal{T}$. We now list physical reasons that could induce differences between source and target datasets and link them to parameter changes of the corresponding generative models (2.8).

1. If we consider two different populations, the head morphology may vary, and the subject-averaged mixing matrices $\boldsymbol{A}_{\mathcal{S}}$ and $\boldsymbol{A}_{\mathcal{T}}$ would differ.

2. Having different populations in both datasets would also imply that they will not have the same mixing matrices distribution: $\sigma_{\mathcal{S},\boldsymbol{A}} \neq \sigma_{\mathcal{T},\boldsymbol{A}}$.

3. When data are recorded with different devices, the recording conditions and noise might not be the same, resulting in different signal-to-noise ratio (SNR): $\sigma_{\mathcal{S},\boldsymbol{n}} \neq \sigma_{\mathcal{T},\boldsymbol{n}}$.

4. Clinical outcomes e.g., neuropsychological testing scores can be noisy. This noise could differ from one dataset to another: $\sigma_{\mathcal{S},\epsilon} \neq \sigma_{\mathcal{T},\epsilon}$.

Because of all those possible causes of variability in the model parameters, machine learning approaches may fail to provide good predictions across datasets. In this work, we focus on shifts that only affect the data, and we assume that the regression coefficients β_j are the same for source and target. In particular, we are interested in understanding changes related to the mixing matrix and the variance of the statistical generators. The variability of these parameters across subjects and datasets affects the observed signals and results in variability in the data distribution. Below, we discuss which statistical methods could help reduce these different shifts between the data distributions of two different datasets.

2.4.2 Alignment methods

We aim to learn a regression model from one dataset, the source domain, that will perform well on another, the target domain. As we focus on shifts affecting the data distribution, we investigate domain adaptation methods that align the source and the target distributions using geometrical transformations. The methods we chose for understanding and reducing dataset shifts are articulated in three alignment steps: re-centering, equalizing dispersion, and rotation correction.

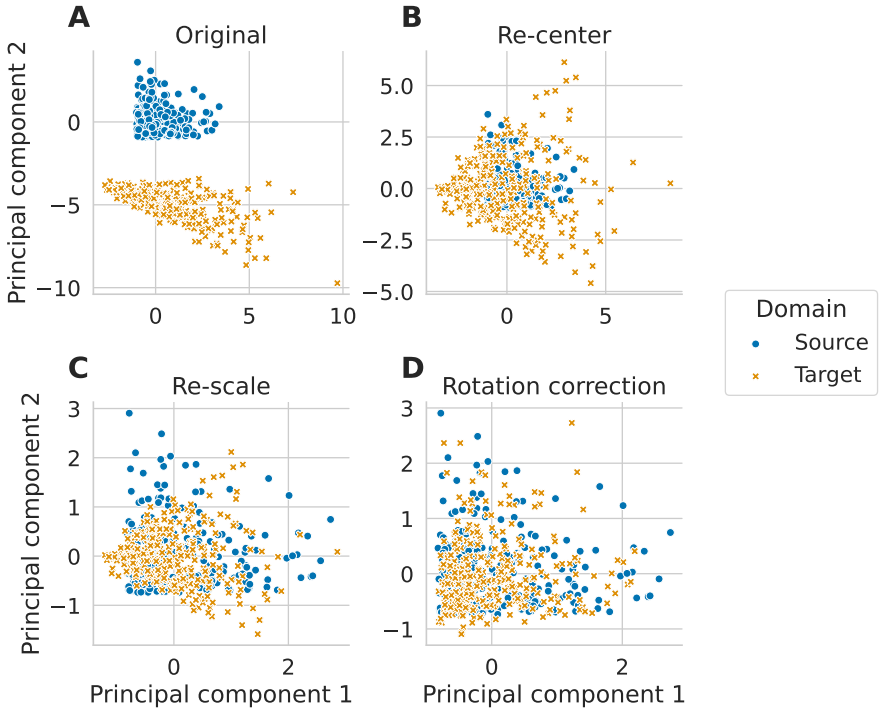

Figure 2.1: **Alignment steps illustrated on simulated data.** The three alignment steps are applied to data simulated following the generative model, as detailed in Section 2.5.1. We set the size of the matrices to $P = 2$ and generated $N = 300$ matrices in each domain. Each new step is applied on top of the previous one. The plots correspond to the two first principal components of the tangent vectors. (**A**) The simulated data are plotted on the tangent space before any alignment steps. (**B**) The original simulated data are centered to a common point, (**C**), then their distributions are equalized, and (**D**) finally, a rotation correction is applied.

This choice was inspired by transfer learning methods used in Brain-Computer Interfaces (BCI) application and, more specifically, by the Riemannian Procrustes Analysis (RPA) of [Rodrigues et al., 2019]. These steps can be used independently, usually by only re-centering the data, or combined. In the following, we detail these alignment functions in a general manner.

Step 1: re-centering The most commonly used method of transfer learning on symmetric positive definite (SPD) matrices is to re-center each dataset in a common reference point on the Riemannian manifold [Zanini et al., 2018, Li et al., 2021]. This reference can be chosen as one of the domains' geometric mean or an arbitrary point on the manifold. Here we propose to re-center each domain to the

Identity by whitening them with their respective geometric mean $\overline{\mathbf{\Sigma}}_{\mathcal{S}}$ and $\overline{\mathbf{\Sigma}}_{\mathcal{T}}$:

$$h^{(\text{rct})}_{\overline{\mathbf{\Sigma}}_{\mathcal{S}}}\left(\mathbf{\Sigma}_{\mathcal{S},i}\right) = \overline{\mathbf{\Sigma}}_{\mathcal{S}}^{-\frac{1}{2}}\mathbf{\Sigma}_{\mathcal{S},i}\overline{\mathbf{\Sigma}}_{\mathcal{S}}^{-\frac{1}{2}} \tag{2.9}$$

$$h^{(\text{rct})}_{\overline{\mathbf{\Sigma}}_{\mathcal{T}}}\left(\mathbf{\Sigma}_{\mathcal{T},i}\right) = \overline{\mathbf{\Sigma}}_{\mathcal{T}}^{-\frac{1}{2}}\mathbf{\Sigma}_{\mathcal{T},i}\overline{\mathbf{\Sigma}}_{\mathcal{T}}^{-\frac{1}{2}} \quad . \tag{2.10}$$

Differently put, re-centering applies separate whitening for source versus target data. This helps avoid errors in the tangent space projection when the average covariance is different for the source and target, e.g., because the mixing matrices are different, as in Figure 2.1 (**B**). This is a Riemannian equivalent of the centering step in classical z-scoring.

Step 2: equalizing the dispersion In this second step, the idea is to re-scale the covariances distribution around their mean $\overline{\mathbf{\Sigma}}$ as illustrated in Figure 2.1(**C**). We first compute the mean dispersion d of the covariances as the sum of the square distance between each matrix of the set and their geometric mean $\overline{\mathbf{\Sigma}}$ over the number of samples in the dataset:

$$d_{\mathcal{S}} = \frac{1}{N}\sum_{i=1}^{N}\delta_R^2\left(\mathbf{\Sigma}_{\mathcal{S},i}, \overline{\mathbf{\Sigma}}_{\mathcal{S}}\right) \tag{2.11}$$

$$d_{\mathcal{T}} = \frac{1}{N}\sum_{i=1}^{N}\delta_R^2\left(\mathbf{\Sigma}_{\mathcal{T},i}, \overline{\mathbf{\Sigma}}_{\mathcal{T}}\right) \quad , \tag{2.12}$$

with δ_R the Riemannian distance defined in (1.8). Then, we re-scale all covariances with $\sqrt{\frac{1}{d}}$ so that the distribution has a dispersion of 1:

$$h^{(\text{str})}_{d_{\mathcal{S}}}\left(\mathbf{\Sigma}_{\mathcal{S},i}\right) = \mathbf{\Sigma}_{\mathcal{S},i}^{\sqrt{1/d_{\mathcal{S}}}} \tag{2.13}$$

$$h^{(\text{str})}_{d_{\mathcal{T}}}\left(\mathbf{\Sigma}_{\mathcal{T},i}\right) = \mathbf{\Sigma}_{\mathcal{T},i}^{\sqrt{1/d_{\mathcal{T}}}} \quad . \tag{2.14}$$

This is a Riemannian equivalent of the re-scaling step in classical z-scoring. By analogy, univariate rescaling of two groups modifies the data so that, e.g. a t-test would find its assumptions of equal variances met while not detecting any difference of means between the two datasets.

Step 3: rotation correction Until now, we have applied correction measures that process source and target data independently. This is not the case in this third step which implies shared information between source and target datasets. The rotation correction is the most delicate of the three steps. It requires estimating many more parameters than the others, and the source and target feature spaces must be the same size ($P_{\mathcal{S}} = P_{\mathcal{T}} = P$). In the literature, several methods for rotation estimation exist. In the following, we detail two of the methods we selected:

1. The first rotation correction we implemented is inspired by [Maman et al., 2019]. The covariances are first vectorized by mapping them in the tangent space (at identity after re-centering). Then we compute the Singular Value Decomposition (SVD) of these tangent vectors $\boldsymbol{z}_i \in \mathbb{R}^{P(P+1)/2}$, $\boldsymbol{Z}_\mathcal{S} = \{\boldsymbol{z}_{\mathcal{S},i}, i = 1 \dots N_\mathcal{S}\}$ and $\boldsymbol{Z}_\mathcal{T} = \{\boldsymbol{z}_{\mathcal{T},i}, i = 1 \dots N_\mathcal{T}\}$:

$$\boldsymbol{Z}_\mathcal{S} = \boldsymbol{U}_\mathcal{S}^\top \boldsymbol{S}_\mathcal{S} \boldsymbol{V}_\mathcal{S} \tag{2.15}$$

$$\boldsymbol{Z}_\mathcal{T} = \boldsymbol{U}_\mathcal{T}^\top \boldsymbol{S}_\mathcal{T} \boldsymbol{V}_\mathcal{T} \ . \tag{2.16}$$

The columns of the $\boldsymbol{U} \in \mathbb{R}^{P(P+1)/2 \times P(P+1)/2}$ matrices are the left singular vectors ordered from largest to smallest singular values. The SVD is done separately on source and target distributions so the resulting singular vectors will unlikely have the same direction. As we desire for corresponding singular vectors between source and target to have an acute angle, we reorient them with a sign correction applied to the columns of $\boldsymbol{U}_\mathcal{T}$:

$$\boldsymbol{u}_{\mathcal{T},j} = \operatorname{sign}\left(\boldsymbol{u}_{\mathcal{S},j} \boldsymbol{u}_{\mathcal{T},j}\right) \boldsymbol{u}_{\mathcal{T},j}, \quad \forall j \tag{2.17}$$

Finally, the $\boldsymbol{U}$ matrices are used for rotation correction:

$$h^{\text{rot}}_{\boldsymbol{U}_\mathcal{S}}\left(\boldsymbol{Z}_\mathcal{S}\right) = \boldsymbol{U}_\mathcal{S}^\top \boldsymbol{Z}_\mathcal{S} \tag{2.18}$$

$$h^{\text{rot}}_{\boldsymbol{U}_\mathcal{T}}\left(\boldsymbol{Z}_\mathcal{T}\right) = \boldsymbol{U}_\mathcal{T}^\top \boldsymbol{Z}_\mathcal{T} \ . \tag{2.19}$$

We will refer to this rotation correction method as *unpaired*.

2. The second way to estimate the rotation that we used is inspired by [Bleuzé et al., 2021]. In this paper, they consider a classification question and propose to correct the rotation between source and target distributions by matching their respective classes' mean. This is done by solving the Procrustes problem

$$\underset{\boldsymbol{R} \in \mathcal{O}(N_\mathcal{T})}{\arg\min} \left\| \boldsymbol{R} \bar{\boldsymbol{Z}}_\mathcal{T} - \bar{\boldsymbol{Z}}_\mathcal{S} \right\|_F \ , \tag{2.20}$$

where $\mathcal{O}(N_\mathcal{T})$ is the orthogonal group, $\bar{\boldsymbol{Z}}_\mathcal{S}$ the concatenation of the classes' mean tangent vector from the source domain, and similarly for $\bar{\boldsymbol{Z}}_\mathcal{T}$ with the available labeled data of the target domain. Then, to correct the rotation, the target tangent vectors are transformed using the solution of the Procrustes problem

$$h^{\text{rot}}_{\boldsymbol{R}}\left(\bar{\boldsymbol{Z}}_\mathcal{T}\right) = \boldsymbol{R} \bar{\boldsymbol{Z}}_\mathcal{T} \ . \tag{2.21}$$

As we wish to be in a regression context without access to target labels, we modified this method by solving the Procrustes problem on all the tangent vectors

$$\underset{\boldsymbol{R} \in \mathcal{O}(N_\mathcal{T})}{\arg\min} \left\| \boldsymbol{R} \boldsymbol{Z}_\mathcal{T} - \boldsymbol{Z}_\mathcal{S} \right\|_F \ . \tag{2.22}$$

In practice, this solution is found by computing the SVD of the product of the source and target tangent vectors

$$\boldsymbol{Z}_{\mathcal{T}}\boldsymbol{Z}_{\mathcal{S}}^{\top} = \boldsymbol{U}^{\top}\boldsymbol{S}\boldsymbol{V} \quad , \tag{2.23}$$

and is $\boldsymbol{R} = \boldsymbol{V}\boldsymbol{U}^{\top}$. In this step, we include the information on which source point should be matched to which target point. It means the source and the target dataset should have the same number of observations ($N_{\mathcal{S}} = N_{\mathcal{T}} = N$) and be composed of "matching" observations (for example, the same set of subjects but different tasks/recording conditions/devices). We will refer to this method as *paired*.

This step allows us to align source and target distributions in a shared space. The rotation correction is helpful when the mixing matrices are different between the domains (Figure 2.1 (**D**)).

2.5 Numerical evaluation: simulated data

2.5.1 Data simulation

In the simulation study, we generated simulated data with the generative model presented in Section 2.3.1. We set the dimension of the matrices to $P = 20$ to have a matrix size coherent with real data, and the number of statistical generators to $Q = P$, in other words, we considered signals without noise. Mixing matrices $\boldsymbol{A}$ were generated as Gaussian random matrices in $\mathbb{R}^{P\times P}$ from $\mathcal{N}(0, 1)$. Instead of generating signals $\boldsymbol{s}$, we directly computed their powers $\boldsymbol{p}$ as random numbers from a uniform distribution in $[0, 1)$. The same powers were used for both the source and the target sets. We then constructed the covariance matrices $\boldsymbol{\Sigma}_{\mathcal{S}} = \{\boldsymbol{\Sigma}_{\mathcal{S},i}, i = 1, \ldots, N\}$ and $\boldsymbol{\Sigma}_{\mathcal{T}} = \{\boldsymbol{\Sigma}_{\mathcal{T},i}, i = 1, \ldots, N\}$, and the outcome to predict as in equations (2.4) (with $\boldsymbol{H}_i = \mathrm{diag}(\boldsymbol{p}_i)$ because $P = Q$) and (2.6) (with $\epsilon_i = 0, \forall i$). We designed several shift scenarios by altering either the mixing matrices or the powers in order to evaluate the alignment methods. In practice, the shifts were created by first generating the source data and then building the target data as a modified version of the source data. By doing so, one point of the source set corresponds to one point of the target set. This way, it is possible to evaluate the paired rotation correction. More details about the shift scenarios are presented in the following paragraphs.

2.5.2 Simulation scenarios

We detail the changes we introduced for each scenario between source and target distributions. As stated in Section 2.4.1, we focused on modeling shifts involving changes in the mixing matrix or the variance of the statistical generators. The parameters that are not mentioned were the same for source and target. For the first three scenarios, we aimed to find transformations/shifts between source

and target to which each alignment step is robust. Table 2.1 summarizes the scenarios and the associated parameter changes in source and target.

Translation In this scenario, we wish to assess how robust the alignment methods are when the source and target mixing matrices are different. Specifically, we built the target mixing matrix as $\boldsymbol{A}_{\mathcal{T}} = \boldsymbol{B}^{\alpha}\boldsymbol{A}_{\mathcal{S}}$ with $\boldsymbol{B} \in \mathbb{S}_P^{++}$ and $\alpha \in \mathbb{R}_*^+$. Here, one considered that the mixing matrix perturbation is done by an SPD matrix to decompose the case of $\boldsymbol{A}_{\mathcal{T}} \neq \boldsymbol{A}_{\mathcal{S}}$ into translation and rotation effects. The benchmark extends the previous simulations from [Sabbagh et al., 2020]. More explanations are provided about this decomposition in the **translation and rotation** paragraph. The parameter α controls the strength of the perturbation and thus how $\boldsymbol{A}_{\mathcal{T}}$ is different from $\boldsymbol{A}_{\mathcal{S}}$ (if $\alpha = 0$, $\boldsymbol{A}_{\mathcal{T}} = \boldsymbol{A}_{\mathcal{S}}$).

Scale We wanted to create a scenario in which the source and target distributions have different dispersions. In this scenario, we constructed the target covariances with an exponent on the powers: $\boldsymbol{p}_{\mathcal{T},i} = (\boldsymbol{p}_{\mathcal{S},i})^{\sigma_p}$ with $\sigma_p > 0$. The parameter σ_p controls how different the dispersions are. This modification was only applied to the data, so the outcome values y were unchanged.

Translation and rotation For this scenario, we built the source and target data from completely different mixing matrices and thus generalized the **translation** scenario. To evaluate how alignment methods performed for a growing difference between the source and the target mixing matrices, we defined a parameter m such as $\boldsymbol{A}_{\mathcal{T}} = m\boldsymbol{A}_t + (1-m)\boldsymbol{A}_{\mathcal{S}}$. $\boldsymbol{A}_t$ was fixed and generated as a random matrix in

	Source	**Target**
Translation	$\boldsymbol{A}_{\mathcal{S}}$ fixed	$\boldsymbol{A}_{\mathcal{T}} = (\boldsymbol{B})^{\alpha}\boldsymbol{A}_{\mathcal{S}}$ with $\alpha > 0$
Scale	$\boldsymbol{p}_{\mathcal{S},i}$ fixed	$\boldsymbol{p}_{\mathcal{T},i} = (\boldsymbol{p}_{\mathcal{S},i})^{\sigma_p}$ with $\sigma_p > 0$
Translation and rotation	$\boldsymbol{A}_{\mathcal{S}}$ fixed, $\boldsymbol{A}_t \neq \boldsymbol{A}_{\mathcal{S}}$ fixed	$\boldsymbol{A}_{\mathcal{T}} = m\boldsymbol{A}_t + (1-m)\boldsymbol{A}_{\mathcal{S}}$ with $m \in [0,1]$
Noise on mixing matrix	$\boldsymbol{A}_{\mathcal{S},i} = \boldsymbol{A}_{\mathcal{S}} + \boldsymbol{E}_{\mathcal{S},i}$ $\sigma_{\mathcal{S},A} = 10^{-2}$ fixed	$\boldsymbol{A}_{\mathcal{T},i} = \boldsymbol{A}_{\mathcal{T}} + \boldsymbol{E}_{\mathcal{T},i}$ $\sigma_{\mathcal{T},A} > 0$

Table 2.1: **Summary of the simulation scenarios.**

$\mathbb{R}^{P\times P}$ from $\mathcal{N}(0,1)$. In this manner, we created an interpolation between $\boldsymbol{A}_t$ and $\boldsymbol{A}_\mathcal{S}$ to generate $\boldsymbol{A}_\mathcal{T}$: if $m=0$, $\boldsymbol{A}_\mathcal{T}=\boldsymbol{A}_\mathcal{S}$ and if $m=1$, $\boldsymbol{A}_\mathcal{T}=\boldsymbol{A}_t$.

In this scenario, $\boldsymbol{A}_\mathcal{T}\neq\boldsymbol{A}_\mathcal{S}$ but we constructed the source and target covariances with the same $\boldsymbol{H}_i$ matrices following equation (2.4). Thus we can write:

$$\boldsymbol{H}_i=[\boldsymbol{A}_\mathcal{S}^\top]^{-1}\boldsymbol{C}_{\mathcal{S},i}[\boldsymbol{A}_\mathcal{S}]^{-1} \tag{2.24}$$

We can replace this $\boldsymbol{H}_i$ expression in the target covariances to get:

$$\begin{aligned}\boldsymbol{\Sigma}_{\mathcal{T},i}&=\boldsymbol{A}_\mathcal{T}\boldsymbol{H}_i\boldsymbol{A}_\mathcal{T}^\top &(2.25)\\ &=\boldsymbol{A}_\mathcal{T}[\boldsymbol{A}_\mathcal{S}^\top]^{-1}\boldsymbol{\Sigma}_{\mathcal{S},i}[\boldsymbol{A}_\mathcal{S}]^{-1}\boldsymbol{A}_\mathcal{T}^\top &(2.26)\\ &=\boldsymbol{D}\boldsymbol{\Sigma}_{\mathcal{S},i}\boldsymbol{D}^\top &(2.27)\end{aligned}$$

with $\boldsymbol{D}=\boldsymbol{A}_\mathcal{T}[\boldsymbol{A}_\mathcal{S}^\top]^{-1}$. The target covariance matrices correspond to the source covariance matrices transformed with the square matrix $\boldsymbol{D}$. A square matrix can be interpreted as a linear transformation: such a matrix can be decomposed into the product of an orthogonal matrix with a positive semi-definite Hermitian matrix (polar decomposition, a.k.a. QR factorization). Thus, we can interpret this scenario as the **translation** scenario (SPD matrix of the polar decomposition) with an additional perturbation by an orthogonal matrix.

Noise on mixing matrix We finally introduced individual noise in the mixing matrix to get a more realistic scenario: $\boldsymbol{A}_{\mathcal{S},i}=\boldsymbol{A}_\mathcal{S}+\boldsymbol{E}_{\mathcal{S},i}$ assuming that entries of $\boldsymbol{E}_{\mathcal{S},i}$ are drawn from $\mathcal{N}\left(0,(\sigma_{\mathcal{S},A})^2\right)$ (and similarly for the target mixing matrices). The source data were generated with a fixed noise value $\sigma_{\mathcal{S},A}=10^{-2}$, and the tested $\sigma_{\mathcal{T},A}$ values varied from 10^{-3} to 1. Here, the mean mixing matrices $\boldsymbol{A}_\mathcal{S}$ and $\boldsymbol{A}_\mathcal{T}$ were the same. This scenario was inspired by the simulation study of [Sabbagh et al., 2020] in which the same level of noise on the mixing matrix was added in the train and the test sets. Here, we explored the situation in which the noise levels in the train (source) and test (target) mixing matrices were different.

2.5.3 Alignment, vectorization, and regression

Once the covariance matrices were generated according to a given scenario, data of both domains were aligned with the methods detailed in Section 2.4.2. Then, we vectorized the matrices in the tangent space as in (2.7) with $\overline{\boldsymbol{\Sigma}}_\mathcal{S}$ as a reference point for both domains. To avoid numerical issues, we removed low-variance features (see Appendix A.1 for more details). The remaining features were then standardized to get features with zero mean and unit variance. To predict from the standardized vectors in these simulations, for simplicity, we used Ridge regression with its regularization term set to 1. This model was trained on the source data, and predictions were made on the target data. We evaluated these predictions with R^2 scores. Results are presented in Section 2.5.4 and discussed in Section 2.7.

2.5.4 Methods evaluation

We simulated data according to the four shift scenarios detailed in Section 2.5.1. Figure 2.2 presents the results for each alignment method and each scenario.

The top left Panel (**A**) illustrates the scores for each alignment method on data generated following the **translation** scenario. The value of α controls the shift. As expected, the farther apart the source and target mixing matrices were (higher α values), the worse the performance on the unaligned domains method became (in blue). The z-score baseline (light orange) failed even earlier than using Riemannian geometry without alignment. Methods including a re-centering step (green, dark orange, pink, and brown) did not suffer from this shift. This suggests that whitening the source and target distributions by their respective geometric mean mostly compensated for the mixing matrix being perturbed by an SPD matrix. It allowed the regression model to access the log of the powers with little distortion, hence, allowing the linear model to infer the correct function.

Panel (**B**) presents the **scale** scenario in which the log of the target powers were scaled by a parameter σ_p in the signals. When $\sigma_p = 1$, the source and target distributions were exactly the same $\overline{\boldsymbol{\Sigma}}_{\mathcal{S}} = \overline{\boldsymbol{\Sigma}}_{\mathcal{T}}$. In this case, methods including the re-scaling step adjusted for this shift and made accurate predictions, whereas the performance of other methods deteriorated as σ_p increased. Re-centering helped to achieve better predictions compared to no alignment. The z-score method performed slightly better than not aligning the distribution but was still worse than re-centering.

The third Panel (**C**) corresponds to the **translation and rotation** scenario. Here, the target mixing matrix was modified by interpolating between the source mixing matrix $\boldsymbol{A}_{\mathcal{S}}$ and another randomly generated matrix $\boldsymbol{A}_t$. The parameter m controls where the target mixing matrix is located between these two other matrices, thus how different $\boldsymbol{A}_{\mathcal{S}}$ and $\boldsymbol{A}_{\mathcal{T}}$ were. The only method reaching perfect predictions, irrespective of the value of m, was Procrustes paired. However, the unpaired Procrustes method failed where $m > 0.5$ and even fell behind re-centering. When no rotation correction was applied, re-centering helped to compensate for slight differences between the mixing matrices, but the performance dropped as this difference increased. As expected, a re-centering step and a rotation correction were needed to correct a shift consisting of translation and rotation.

In the last scenario **noise on mixing matrix**, displayed in Panel (**D**), we introduced noise in both source and target mixing matrices to simulate individual differences between subjects. $\sigma_{\mathcal{S},A}$ was set at 10^{-2}, and even when $\sigma_{\mathcal{T},A} = \sigma_{\mathcal{S},A}$ we had $\boldsymbol{A}_{\mathcal{T},i} \neq \boldsymbol{A}_{\mathcal{S},i}$. Procrustes unpaired performed worst in this scenario. The unpaired rotation correction was not robust to noise on the mixing matrix. All the other methods performed similarly for low values of $\sigma_{\mathcal{T},A}$. When $\sigma_{\mathcal{T},A} > 10^{-1}$, all methods deteriorated. The z-score method again showed lower R^2 scores than all other methods. Results suggest that the best solution for this scenario is the paired rotation correction.

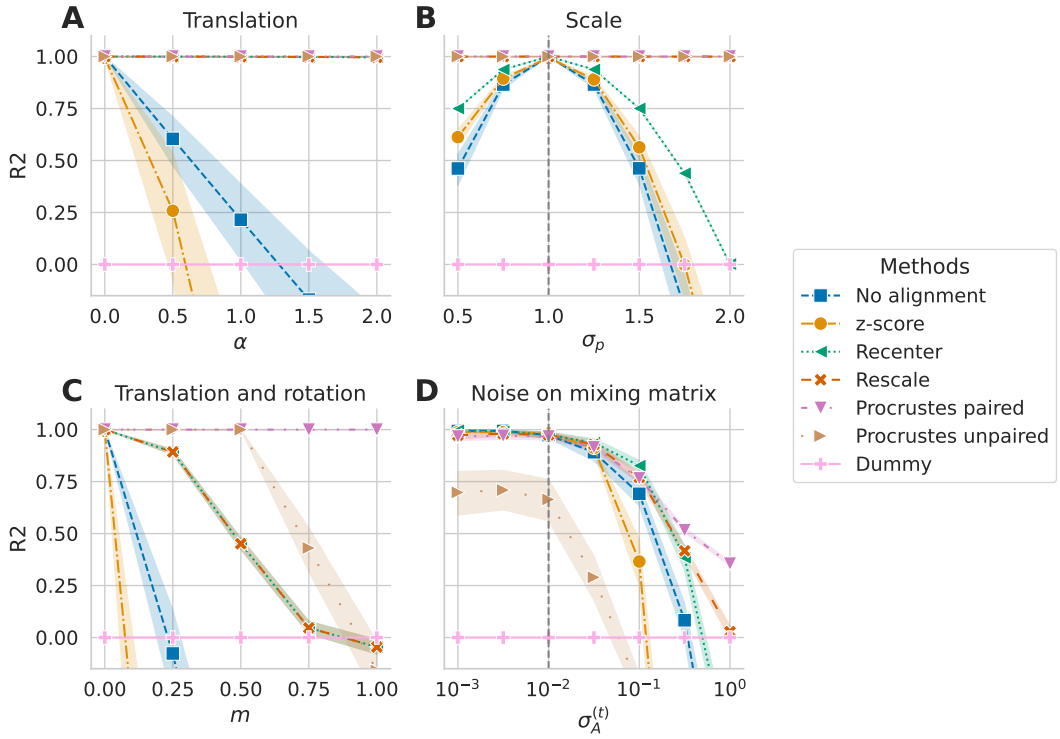

Figure 2.2: **Alignment method comparison across simulated dataset shift scenarios (R^2 score).** Alignment methods were evaluated on four different scenarios with an increasing shift. We generated $N = 300$ matrices per domain to have data sets of the magnitude of real EEG datasets that would be considered as medium to large in terms of operational costs and curation effort. Error bars show standard deviations of the metric obtained with 50 random repetitions. The dashed vertical gray lines on (**B**) and (**D**) indicate the fixed parameter's value of the source set. Panel (**A**) displays the performance achieved when the target covariance matrices were created by multiplying the source mixing matrix with an SPD matrix: $\boldsymbol{A}_{\mathcal{T}} = \boldsymbol{B}^{\alpha}\boldsymbol{A}_{\mathcal{S}}$ with $\boldsymbol{B} \in \mathbb{S}_P^{++}$. All methods that included re-centering the distributions on the same reference point performed well. (**B**) displays the performance achieved when the dispersion of covariances differs between source and distributions ($\sigma_p \neq 1$). Here, the re-scaling step was essential to align the distributions correctly. (**C**) In this scenario, $\boldsymbol{A}_{\mathcal{S}} \neq \boldsymbol{A}_{\mathcal{T}}$, which led to a translation and a rotation of the target set compared to the source set. Re-centering was not insufficient, and a rotation correction was needed to achieve good performance. Interestingly, while Procrustes paired performed well, the unpaired correction broke as the difference between the mixing matrices increased. (**D**) In this scenario, different levels of individual noise were added to the mixing matrices of both domains. For low $\sigma_{\mathcal{T},A}$ values, all methods except the unpaired rotation correction performed similarly with R^2 scores decreasing slowly. For higher values, the scores dropped, and correcting the rotation with the paired method performed best.

The paired rotation correction method performed best in all scenarios but requires the source and target sets to be the same size and have corresponding/paired points. When this is not the case (for datasets with different subjects, for example), re-centering and re-scaling should be the best solution for improving performance. The unpaired rotation estimation seems particularly unstable when the induced shift is too big or when there is noise.

2.6 Numerical evaluation: M/EEG data

In the following empirical benchmarks, we focused on one MEG and two EEG datasets for evaluating our alignment methods with real-world data. We first describe these datasets and their preprocessing, then explain how we computed the covariance matrices of the signals. We followed the same preprocessing and processing steps as in [Engemann et al., 2022] for the 'filterbank-riemann' pipeline. Finally, we detail the design of each benchmark.

2.6.1 Datasets

Cam-CAN MEG data The Cambridge Center of Aging and Neuroscience (Cam-CAN) dataset [Taylor et al., 2017] consists of MEG recordings from a healthy population covering a wide age range. These data were recorded for each subject during resting state with eyes closed, an audio-visual (passive) task with visual and auditory stimuli presented separately, and a sensorimotor (smt) task with the same stimuli as the previous task combined with a manual response. All data were collected with the same 306-channel VectorView MEG system (Elekta Neuromag, Helsinki) with a sampling rate of 1 kHz.

Sample description: We included 646 subjects (319 female, 327 male) with all three recordings. Their age distribution is from 18.5 to 88.9 years with an average of 54.9 ± 18.4 years and an almost uniform spread over the age range. There was no exclusion of participants. The set of subjects of each benchmark only depends on the availability of recordings for the source and the target tasks and the success of the preprocessing and the feature extraction. Thus some subjects with only two tasks recorded are not included in all benchmarks leading to small variations of the subject sample between benchmarks.

Preprocessing: We applied a FIR band-pass filter between 0.1 and 49 Hz to all data. We decimated the signals with a factor of 5 to get a sampling frequency of 200 Hz. To compensate for environmental noise, we performed a temporal signal-space-separation (tSSS) method [Taulu et al., 2005] with a chunk duration of 10 seconds and a correlation threshold of 98%. We only picked channels corresponding to magnetometers (after tSSS signals from magnetometers and gradiometers are mixed and linearly related).

TUAB EEG data The Temple University (TUH) EEG Corpus [Harati et al., 2014] is a large publicly available dataset of clinical EEG recordings. This dataset includes socially and ethnically diverse subjects. In this work, we focus on the Temple University Hospital Abnormal EEG Corpus (TUAB) [Obeid and Picone, 2016], a subset of the TUH EEG Corpus in which recordings were labeled as normal or abnormal by medical experts. Data were collected using several Nicolet EEG devices between 24 and 36 channels and sampled at 500 Hz. The subjects were at rest during the recording.

Sample description: We only included healthy subjects with normal EEG in our benchmark. This led to a sample of 1385 subjects (female = 775 and male = 610) with ages between 0 and 95 years (mean = 44.4 years and std = 16.5 years).

Preprocessing: Data were band-pass filtered between 0.1 and 49 Hz with a zero-phase finite impulse response (FIR) filter using the firwin with Hamming window featuring a 0.0194 passband ripple, 53 dB stopband attenuation, a 0.1 Hz lower transition bandwidth and a 12.25 Hz upper transition bandwidth, and a filter length of 6601 samples (33.005 s). Data were then resampled to 200 Hz. We selected a subset of 21 channels common to all recording devices used in this dataset. When several recordings were available for one patient, we picked the first to get only one recording per subject.

LEMON EEG data The Leipzig Mind-Brain-Body database provides multimodal data from healthy groups of young and elderly subjects [Babayan et al., 2019]. In our benchmark, we only used EEG recordings from this dataset. They were recorded with a 62-channel ActiCAP device and sampled at 2500 Hz. Each subject did two recordings at rest with two conditions: eyes closed and eyes open.

Sample description: We included 213 subjects from the LEMON database in our benchmark. No selection criteria were applied, and we kept the data for which the processing and the feature extraction were successful. This led to a cohort with 134 males and 79 females aged from 20 to 77 years. The age distribution of the LEMON presents a peculiarity: it is split into two separate age groups, one with individuals being between 20 and 35 years old and the second between 55 and 77 years old.

Preprocessing: A band-pass filter between 0.1 and 49 Hz was applied to the data and resampled to 200 Hz. To keep a maximum of data, recordings with eyes closed and eyes open were pooled before feature extraction.

2.6.2 Processing and feature extraction

Name	low	δ	θ	α	β_{low}	β_{mid}	β_{high}
Range (Hz)	0.1 — 1	1 — 4	4 — 8	8 — 15	15 — 26	26 — 35	35 — 49

Table 2.2: **Definition of frequency bands.**

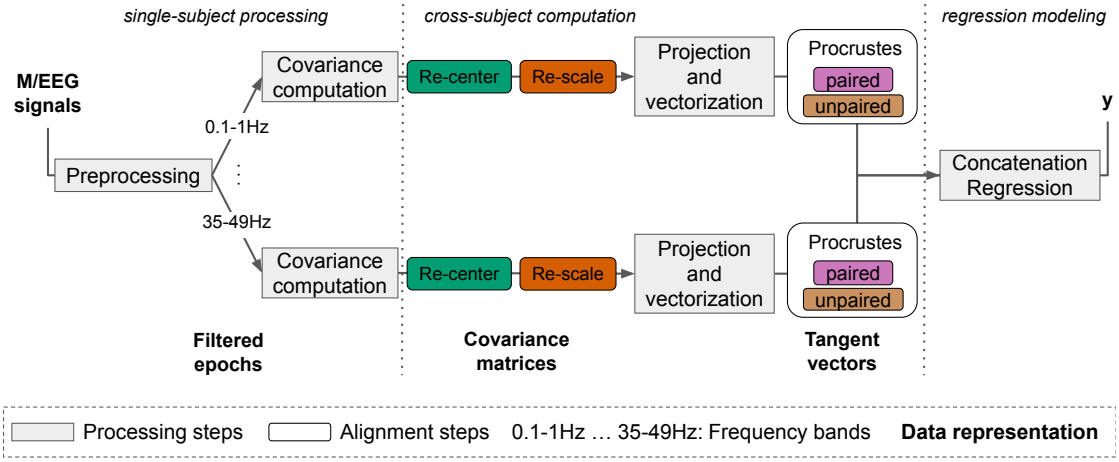

Figure 2.3: **Pipeline for regression modeling with M/EEG with different dataset harmonization steps.** For every subject, we summarize the M/EEG recording by the covariance matrix after performing artifact cleaning (Section 2.6.1). The covariances computation, alignments steps, projection to the tangent space, and vectorization steps are done separately for seven frequency bands of Table 2.2. Alignment steps detailed in Section 2.4.2 are computed from the covariance distribution across all subjects. The re-center and re-scale steps are performed separately for source and target datasets. The Procrustes steps combine information across source and target datasets. Finally, the seven resulting tangent vectors are concatenated to form one vector per subject used for regression.

After preprocessing, each filtered recording was segmented in 10 s epochs without overlap. Epochs were then filtered into seven frequency bands as defined in Table 2.2 as in [Engemann et al., 2022]. We performed artifact rejection by thresholding extreme peak-to-peak amplitudes on single epochs using the local autoreject method [Jas et al., 2017]. Subsequently, we computed covariance matrices from the set of artifact-free epochs with the Oracle Approximating Shrinkage (OAS) estimator [Chen et al., 2010]. The ensuing regression pipeline, including all alignment steps, is illustrated in Figure 2.3.

For MEG signals, the tSSS method reduces noise by projecting them in a subspace mainly containing the signal, leading to rank-deficient covariance matrices. As a result, it is not possible to correctly apply our alignment methods directly, as rank-deficient covariance matrices are not SPD matrices. To extract valid SPD matrices, we follow the approach from [Sabbagh et al., 2019] introduced in Section 1.3.3 and apply Principal Component Analysis to reduce the dimensionality of the covariance matrices, which renders them full rank. Matrices of both domains are transformed following (1.25) as:

$$\boldsymbol{\Sigma}_{\mathcal{S},\text{filtered},i} = \boldsymbol{W}_{\mathcal{S}} \boldsymbol{\Sigma}_{\mathcal{S},i} \boldsymbol{W}_{\mathcal{S}}^{\top} \in \mathbb{R}^{R \times R} \tag{2.28}$$

$$\boldsymbol{\Sigma}_{\mathcal{T},\text{filtered},i} = \boldsymbol{W}_{\mathcal{S}} \boldsymbol{\Sigma}_{\mathcal{T},i} \boldsymbol{W}_{\mathcal{S}}^{\top} \in \mathbb{R}^{R \times R} \tag{2.29}$$

For Cam-CAN data, we set $R = 65$. This spatial filter is applied to each frequency band separately. We did not apply this procedure to EEG data.

2.6.3 Alignment, vectorization, and regression

The matrices from both domains were first aligned with the methods described in Section 2.4.2. We projected the aligned data in the tangent space at $\overline{\boldsymbol{\Sigma}}_{\mathcal{S}}$ to get tangent vectors. Tangent vectors from all frequency bands were concatenated. Then we applied ridge regression after standardizing (z-scoring) all the features. To select the regularization hyperparameter, we used a generalized (leave-one-out) cross-validation [Gene H. Golub and Wahba, 1979] on a logarithmically spaced grid of 100 points from 10^{-5} to 10^{10}. For quantifying prediction performance, we use the R^2 score.

2.6.4 Cam-CAN (MEG): same subjects

For this benchmark, we used the experimental tasks of the Cam-CAN dataset for defining the different domains. Here, the source subjects were the same as the target subjects. Only the experimental task (e.g., audiovisual VS audiovisual + motor) changed from one domain to the other. All subjects included underwent MEG recordings for both the source and the target domain. Therefore, importantly, the mixing matrix and the age distributions were the same for the source and target. As we dealt only with healthy participants, we aimed to minimize the error in age prediction when learning on one task (source domain) and predicting

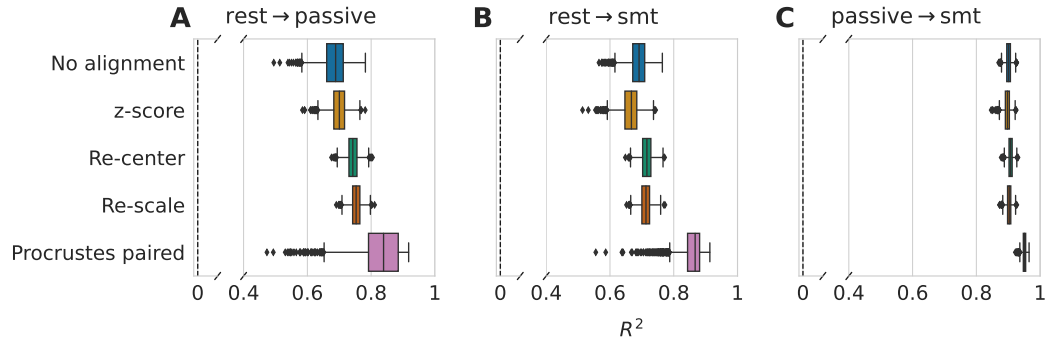

Figure 2.4: **Impact of data alignment on age prediction across different tasks on the same subjects from Cam-CAN dataset (R^2 score).** Alignment methods comparison for three different source-target tasks using 2000 repeat-bootstrap to select the subjects. Both domains contained the same subjects, only their task was different. Models are depicted along the y-axis, and standard boxplots represent their associated R^2 score. The dashed black lines represent chance-level performance. (**A**) Generalization of age prediction regression model from resting state to the passive task. Re-centering and the paired rotation correction led to an increased R^2 score with no obvious benefits for additional re-scaling. (**B**) The regression model was trained on resting-state data, and predictions were made on the recordings of the somatosensory task. Re-centering the data led to slightly improved R^2 scores. Again, the re-scaling step did not lead to further improvements. Correcting the rotation with the paired method contributed to improving 99% of the splits in comparison to only re-centering. (**C**) Here we used the data from the passive task as the source domain and the somatosensory task as the target domain. Re-centering and re-scaling steps did not affect the prediction performance. The paired rotation correction improved the scores in all splits.

on another (target domain). To estimate standard deviation, we did a bootstrap with 100 repetitions.

We first focused on a regression problem for which there was an individual noise on the mixing matrices, but their distribution was the same in source and target sets because the subjects were the same. It also implies that the age distribution was identical for both domains. The interest of this analysis is to assess what kind of shift is produced when only the task changes and if alignment methods can rectify this. Results for each alignment method on age prediction for three source-target tasks associations are displayed in Figure 2.4.

The z-score method led to scores similar or lower to what is obtained without alignment across all three Panels, as expected from the simulation results. For the two first Panels (**A** and **B**), the source domain contained the resting-state recordings, and the target tasks were, respectively, the passive and the somatosensory tasks. The R^2 scores we obtained after these two benchmarks were highly similar. When no alignment was done, the mean R^2 score was around 0.7. Re-centering the distributions led to a reduced standard deviation and an increased mean R^2 score in both situations, even though this was more pronounced in Panel (**A**). The re-scaling step had no obvious impact on performance. The paired rotation correction led to improved prediction scores on 87.4% of the bootstrap iterations in Panel (**A**) and on 99.1% of the iterations in Panel (**B**) compared to only re-centering. In Panel (**C**), the source domain was the passive task, and we made predictions on the somatosensory task, leading to quite different results. The performance reached with no alignment was already very high, with a mean R^2 of 0.9. Re-centering and re-scaling gave the same results as not performing any alignment. Then, the paired rotation correction step induced increased scores for all bootstrap iterations ($R^2 = 0.951 \pm 0.005$). Going from rest to tasks affects the geometric mean of the covariance matrix distributions, but not when going between passive and smt tasks. In all three situations, the performance gain obtained with Procrustes paired implies the presence of a rotation of the tangent vector distribution.

2.6.5 Cam-CAN (MEG): different subjects

In this second benchmark on the Cam-CAN data, we again defined the different domains in terms of the experimental MEG tasks performed by the subjects. Yet, the critical difference with the previous benchmarks is that the source subjects and the target subjects were distinct persons. To implement this analysis, we randomly divided all Cam-CAN subjects into subsets of 80% forming the source subjects, and the left-out 20 % forming the target subjects. A stratification was performed by age decade to maintain similar age distributions between splits. We repeated this split with 100 different random initializations.

In this second benchmark, we focused on domain-shift differences between MEG tasks in non-overlapping samples of distinct subjects. As a consequence, the distributions of mixing matrices, necessarily, differed for the source and target domains. We performed a cross-validation in which 80% of the subjects were assigned to

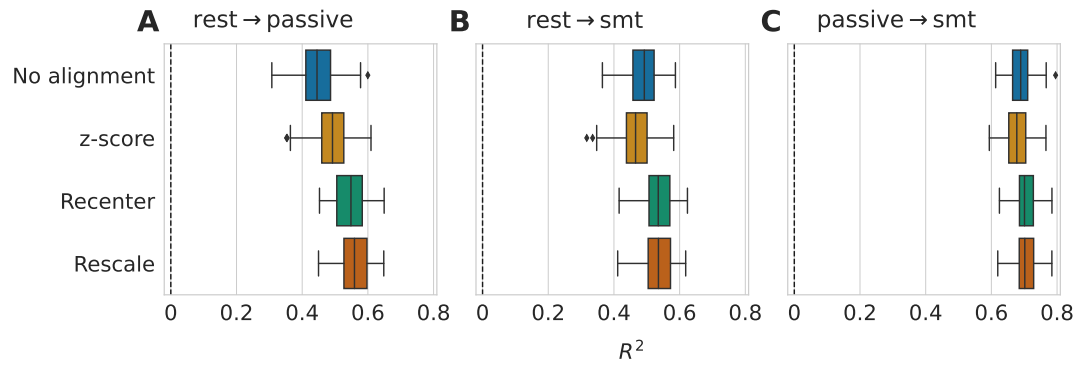

Figure 2.5: **Impact of data alignment on age prediction across different tasks for different subjects from Cam-CAN dataset (R^2 score).** Alignment methods comparison for three different source-target tasks using 100 stratified Monte Carlo cross-validation (shuffle split) iterations to determine which subjects form the source and the target sets. We depict the models along the y-axis and represent the R^2 scores with standard boxplots. The dashed black lines represent chance-level performance. (**A**) The model was trained on the rest task, and predictions were made on the passive task recordings. When re-centering source and target distributions, prediction performance substantially improved, whereas re-scaling did change performance. (**B**) The target set was composed of recordings from the somatosensory task. The improvement of the re-centering step was smaller but still present. Re-scaling, still, did not lead to obvious improvements. (**C**) In the last Panel, the passive task was the source domain, and the somatosensory task was the target. In this case, aligning was not helpful and led to the same performance as not performing any alignment.

the source domain and 20% to the target domain. To keep relatively similar age distributions in the train and test splits, we did a stratification on the age decades of the subjects (cf `StratifiedShuffleSplit` of the Scikit-Learn software). Applying the paired rotation correction in this setup was no longer possible. Thus, it is impossible to analyze whether a rotation exists in the shift. We present the results of each alignment step in Figure 2.5.

When covariance matrices were not aligned, generalizing from rest to passive tasks led to a R^2 score of 0.55 ± 0.05 (**A**), and when the target task was the smt task, we observed an $R^2 = 0.54 \pm 0.04$ (**B**). The z-score method performed again similarly to the procedure without alignment. The re-centering step led to comparable results across generalization scenarios involving resting state and any event-related task (Panels (**A**) and (**B**)). Again, matching the source and target dispersions with re-scaling was not helpful. Finally, all methods showed similar performance when the passive and the smt tasks were the source and the target tasks, respectively (**C**). Our observations for this benchmark match those we made when the subjects were the same for the source and target sets. The R^2 scores reached after alignment in Figure 2.5 are considerably lower than in Figure 2.4. Having different subjects in the source and the target domain clearly creates a more difficult-to-reachable shift.

2.6.6 TUAB to LEMON (EEG): different subjects

In this benchmark, we gauged the performance of alignment methods when the source and target domains are two different datasets. Here, the source domain was composed of data from TUAB, and the target one of data from LEMON. These datasets were not recorded with the same device. However, they had 15 channels in common. We picked the same channels on both datasets to define covariance matrices of the same shape and similar information. The target set was kept fixed, and we implemented a bootstrap procedure on the source subjects to estimate standard deviations. In this setup, in addition to evaluating the alignment methods on the Riemannian regression model, we also applied them with a regression model based on Source Power Comodulation (SPoC). SPoC is a supervised spatial filtering method in which the filters $\boldsymbol{W}_{\text{SPoC}}$ maximize the covariance between the power of the filtered signals and the outcome y [Dähne et al., 2014]. Denoting by $\widetilde{\boldsymbol{\Sigma}} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{\Sigma}_i$ the Euclidean average covariance matrix and $\boldsymbol{\Sigma}_y = \frac{1}{N}\sum_{i=1}^{N} y_i \boldsymbol{\Sigma}_i$ the weighted average covariance matrix, the first filter $\boldsymbol{w}_{\text{SPoC}}$ is given by:

$$\boldsymbol{w}_{\text{SPoC}} = \arg\max_{\boldsymbol{w}} \frac{\boldsymbol{w}^\top \boldsymbol{\Sigma}_y \boldsymbol{w}}{\boldsymbol{w}^\top \widetilde{\boldsymbol{\Sigma}} \boldsymbol{w}}. \tag{2.30}$$

The same idea was proposed by [de Cheveigné and Parra, 2014]. As $\boldsymbol{W}_{\text{SPoC}}$ recovers the inverse of the mixing matrix $\boldsymbol{A}$ [Sabbagh et al., 2020], the SPoC regression model is defined as:

$$\boldsymbol{z}_i = \text{diag}(\log(\boldsymbol{W}_{\text{SPoC}} \boldsymbol{\Sigma}_i \boldsymbol{W}_{\text{SPoC}}^\top)) \tag{2.31}$$

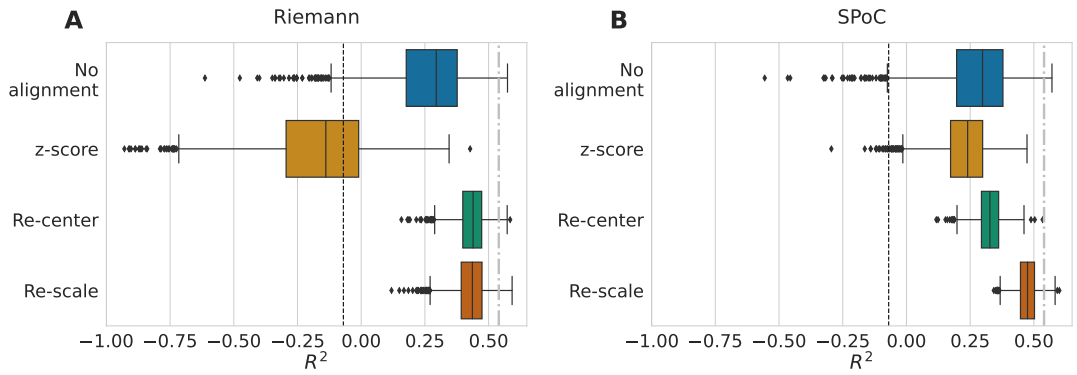

Figure 2.6: **Impact of data alignment on age prediction across different EEG datasets (R^2 score).** Data from the TUAB dataset were used as the source domain, and from the LEMON data as the target domain. We compare the alignment methods across 2000 bootstrap iterations on the source data ($N = 1385$). The target set was always the same ($N = 213$). The methods are represented along the y-axis, and we depict their associated R^2 scores with standard boxplots. The dashed black lines represent chance-level performance. (**A**) Results of alignment methods combined with the Riemannian approach of (2.7) (as for all the results we have previously presented). Without alignment, the prediction made on the LEMON data led to R^2 scores far lower than what was reported in [Engemann et al., 2022] (10-fold cross-validation on LEMON data only: 0.54 ± 0.13 represented by the dashed gray line). When both domains are re-centered to identity, we reached performances similar to when the model is trained on LEMON. Re-scaling did not visibly improve results. (**B**) Results when the regression model follows the SPoC approach. Not aligning led again to poor R^2 scores. Unlike the first panel, the z-score method improved the predictions similarly to re-centering. Re-scaling helped to reach performances on par with the Riemannian model trained on LEMON.

We now consider the resting-state data from two different EEG datasets. The source and target populations are different, as well as the recording devices, but all recordings were done at rest. The source dataset is the larger TUAB dataset (N=1385), and the target dataset is the smaller LEMON dataset (N=213). TUAB also has a broader age range. This way, the regression model will be asked to predict age values that fall within the range observed during model training. We performed a bootstrap with 2000 iterations on TUAB data. The results are reported in Figure 2.6. In addition to the Riemannian approach we focused on in this work (**A**), we were also interested in the impact of the alignment methods on a non-Riemannian model like SPoC (**B**).

Without alignment, the Riemannian model and SPoC led to poor results with mean R^2 scores around 0.26. On Panel (**A**), the z-score method performed at the level of the dummy model. Re-centering the data drastically improved the age prediction performances with R^2 scores of 0.44 ± 0.06 with a visibly reduced variance. Adding the re-scaling step on top of re-centering did not bring any improvement in performance. In [Engemann et al., 2022], the 'filterbank-riemann' pipeline trained on LEMON data only with a 10-fold cross-validation led R^2 scores of $R^2 = 0.54 \pm 0.13$. Here, the training dataset only consisted of data from the TUAB dataset. The Riemannian re-center step made it possible to reach performance comparable to a model trained within the same dataset. With SPoC (**B**), re-centering led to a reduced standard deviation compared to no alignment. The highest R^2 scores were achieved when the re-scaling step was added to the alignment procedure and almost reached the performance of the Riemannian model trained on LEMON.

Aligning the covariances distribution helped improve prediction performance even with a regression model like SPoC that does not leverage the geometry of the covariance matrices. This observation motivated an examination of how alignment affects the SPoC patterns, the inverse of the SPoC spatial filters $\boldsymbol{W}_{\text{SPoC}}$, and the resulting powers. As re-centering is a linear transformation, it is possible to combine it with the SPoC patterns for visualization. Thus this is the alignment method we used for the results displayed in Figure 2.7. The first two rows of (**A**) illustrate the five first SPoC patterns of unaligned source (TUAB) and target (LEMON) data. Without alignment, the source patterns of the first row were directly applied to the unaligned source and target data, resulting in the log powers represented by the blue dots in the scatterplot (**B**). The target log powers covered a wider range of values than the source log powers and did not match the identity line. We then trained the model on the aligned source data and applied it to the aligned source and target data to get the log powers values represented as orange crosses on the scatterplot (**B**). Re-centering each domain independently resulted in more comparable source and target log powers on average across subjects. To visualize the patterns associated with the aligned log powers and compare them to the unaligned source and target patterns, we displayed on the third row of Figure 2.7 (**A**) the SPoC patterns of the aligned source data adjusted with the target

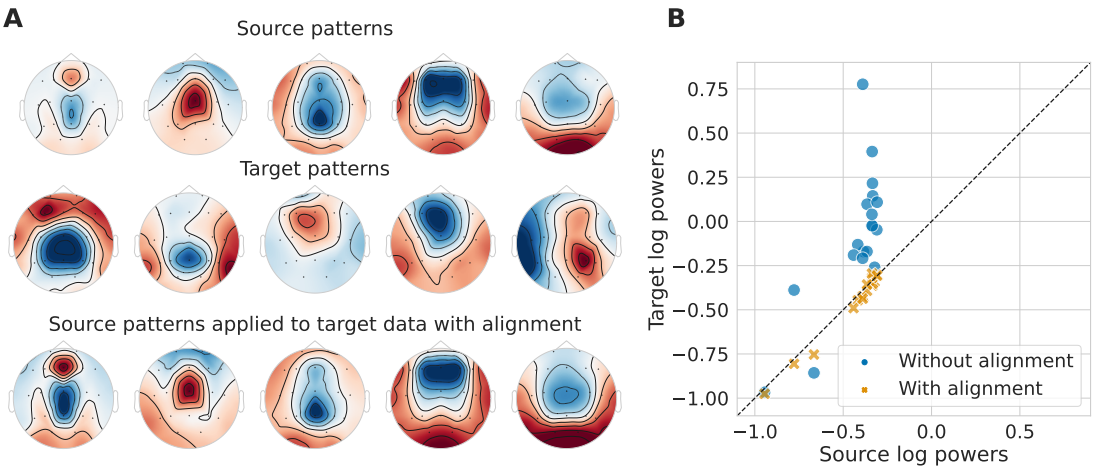

Figure 2.7: **Impact of alignment of different EEG datasets on their SPoC patterns and source powers.** TUAB data were used as the source domain, and LEMON data as the target domain. Alignment refers to re-centering the source and the target distribution by whitening them respectively by their geometric mean. To obtain these figures, data were filtered in the alpha band. We included 19 channels (15 commons and 4 with similar locations on the scalp) in both datasets. (**A**) Topographic maps of the five first SPoC source patterns without alignment (first row) and target patterns without alignment (second row). The third row corresponds to the aligned source patterns adjusted with the target whitening inverse filter. These are the patterns applied to unaligned target data to obtain the target powers with alignment. The color map is normalized across each row. (**B**) Scatter plot of the target log powers as a function of the source log powers without and with alignment averaged across subjects. The dashed black line is the identity line. Alignment makes target and source log-powers more comparable.

whitening inverse filter $\overline{\mathbf{\Sigma}}_{\mathcal{T}}^{\frac{1}{2}}$. In other words, these adjusted patterns correspond to the SPoC filters applied to the unaligned target data to obtain the target log powers with alignment. The shapes of the adjusted patterns look similar to the source patterns of the first rows without any clear transformation in the direction of the target patterns. Even though this analysis was performed in the alpha band, we made the same observations in all other frequency bands.

2.7 Discussion

In this study, we thoroughly explored domain adaptation methods that align M/EEG covariance matrix distributions for regression problems on both simulations and large datasets. We considered methods from BCI applications [Rodrigues et al., 2019, Maman et al., 2019, Bleuzé et al., 2021] articulated in three alignment steps: re-center the geometric means, equalizing dispersions, and rotation correction. These alignment steps are evaluated in the regression context of generalizing age prediction across different domains. We investigated how dataset shifts can occur by analyzing a statistical generative model of M/EEG data. We presented simulated dataset shift scenarios based on this model for which alignment steps can effectively compensate the shift, plus a noise scenario to get a sense of how the methods would perform with real data. The simulation results showed that Procrustes paired is the most effective method in all scenarios. It was expected as it includes the three alignment steps and a rotation correction informed by the pairing of source and target subjects. We then designed M/EEG benchmarks with different domain definitions to determine the alignment methods' efficiency in those various settings. Coherently with the simulation results, Procrustes paired achieved the best performance, but since it cannot be applied in all situations, re-centering is the best option.

We compared the alignment steps leveraging Riemannian geometry with a z-score method that transforms the covariance matrices into correlation matrices. This method systematically performed worse than all the others. Taking into consideration the geometry of the data space is essential. Among the three Riemannian alignment steps, re-centering and the paired rotation correction of the source and target distributions help to improve the prediction performance in the M/EEG benchmarks. Re-centering and the paired rotation correction were shown to compensate for changes in the mixing matrices of the generative model, so we expected these steps to reduce the shift in benchmarks where the target population is not the same as the source population. In the first benchmark on Cam-CAN data, the source and target subjects were the same, but we still observed that re-centering and Procrustes paired led to better scores. On the other hand, equalizing their dispersions did not bring clear gains in performance in any benchmark. In the Cam-CAN benchmarks, the scores reached when the subjects are d ifferent in the domains are distinctly lower than with the same subjects: the shift is bigger (lower no alignment baseline) and harder to recover. For the EEG benchmark, we

used the TUAB dataset as the source and the LEMON dataset as the target. Here, all recordings were done at rest but with different recording devices and in different populations. Re-centering the distributions in the EEG benchmark exceeded our expectations. Re-centering was sufficient to recover performance close to what is reached when training the Riemannian model on LEMON [Engemann et al., 2022]. The re-centering step is simple to implement and has already been very effective in BCI classification to deal with variability between sessions [Barachant et al., 2013] but also between subjects [Zanini et al., 2018]. Our results suggest it is also effective in a regression context with variability between populations, tasks, and recording devices.

We extended our evaluation of the impact of alignment methods on different EEG datasets to the SPoC model [Dähne et al., 2014]. In this setting, the z-score method and re-centering performed both equally better than no alignment. Interestingly, re-scaling was beneficial and helped to reach performance close to the Riemannian model trained on LEMON. By inspecting the SPoC patterns and the associated log powers, we demonstrated that the observed gain in the performance of re-centering was enabled by more similar log powers between source and target than without alignment. In other words, data alignment adapts the target features to the regression equations fitted on the source data, which explains generalization.

Unfortunately, the two last benchmarks are missing a rotation correction method. As Procrustes paired led to an apparent score increase in the first benchmark, we expect a rotation correction to be beneficial in the other benchmarks. Yet, the simulation study showed that the unpaired Procrustes method failed to correctly estimate the rotation when there is noise or when the shift gets too large. The result suggests that this method would likely fail with M/EEG data. The condition of matching subjects between the source and target in Procrustes paired is too restrictive and is not applicable in many settings. The supervised rotation correction methods developed for classification problems [Maman et al., 2019, Bleuzé et al., 2021, Rodrigues et al., 2019] are unsuitable for regression. Further investigations are needed to fill the lack of rotation correction in regression contexts.

Another limitation of this work is that we only performed benchmark that involved source and target covariance matrices formed from the same set of sensors. The dimensionalities of the source and the target data must be equal to apply the predictive model to it. It has been proposed to deal with different dimensionalities of covariance matrices via zero-padding [Rodrigues et al., 2021]. However, this method is not applicable if there is no rotation correction afterward, so we could not use it. We also observed that our framework is not robust to sensor permutation in the covariance matrices, even with the same sets of sensors. In our EEG benchmark, we had to select the common channels between the two datasets and sort them to reach acceptable performance even for the re-center step. In addition to leading to a gain in performance, having a proper rotation correction would help to deal with issues related to different numbers or types of sensors in the source and target datasets.

Besides the limitations linked to the rotation correction, further points would deserve future studies. First, we explored unsupervised alignment methods that do not explicitly share any information between the source and the target domains. Comparing our results with supervised methods could allow us to quantify the gain of supervision and to have additional insights into the trade-off between the approaches. A second element to consider is that our goal was to evaluate the alignment methods on a regression problem by minimizing the prediction error. We focused on brain age prediction as age is a label that is easy to collect. But other prediction targets should equally benefit from the methods presented in this work. Importantly, we conducted our benchmarks on healthy participants sampled from the general population. Yet, the biggest impact of our results may be seen when bridging datasets from heterogeneous clinical populations, which remains to be demonstrated. Finally, in this work we focus on linear regression model but recent work demonstrated that this Riemannian framework can also be applied with non-linear models [Bonet et al., 2023]. Kernel-based models have been shown to perform well on brain age prediction but were not investigated in a domain adaptation context.

3 Joint shift in data and label distributions

Contents

In Chapter 1 and Chapter 2, we introduced the problem of dataset shift in EEG data and highlighted the need for domain adaptation methods to effectively manage these shifts. Until now, our focus was on situations where shifts predominantly occur in the input data X. We analyzed the mechanisms and causes of these shifts, and investigated domain adaptation techniques designed to handle differences in in the distribution of input data. We observed that implementing a riemannian re-centering alignment improved prediction performance when there were changes in tasks performed during recording, or different participants between the source and target domains.

In this chapter, we extend our investigation to scenarios where shifts occur in both the input data X and the outcome variable y. This situation is more reflective of real-world EEG data analysis, where participant populations can vary considerably between different recording sites. We illustrate this scenario using the HarMNqEEG dataset [Li et al., 2022], which includes EEG data from various recording sites with differing age ranges and distributions. Through our experimental analysis, we demonstrate that the alignment approaches used in the previous chapter are inadequate for this context and fail to deliver optimal performance.

To address this limitation, we propose a novel domain adaptation method capable of handling shifts in both X and y distributions. Our results show that this new method outperforms previous techniques on the HarMNqEEG dataset, thereby offering a more robust solution for EEG data analysis in the situation of joint dataset shifts.

3.1 Dealing with complex dataset shift in EEG data

Machine learning (ML) has enabled advances in the analysis of complex biological signals, such as magneto- and electroencephalography (M/EEG), in diverse applications including biomarker exploration [Wu et al., 2020, Hakeem et al., 2022, Yang et al., 2022] or developing Brain-Computer Interface (BCI) [Wolpaw et al., 1991, Forenzo et al., 2024, Allahgholizadeh Haghi et al., 2019, Anumanchipalli et al., 2019]. However, a major challenge in applying ML to these signals arises from their inherent variability, a problem commonly referred to as dataset shift [Dockès et al., 2021]. In the case of M/EEG data this variability can be caused by differences in recording devices (electrode positions and amplifier configurations), recording protocols, population demographics, and inter-subject variability [Mellot et al., 2023, Engemann et al., 2018, Heremans et al., 2022, Jiang et al., 2023]. Notably, shifts can occur not only in the data X but also in the biomedical variable y we aim to predict, further complicating the use of ML algorithms.

Riemannian geometry has significantly advanced EEG data analysis by enabling the use of spatial covariance matrices as EEG descriptors [Barachant et al., 2010, 2013, 2012, Nguyen et al., 2017, Kobler et al., 2022, Lopez Naranjo et al., 2024, Li et al., 2022, Sabbagh et al., 2020, Gemein et al., 2020, Wilson et al.,

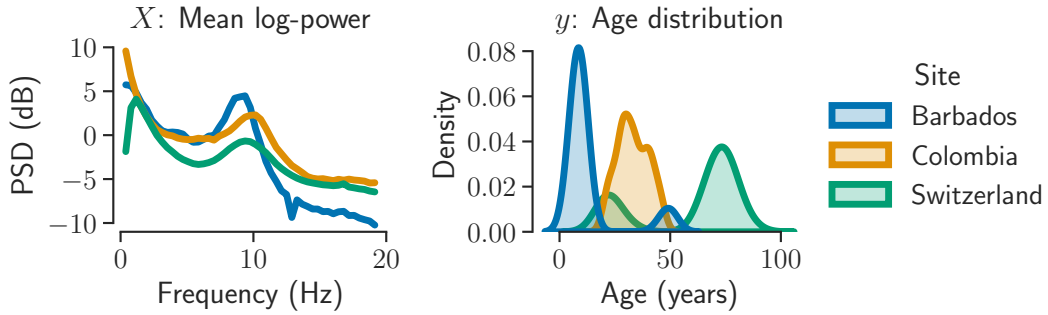

Figure 3.1: **Joint shift in X and y distributions on the HarMNqEEG dataset [Li et al., 2022].** Subset of mean PSDs (**A**) and age distributions (**B**) from three recording sites used for the empirical benchmarks.

2024]. In Barachant et al. [2013, 2012], the authors introduced a classification framework for BCI based on the Riemannian geometry of covariance matrices. These methods classify EEG signals directly on the tangent space using the Riemannian manifold of symmetric positive definite (SPD) matrices ($\mathbb{S}_P^{++}$), effectively capturing spatial information. More recently, Sabbagh et al. [2019, 2020], Bomatter et al. [2024] extended this framework to regression problems from M/EEG data in the context of biomarker exploration. Furthermore, Sabbagh et al. [2019, 2020] proved that Riemannian metrics lead to regression models with statistical guarantees in line with log-linear brain dynamics [Buzsáki and Mizuseki, 2014] and are, therefore, well-suited for neuroscience applications. Across various biomarker-exploration tasks and datasets, recent work has shown that Riemannian M/EEG representations offer parameter-sparse alternatives to non-Riemannian deep learning architectures [Engemann et al., 2022, Paillard et al., 2024, Gemein et al., 2020].

Domain adaptation addresses the challenges posed by differences in data distributions between source and target domains, e.g., when data are recorded with different cameras in computer vision [Wang and Deng, 2018], different writing styles in natural language processing [Li et al., 2019a], or varying sensor setups in time series analysis [He et al., 2023]. In particular, DA considers *target shift* where the shift is in the outcome variable y. For classification it means source and target data share the same labels but in different proportions [Li et al., 2019b]. Target shift is also frequent in the context of multicenter neuroscience studies, as the studied population of one site may vary significantly from the studied population of another site. Figure 3.1 Illustrates an example of this situation with data from the HarMNqEEG dataset Li et al. [2022]. To tackle various sources of variability in neurophysiological data like EEG, there is a need for a DA approach that can deal with a joint shift in X and y.

The rest of the chapter is organized as follows. In Section 3.3, we discuss works related to the approach we propose in this chapter. We first review covariance-based methods used for domain adaptation in the BCI context. We interpret the predictive pipeline used throughout the manuscript on $\mathbb{S}_P^{++}$ as parallel transports combined with Riemannian logarithmic mappings. We also discuss mixed-effects models, which are to be applied in context where the distribution of the input data X and the outcome variable y are shifted. In Section 3.4 we present our proposed method, Geodesic Optimization for Predictive Shift Adaptation (`GOPSA`), which learns to parallel transport each domain along a geodesic. We detail the optimization problem and the train-time and test-time algorithms associated with `GOPSA`. Finally, in Section 3.5, we apply `GOPSA` as well as different baselines on the HarMNqEEG dataset. In addition, we provide a model inspection to understand the behavior of `GOPSA` and its impact on the data.

3.2 Contributions

The content of this chapter is based on the work presented in the preprint:

Apolline Mellot, Antoine Collas, Sylvain Chevallier, Alexandre Gramfort, and Denis A. Engemann. "Geodesic Optimization for Predictive Shift Adaptation on EEG data." under review,

In this work, we address the challenging problem of multi-source domain adaptation with predictive shifts on the SPD manifold, focusing on distribution shifts in both the input data X and the variable to predict y. We propose a novel test-time domain adaptation method called Geodesic Optimization for Predictive Shift Adaptation (`GOPSA`), which adapts models to new domains during inference without retraining on source data. It enables mixed-effects modeling by jointly learning parallel transport along a geodesic for each domain and a global regression model common to all domains, with the assumption that the mean $\bar{y}_{\mathcal{T}}$ of the target domain is known. `GOPSA` aims to advance the state of the art by: *(i)* addressing shifts in both covariance matrices and the outcome variable y, *(ii)* being tailored for regression problems, and *(iii)* being a multi-source test-time domain adaptation method, meaning that once trained on source data, it can generalize to any target domain without requiring access to source data or retraining a new model.

The scripts of the baseline methods and the proposed approach are available available on GitHub along with the codes for the experimental evaluation on the HarMNqEEG dataset:

`https://github.com/apmellot/GOPSA.git`

3.3 Related work

3.3.1 Covariance-based transfer learning

In Zanini et al. [2018], the authors addressed domain adaptation for EEG-based BCI using re-centering affine transformation of covariance matrices, presented in Section 2.4.2, to align data from different sessions or human participants, improving classification accuracies. Yair et al. [2019] extended this with parallel transport showing its effectiveness in EEG analysis, Peng et al. [2022] introduced a domain-specific regularizer based on the Riemannian mean. Notably, this parallel transport approach reduces to Zanini et al. [2018] when the common reference point is the identity.

In a deep learning context, Kobler et al. [2022] proposed to do a per-domain online re-centering which can be seen as a domain specific Riemannian batch norm. Going beyond re-centering, Riemannian Procrustes Analysis (RPA) [Rodrigues et al., 2019] was proposed for EEG transfer learning, using three steps: mean alignment, dispersion matching, and rotation correction. However, the rotation step is unsuitable for regression problems and RPA adapts only a single source to a target domain. In Chapter 2, we demonstrated the benefits of re-centering for regression problems, showing improvements in handling task variations in MEG and enhancing across-dataset inference in EEG.

3.3.2 Mixed-effects models

On the other hand, mixed-effects models (or multilevel models) have been successfully used to tackle data shifts in X and y [Gelman, 2006, Hox, 1998]. In biomedical data, mixed-effects models are crucial due to the presence of common effects, such as disease status and age. These common effects are often intertwined with site-specific effects, like differences in recording devices and protocols. Riemannian mixed-effects models have been used to analyze observations on Riemannian manifolds, accommodating individual trajectories with mixed effects at both group and individual levels [Kim et al., 2017, Schiratti et al., 2015, 2017]. These models adapt a base point on the manifold for each data point and utilize parallel transport for this adaptation, which is necessary for accurate trajectory modeling. However, they differ significantly from the problem we address in this work. Notably, the input data X are covariates (e.g., age or disease status) which belong to a Euclidean space and the variables y to predict belong to the manifold (e.g., MRI diffusion tensors on $\mathbb{S}_P^{++}$) which is the opposite of the paper's studied problem. This distinction is critical as it highlights that while both methods use the geometry of Riemannian manifolds, the nature of the predicted variables and the type of data used differ from existing Riemannian mixed-effects models.

3.4 Learning to re-center from highly shifted y distributions

In this section, we introduce a novel multi-source domain adaptation method, called Geodesic Optimization for Predictive Shift Adaptation (`GOPSA`), that operates on the $\mathbb{S}_P^{++}$ manifold and is capable of handling vastly different distributions of y. Our approach implements a Riemannian mixed-effects model, which consists of two components: a single parameter estimating a geodesic intercept specific to each domain and a set of parameters shared across domains. Notably, `GOPSA` is test-time adaptable, meaning that once trained on source data, it can generalize to any target domain without requiring source data or retraining a new model.

At train-time, `GOPSA` jointly learns the parallel transport of each of the K source domains and the regression model shared across domains. At test-time, we assume having access to the target mean response value $\bar{y}_{\mathcal{T}}$ and predict on the unlabeled target domain of covariance matrices $(\boldsymbol{\Sigma}_{\mathcal{T},i})_{i=1}^{N_{\mathcal{T}}}$. `GOPSA` focuses solely on learning the parallel transport of the target domain so that the mean prediction, using the regression model learned at train-time, matches $\bar{y}_{\mathcal{T}}$.

3.4.1 Parallel transport along the geodesic for multi-source domain adaptation

Parallel transport

A classical practice to align distributions is parallel transport of covariance matrices from their mean to the identity and then apply the logarithmic mapping (1.12). Parallel transport along a curve allows to move SPD matrices from one point on the curve to another point on the curve while keeping the inner product between the logarithmic mappings with any other vector transported along the same curve constant. The following lemma gives the parallel transport of $\boldsymbol{\Sigma}'$ from $\boldsymbol{\Sigma}$ to $\boldsymbol{I}_P$ along the geodesic between these two points (See proof in Appendix B.1).

Lemma 3.4.1 (Parallel transport to the identity). *Given $\boldsymbol{\Sigma}, \boldsymbol{\Sigma}' \in \mathbb{S}_P^{++}$, the parallel transport of $\boldsymbol{\Sigma}'$ along the geodesic from $\boldsymbol{\Sigma}$ to the identity $\boldsymbol{I}_P$ at $\alpha \in [0,1]$ is*

$$\mathrm{PT}\left(\boldsymbol{\Sigma}', \boldsymbol{\Sigma}, \alpha\right) \triangleq \boldsymbol{\Sigma}^{-\alpha/2} \boldsymbol{\Sigma}' \boldsymbol{\Sigma}^{-\alpha/2} \,.$$

Learning on $\mathbb{S}_P^{++}$

In (1.15) we introduced a classical non-linear feature extraction [Barachant et al., 2012, Mellot et al., 2023, Bonet et al., 2023] of a dataset $\{\boldsymbol{\Sigma}_i\}_{i=1}^N$, with $\overline{\boldsymbol{\Sigma}}$ the Riemannian mean of the dataset. As the logarithmic mapping (1.12) at the identity is simply the matrix logarithm, this feature extraction can be written as the

combination of parallel transport and logarithmic mapping at the identity:

$$\phi\left(\boldsymbol{\Sigma}_i, \overline{\boldsymbol{\Sigma}}\right) = \operatorname{uvec}\left(\log\left(\overline{\boldsymbol{\Sigma}}^{-1/2}\boldsymbol{\Sigma}_i\overline{\boldsymbol{\Sigma}}^{-1/2}\right)\right) \tag{3.1}$$

$$= \operatorname{uvec}\left(\log_{\boldsymbol{I}_P}\left(\mathrm{PT}\left(\boldsymbol{\Sigma}_i, \overline{\boldsymbol{\Sigma}}, 1\right)\right)\right) \in \mathbb{R}^{P(P+1)/2} \tag{3.2}$$

where uvec vectorizes the upper triangular part with off-diagonal elements multiplied by $\sqrt{2}$ to preserve the norm.

In the multi-source domain adaptation scenario, we have access to K labeled source domains, each consisting of N_k covariance matrices, along with their corresponding response values, denoted by $\{(\boldsymbol{\Sigma}_{k,i}, y_{k,i})\}_{i=1}^{N_k}$. Correcting dataset shifts by re-centering all source datasets to a common point on the manifold [Zanini et al., 2018], corresponds to parallel transporting data $\{\boldsymbol{\Sigma}_{k,i}\}_{i=1}^{N_k}$ of each domain $k \in [\![1, K]\!]$ from its Riemannian mean $\overline{\boldsymbol{\Sigma}}_k$ to the identity,

$$\phi(\boldsymbol{\Sigma}_{k,i}, \overline{\boldsymbol{\Sigma}}_k) = \operatorname{uvec}\left(\log\left(\overline{\boldsymbol{\Sigma}}_k^{-1/2}\boldsymbol{\Sigma}_{k,i}\overline{\boldsymbol{\Sigma}}_k^{-1/2}\right)\right) . \tag{3.3}$$

This method is the go-to approach for reducing shifts of the covariance matrix distributions coming from different domains and has been applied successfully for brain-computer interfaces [Rodrigues et al., 2019, Yair et al., 2019] and age prediction from M/EEG data [Mellot et al., 2023], as presented in the previous chapter.

In Section 2.4.2, we presented how the re-center alignment is usually performed on $\mathbb{S}_P^{++}$. In particular, (3.3) generalized the re-center formulation for the multi-source context in order to account for the shifts of each domain. However, this operator can only work if the variability between domains is considered as noise. As explained earlier, we are interested in shifts in both features and the response variable. Thus, (3.3) discards shift coming from the response variable and hence harms the performance of the predictive model. Based on the Lemma 3.4.1, we propose to parallel transport features to any point on the geodesic between a domain-specific Riemannian mean $\overline{\boldsymbol{\Sigma}}_k$ and the identity. Indeed, `GOPSA` parallel transports $\boldsymbol{\Sigma}_{k,i}$ on this geodesic with $\alpha \in [0, 1]$ and then applies the Riemannian logarithmic mapping (1.12) at the identity,

$$\phi(\boldsymbol{\Sigma}_{k,i}, \overline{\boldsymbol{\Sigma}}_k, \alpha) \triangleq \operatorname{uvec}\left(\log_{\boldsymbol{I}_P}\left(\mathrm{PT}\left(\boldsymbol{\Sigma}_{k,i}, \overline{\boldsymbol{\Sigma}}_k, \alpha\right)\right)\right) \tag{3.4}$$

$$= \operatorname{uvec}\left(\log\left(\overline{\boldsymbol{\Sigma}}_k^{-\alpha/2}\boldsymbol{\Sigma}_{k,i}\overline{\boldsymbol{\Sigma}}_k^{-\alpha/2}\right)\right) . \tag{3.5}$$

This allows each domain to undergo parallel transport to a certain degree, effectively moving it toward the identity.

3.4.2 Train-time

GOPSA aims to learn simultaneously features from (3.4) and a regression model. To do so, we solve the following optimization problem

$$\underset{\substack{\boldsymbol{\beta}_{\mathcal{S}} \in \mathbb{R}^{P(P+1)/2} \\ \boldsymbol{\alpha}_{\mathcal{S}} \in [0,1]^K}}{\text{minimize}} \sum_{k=1}^{K} \sum_{i=1}^{N_k} \left(y_{k,i} - \boldsymbol{\beta}_{\mathcal{S}}^{\top} \phi \left(\boldsymbol{\Sigma}_{k,i}, \overline{\boldsymbol{\Sigma}}_k, \alpha_k \right) \right)^2 \tag{3.6}$$

with $\boldsymbol{\alpha}_{\mathcal{S}} = [\alpha_1, \ldots, \alpha_K]^{\top}$. This cost function is decomposed into three key aspects. First, covariance matrices undergo parallel transported using Lemma 3.4.1 to account for shifts between domains. Second, they are vectorized, and a linear regression predicts the output variable from these vectorized features. Third, the coefficients of the linear regression $\boldsymbol{\beta}_{\mathcal{S}}$ and the $\boldsymbol{\alpha}_{\mathcal{S}}$ are learned jointly so that the predictor is adapted to the parallel transport and reciprocally. Besides, to enforce the constraint on $\boldsymbol{\alpha}_{\mathcal{S}}$, we re-parameterize it using the sigmoid function, which defines a bijection between $\mathbb{R}$ and $(0, 1)$, thereby ensuring that the resulting $\boldsymbol{\alpha}_{\mathcal{S}}$ values lie within the desired range: $\alpha_k = \sigma(\gamma_k) \triangleq (1 + \exp(-\gamma_k))^{-1}$. Thus, the constrained problem (3.6) can be formulated as the following unconstrained optimization problem

$$\underset{\substack{\boldsymbol{\beta}_{\mathcal{S}} \in \mathbb{R}^{P(P+1)/2} \\ \boldsymbol{\gamma}_{\mathcal{S}} \in \mathbb{R}^{K}}}{\text{minimize}} \sum_{k=1}^{K} \sum_{i=1}^{N_k} \left(y_{k,i} - \boldsymbol{\beta}_{\mathcal{S}}^{\top} \phi \left(\boldsymbol{\Sigma}_{k,i}, \overline{\boldsymbol{\Sigma}}_k, \sigma(\gamma_k) \right) \right)^2, \tag{3.7}$$

with $\boldsymbol{\gamma}_{\mathcal{S}} = [\gamma_1, \ldots, \gamma_K]^{\top}$.

Let us define the matrix $\boldsymbol{Z}_{\mathcal{S}}(\boldsymbol{\gamma}) \in \mathbb{R}^{N_{\mathcal{S}} \times P(P+1)/2}$, with $N_{\mathcal{S}} = \sum_{k=1}^{K} N_k$, as the concatenation of the source data, where each row corresponds to a feature vector:

$$\boldsymbol{Z}_{\mathcal{S}}(\boldsymbol{\gamma}) = \left[\phi \left(\boldsymbol{\Sigma}_{1,1}, \overline{\boldsymbol{\Sigma}}_1, \sigma(\gamma_1) \right), \ldots, \phi \left(\boldsymbol{\Sigma}_{K,N_K}, \overline{\boldsymbol{\Sigma}}_K, \sigma(\gamma_K) \right) \right]^{\top}. \tag{3.8}$$

In the same manner, the source labels are concatenated to $\boldsymbol{y}_{\mathcal{S}} = [y_{1,1}, \ldots, y_{K,N_K}]^{\top} \in \mathbb{R}^{N_{\mathcal{S}}}$. Given a fixed $\boldsymbol{\gamma}_{\mathcal{S}}$, the problem (3.7) is solved with the ordinary least squares estimator. In practice, we choose to regularize the estimation of the linear regression with a Ridge penalty. Thus, (3.7) is rewritten as

$$\begin{aligned} \boldsymbol{\gamma}_{\mathcal{S}}^{\star} \triangleq \underset{\boldsymbol{\gamma} \in \mathbb{R}^K}{\arg\min} & \left\{ \mathcal{L}_{\mathcal{S}}(\boldsymbol{\gamma}) \triangleq \| \boldsymbol{y}_{\mathcal{S}} - \boldsymbol{Z}_{\mathcal{S}}(\boldsymbol{\gamma}) \boldsymbol{\beta}_{\mathcal{S}}^{\star}(\boldsymbol{\gamma}) \|_2^2 \right\} \\ & \text{subject to} \quad \boldsymbol{\beta}_{\mathcal{S}}^{\star}(\boldsymbol{\gamma}) \triangleq \boldsymbol{Z}_{\mathcal{S}}(\boldsymbol{\gamma})^{\top} (\lambda \boldsymbol{I}_N + \boldsymbol{Z}_{\mathcal{S}}(\boldsymbol{\gamma}) \boldsymbol{Z}_{\mathcal{S}}(\boldsymbol{\gamma})^{\top})^{-1} \boldsymbol{y}_{\mathcal{S}} \ , \end{aligned} \tag{3.9}$$

where $\boldsymbol{\beta}_{\mathcal{S}}^{\star}(\boldsymbol{\gamma}) \in \mathbb{R}^{P(P+1)/2}$ are the Ridge estimated coefficients given a fixed $\boldsymbol{\gamma}$ and $\lambda > 0$ is the regularization hyperparameter. The problem (3.9) is efficiently solved with any gradient-based solver [Nocedal and Wright, 1999].

The train-time of GOPSA is summarized in Algorithm 1. The proposed training algorithm begins by calculating the Riemannian mean of covariance matrices for each domain k. It then iteratively optimizes the parameters $\boldsymbol{\gamma}_{\mathcal{S}}$ by computing the feature matrix (3.8), determining Ridge regression coefficients (3.9), and updating $\boldsymbol{\gamma}_{\mathcal{S}}$ using a gradient descent step on the loss function (3.9) until convergence. The output result is the optimized Ridge regression coefficients. For clarity of presentation, Algorithm 1 employs a gradient descent. In practice, we use L-BFGS and obtain the gradient using automatic differentiation through the Ridge solution that is plugged into the loss in (3.9).

3.4.3 Test-time

At test-time, we now have a fitted linear model on source data with coefficients $\boldsymbol{\beta}^{\star}_{\mathcal{S}}(\boldsymbol{\gamma}^{\star}_{\mathcal{S}})$. The goal is to adapt a new target domain $(\boldsymbol{\Sigma}_{\mathcal{T},i})_{i=1}^{N_{\mathcal{T}}}$ for which the average outcome $\bar{y}_{\mathcal{T}}$ is assumed to be known. First, let us define the matrix $\boldsymbol{Z}_{\mathcal{T}}(\gamma) \in \mathbb{R}^{N_{\mathcal{T}} \times P(P+1)/2}$ as the concatenation of the target data

$$\boldsymbol{Z}_{\mathcal{T}}(\gamma) = \left[\phi\left(\boldsymbol{\Sigma}_{\mathcal{T},1}, \overline{\boldsymbol{\Sigma}}_{\mathcal{T}}, \sigma(\gamma)\right), \ldots, \phi\left(\boldsymbol{\Sigma}_{\mathcal{T},N_{\mathcal{T}}}, \overline{\boldsymbol{\Sigma}}_{\mathcal{T}}, \sigma(\gamma)\right)\right]^{\top} . \tag{3.10}$$

Then, GOPSA adapts to this new target domain by minimizing the error between $\bar{y}_{\mathcal{T}}$ and its estimation computed with the fitted linear model. This minimization is performed with respect to $\gamma_{\mathcal{T}} \in \mathbb{R}$ that parametrizes the parallel transport of the target domain, i.e.,

$$\gamma^{\star}_{\mathcal{T}} = \underset{\gamma \in \mathbb{R}}{\arg\min} \left\{ \mathcal{L}_{\mathcal{T}}(\gamma) \triangleq \left(\bar{y}_{\mathcal{T}} - \frac{1}{N_{\mathcal{T}}} \mathbf{1}_{N_{\mathcal{T}}}^{\top} \boldsymbol{Z}_{\mathcal{T}}(\gamma) \boldsymbol{\beta}^{\star}_{\mathcal{S}}(\boldsymbol{\gamma}^{\star}_{\mathcal{S}}) \right)^2 \right\} . \tag{3.11}$$

Finally, the predictions on the target domain are

$$\widehat{\boldsymbol{y}}_{\mathcal{T}} = \boldsymbol{Z}_{\mathcal{T}}(\gamma^{\star}_{\mathcal{T}}) \boldsymbol{\beta}^{\star}_{\mathcal{S}}(\boldsymbol{\gamma}^{\star}_{\mathcal{S}}) \in \mathbb{R}^{N_{\mathcal{T}}} . \tag{3.12}$$

The test-time procedure of GOPSA is summarized in Algorithm 2. The algorithm begins by calculating the Riemannian mean of the target covariance matrices $\{\boldsymbol{\Sigma}_{\mathcal{T},i}\}_{i=1}^{N_{\mathcal{T}}}$. It then iteratively optimizes the parameter $\gamma_{\mathcal{T}}$ by computing the feature matrix (3.10), the derivative of the loss function (3.11), and updating $\gamma_{\mathcal{T}}$ using a gradient descent step until convergence. The algorithm determines the estimated target outcomes, $\widehat{\boldsymbol{y}}_{\mathcal{T}}$, by using the optimized $\gamma^{\star}_{\mathcal{T}}$ on the feature matrix, combined with the pre-trained regression coefficients $\boldsymbol{\beta}^{\star}_{\mathcal{S}}(\boldsymbol{\gamma}^{\star}_{\mathcal{S}})$. The output result is the predicted target outcomes $\widehat{\boldsymbol{y}}_{\mathcal{T}}$. It should be noted that, once again, for clarity of presentation, Algorithm 2 employs a gradient descent, but other derivative-based optimization methods can be used.

Algorithm 1: Train-Time `GOPSA`

Input: For all $k \in [\![1, K]\!]$, $\{(\boldsymbol{\Sigma}_{k,i}, y_{k,i})\}_{i=1}^{N_k}$, initialization of $\boldsymbol{\gamma}_{\mathcal{S}}$, step-sizes $\{\xi_t\}_{t\geq 1}$

for $k = 1 \to K$ **do**

  $\overline{\boldsymbol{\Sigma}}_k \leftarrow$ Riemannian mean of $\{\boldsymbol{\Sigma}_{k,i}\}_{i=1}^{N_k}$

end

$t \leftarrow 1$

while *not converged* **do**

  $\boldsymbol{Z}_{\mathcal{S}}(\boldsymbol{\gamma}_{\mathcal{S}}) \leftarrow$ Compute features with (3.8)

  $\boldsymbol{\beta}^{\star}_{\mathcal{S}}(\boldsymbol{\gamma}_{\mathcal{S}}) \leftarrow$ Compute Ridge coeff. with (3.9)

  $\nabla\mathcal{L}_{\mathcal{S}}(\boldsymbol{\gamma}_{\mathcal{S}}) \leftarrow$ Compute loss gradient of (3.9)

  $\boldsymbol{\gamma}_{\mathcal{S}} \leftarrow \boldsymbol{\gamma}_{\mathcal{S}} - \xi_t \nabla\mathcal{L}_{\mathcal{S}}(\boldsymbol{\gamma}_{\mathcal{S}})$

  $t \leftarrow t + 1$

end

return $\boldsymbol{\beta}^{\star}_{\mathcal{S}}(\boldsymbol{\gamma}^{\star}_{\mathcal{S}})$

Algorithm 2: Test-Time `GOPSA`

Input: $\{\boldsymbol{\Sigma}_{\mathcal{T},i}\}_{i=1}^{N_{\mathcal{T}}}$, mean outcome value $\bar{y}_{\mathcal{T}}$, initialization of $\gamma_{\mathcal{T}}$, trained Ridge coeff. $\boldsymbol{\beta}^{\star}_{\mathcal{S}}(\boldsymbol{\gamma}^{\star}_{\mathcal{S}})$, step-sizes $\{\xi_t\}_{t\geq 1}$

$\overline{\boldsymbol{\Sigma}}_{\mathcal{T}} \leftarrow$ Riemannian mean of $\{\boldsymbol{\Sigma}_{\mathcal{T},i}\}_{i=1}^{N_{\mathcal{T}}}$

$t \leftarrow 1$

while *not converged* **do**

  $\boldsymbol{Z}_{\mathcal{T}}(\gamma_{\mathcal{T}}) \leftarrow$ Compute features with (3.10)

  $\mathcal{L}'_{\mathcal{T}}(\gamma_{\mathcal{T}}) \leftarrow$ Compute loss derivative of (3.11)

  $\gamma_{\mathcal{T}} \leftarrow \gamma_{\mathcal{T}} - \xi_t \mathcal{L}'_{\mathcal{T}}(\gamma_{\mathcal{T}})$

  $t \leftarrow t + 1$

end

$\widehat{\boldsymbol{y}}_{\mathcal{T}} \leftarrow \boldsymbol{Z}_{\mathcal{T}}(\gamma^{\star}_{\mathcal{T}})\boldsymbol{\beta}^{\star}_{\mathcal{S}}(\boldsymbol{\gamma}^{\star}_{\mathcal{S}})$

return $\widehat{\boldsymbol{y}}_{\mathcal{T}}$

3.5 Empirical benchmarks

In this section, we built empirical benchmark to evaluate the performance of `GOPSA`. We first present the simulated data that we used to illustrate the relevance of our method when there is a joint distrbution shift of the data and the labels. Then, we present the EEG dataset that we used to evaluate the performance of `GOPSA` with real data from different recording sites. Finally, we present the baseline methods that are compared with `GOPSA`.

3.5.1 Simulated data

To generate simulated data, we used the generative model described in Section 2.3.1. The data are generated following the classical instantaneous mixing model:

$$\boldsymbol{x}_i(t) = \boldsymbol{A}\boldsymbol{s}_i(t) \tag{3.13}$$

where $\boldsymbol{x}_i(t) \in \mathbb{R}^P$ are the observed time-series, $\boldsymbol{s}_i(t) \in \mathbb{R}^P$ are the underlying signal of the neural generators and $\boldsymbol{A}$ is the mixing matrix whose columns are the observed spatial patterns of the neural generators. Furthermore, we use a log-linear model to construct y:

$$y_i = \beta_0 + \sum_{\ell=1}^{P} \beta_\ell \log(p_{\ell i}) \tag{3.14}$$

where $p_{\ell i} \in \mathbb{R}$ is the variance of the ℓ-th element of the underlying signal $\boldsymbol{s}_i(t)$ as introduced in Chapter 2 [Sabbagh et al., 2019, 2020, Mellot et al., 2023].

From this, we generate domains by applying two shifts. One on $\boldsymbol{X}$ that changes the mixing matrix, and one on y by shifting the variances. In addition, we wish to generate several domains with shifts on $\boldsymbol{X}$ and y. We denote the domains by k and the number of domains by K.

Simulated shift in the data distribution: We introduced a shift in the data distribution by applying a affine transformation to the covariance matrices $\boldsymbol{\Sigma}$ of $\boldsymbol{x}$:

$$\boldsymbol{\Sigma} \mapsto \boldsymbol{B}_k^{\xi} \boldsymbol{\Sigma} \boldsymbol{B}_k^{\xi} \tag{3.15}$$

with $\boldsymbol{B}_k \in \mathbb{S}_P^{++}$ and $\xi \geq 0$ controlling the amplitude of the shift.

Simulated shift in the label distribution: We introduced a shift in the label distribution by modifying the variance of the underlying signal $p_{\ell i}$:

$$p_{\ell i} \mapsto p_{\ell i}^{1+k\xi} \tag{3.16}$$

with $\xi \geq 0$ still controlling the amplitude of the shift. Thus, the distribution of y is shifted per domain because of the log-linear relationship of (3.14). It should be noted that $\boldsymbol{\beta}$ is kept constant across domains.

3.5.2 EEG dataset

Dataset

The HarMNqEEG dataset [Li et al., 2022] was used for our numerical experiments. This dataset includes EEG recordings collected from 1564 participants across 14 different study sites, distributed across 9 countries. In our analysis, we consider each study site as a distinct domain. Figure 3.2 provides detailed demographic information. The EEG data were recorded with the same montage of 19 channels of the 10/20 International Electrodes Positioning System. The dataset provides pre-computed cross-spectral tensors for each participant rather than raw data, and anonymized metadata including the age and the sex of the participants. More precisely, the shared data consists of cross-spectral matrices with a frequency range of 1.17 Hz to 19.14 Hz, sampled at a resolution of 0.39 Hz.

Pre-processing

A standardized recording protocol was enforced to ensure the consistency across EEG recording of the dataset. In addition to recording constraints, this protocol included artifact cleaning procedures. The cross-spectrum were computed using Bartlett's method (See Appendix B.2). Our pre-processing steps were guided by the pre-processing pipeline outlined in Li et al. [2022]. First, we performed a common average reference (CAR) on all cross-spectrum (See Appendix B.2) as different EEG references were used across domains. Subsequently, we extracted the real part of the cross-spectral tensor to obtain co-spectrum tensors containing frequency-specific covariance estimates along the frequency spectrum. Due to the linear dependence between channels introduced by the CAR, the covariance matrices are rank deficient. To address this, we applied a shrinkage regularization with a coefficient of 10^{-5} to the data. Additionally, we implemented a global-scale factor (GSF) correction, which compensates for amplitude variations between EEG recordings by scaling the covariance matrices with a subject-specific factor [Hernández et al., 1994, Li et al., 2022] (See Appendix B.2). Following these pre-processing steps, we obtained a set of 49 covariance matrices for each EEG recording, with each matrix corresponding to a specific frequency bin of the EEG signal. This pre-processed co-spectrum served as the input data for our domain adaptation study.

Performance evaluation and hyperparameter selection

To evaluate the performance of the compared methods, we conducted experiments across several combinations of source and target sites. We selected source domains such that the union distribution of their predictive variable y encompasses a broad age range. All remaining sites were assigned as target domains. For each source-target combination we performed a stratified shuffle split approach with 100

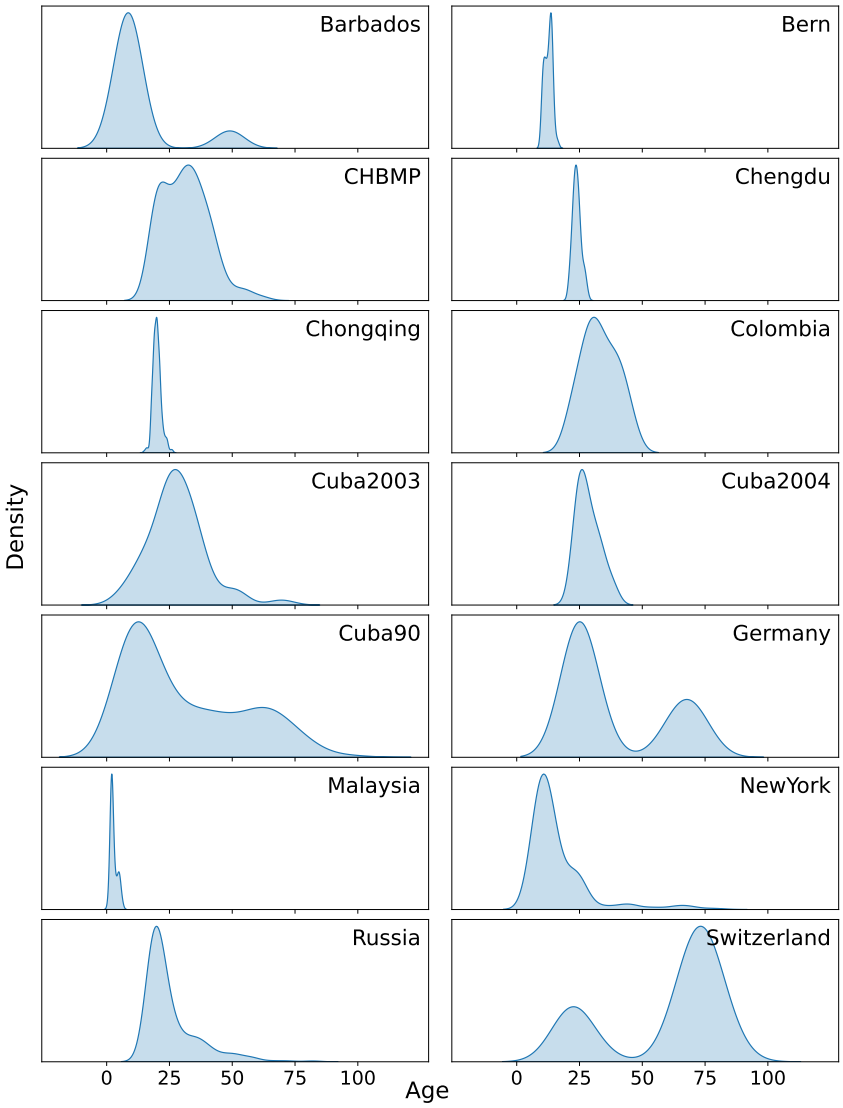

Figure 3.2: **Age distribution of the 14 sites of the HarMNqEEG dataset [Li et al., 2022].** The distributions are represented with a kernel density estimate. The y-scales are not shared for visualization purpose.

repetitions on the target data. Stratification was based on the recording sites to ensure that each split contained a balanced proportion of participants from each site. The regularization parameter λ in Ridge regression was selected with a nested cross-validation (grid search) over a logarithmic grid of values from 10^{-1} to 10^{5}.

3.5.3 Baseline methods

To evaluate the benefit of `GOPSA`, we compared it against four baselines. The four baseline methods used in this work are detailed in the following. For every methods we have access to K labeled source domains, each including N_k covariance matrices and their corresponding variables of interest $(\boldsymbol{\Sigma}_{k,i}, y_{k,i})_{i=1}^{N_k}$. The method `DO Dummy` and the mixed-effects model baseline `DO Intercept` both access the mean value $\bar{y}_{\mathcal{T}}$ of the target domain variable to predict. We remind that as the dataset used in the empirical benchmarks is constituted of several frequency bands, each method is applied to each frequency band independently and then computed vectors are concatenated. For each baseline method, the regression task was performed with Ridge regression.

Domain-aware dummy model (`DO Dummy`)

As `GOPSA` requires access to the mean $\bar{y}_k$ of each domain, we used a domain-aware dummy model predicting always the mean $\bar{y}_k$ of each domain.

No re-center / No domain adaptation (`No DA`)

This second baseline method involves applying the regression pipeline outlined in Sabbagh et al. [2019, 2020] without any re-centering. In this setup, all covariance matrices are projected to the tangent space at the source geometric mean $\overline{\boldsymbol{\Sigma}}$ computed from all source points, no matter their recording sites.

The covariance matrices are used as input of the regression pipeline without any re-centering. First, the geometric mean of the source matrices is computed:

$$\overline{\boldsymbol{\Sigma}}_{\mathcal{S}} \triangleq \underset{\boldsymbol{\Sigma} \in \mathbb{S}_P^{++}}{\arg\min} \sum_{k=1}^{K} \sum_{i=1}^{N_k} \delta_R(\boldsymbol{\Sigma}, \boldsymbol{\Sigma}_{k,i})^2. \tag{3.17}$$

Then, source feature vectors are extracted with:

$$\phi(\boldsymbol{\Sigma}_{k,i}, \overline{\boldsymbol{\Sigma}}_{\mathcal{S}}) = \operatorname{uvec}\left(\log\left(\overline{\boldsymbol{\Sigma}}_{\mathcal{S}}^{-1/2} \boldsymbol{\Sigma}_{k,i} \overline{\boldsymbol{\Sigma}}_{\mathcal{S}}^{-1/2}\right)\right) \in \mathbb{R}^{P(P+1)/2}. \tag{3.18}$$

Finally, the target feature vectors are extracted from the target data $(\boldsymbol{\Sigma}_{\mathcal{T},i})_{i=1}^{N_{\mathcal{T}}}$ with:

$$\phi(\boldsymbol{\Sigma}_{\mathcal{T},i}, \overline{\boldsymbol{\Sigma}}_{\mathcal{S}}) = \operatorname{uvec}\left(\log\left(\overline{\boldsymbol{\Sigma}}_{\mathcal{S}}^{-1/2} \boldsymbol{\Sigma}_{\mathcal{T},i} \overline{\boldsymbol{\Sigma}}_{\mathcal{S}}^{-1/2}\right)\right) \in \mathbb{R}^{P(P+1)/2}. \tag{3.19}$$

Re-center to a common reference point (`Re-center`)

As introduced in Section 3.4.1, a common transfer learning approach is a Riemannian re-centering of all domains to a common point on the manifold [Zanini et al., 2018, Mellot et al., 2023]. This baseline thus correspond to re-centering each domain k, source and target, independently by whitening them by their respective geometric mean $\overline{\boldsymbol{\Sigma}}_k$.

Domains are re-centered to a common reference point, here we decided to use the identity. First, the geometric mean of each domain is computed:

$$\overline{\boldsymbol{\Sigma}}_k \triangleq \underset{\boldsymbol{\Sigma}\in\mathbb{S}_P^{++}}{\arg\min} \sum_{i=1}^{N_k} \delta_R(\boldsymbol{\Sigma}, \boldsymbol{\Sigma}_{k,i})^2 \quad . \tag{3.20}$$

Then, feature vectors are extracted using the specific geometric mean of each domain:

$$\phi(\boldsymbol{\Sigma}_{k,i}, \overline{\boldsymbol{\Sigma}}_k) = \text{uvec}\left(\log\left(\overline{\boldsymbol{\Sigma}}_k^{-1/2}\boldsymbol{\Sigma}_{k,i}\overline{\boldsymbol{\Sigma}}_k^{-1/2}\right)\right) \in \mathbb{R}^{P(P+1)/2} \quad . \tag{3.21}$$

Covariance matrices of the target domain $(\boldsymbol{\Sigma}_{\mathcal{T},i})_{i=1}^{N_\mathcal{T}}$ are also re-centered to the identity using their geometric mean :

$$\overline{\boldsymbol{\Sigma}}_\mathcal{T} \triangleq \underset{\boldsymbol{\Sigma}\in\mathbb{S}_P^{++}}{\arg\min} \sum_{i=1}^{M} \delta_R(\boldsymbol{\Sigma}, \boldsymbol{\Sigma}_{\mathcal{T},i})^2 \tag{3.22}$$

$$\phi(\boldsymbol{\Sigma}_{\mathcal{T},i}, \overline{\boldsymbol{\Sigma}}_\mathcal{T}) = \text{uvec}\left(\log\left(\overline{\boldsymbol{\Sigma}}_\mathcal{T}^{-1/2}\boldsymbol{\Sigma}_i\overline{\boldsymbol{\Sigma}}_\mathcal{T}^{-1/2}\right)\right) \in \mathbb{R}^{P(P+1)/2}. \tag{3.23}$$

Re-scale to a common dispersion (`Re-scale`)

This baseline method consists in first re-centering the domains to a common reference point as done in `Re-center` with (3.21) and (3.23). Then re-scaling the covariance matrices distribution of each domain so that they have the same dispersion. The dispersion d_k of each domain are computed as previoulsy presented in Chapter 2, Section 2.4.2 with (2.11), and are then used to re-scale the distribution of each domain to one:

$$\boldsymbol{\Sigma}_{k,i} \mapsto \boldsymbol{\Sigma}_{k,i}^{1/\sqrt{d_k}} \tag{3.24}$$

Finaly the feature vectors are extracted from the recentered and rescaled data:

$$\phi(\boldsymbol{\Sigma}_{k,i}, \overline{\boldsymbol{\Sigma}}_k) = \text{uvec}\left(\log\left((\overline{\boldsymbol{\Sigma}}_k^{-1/2}\boldsymbol{\Sigma}_{k,i}\overline{\boldsymbol{\Sigma}}_k^{-1/2})^{1/\sqrt{d_k}}\right)\right) \in \mathbb{R}^{P(P+1)/2} \tag{3.25}$$

The same transformations are applied to the target domain with its dispersion $d_\mathcal{T}$:

$$\boldsymbol{\Sigma}_{\mathcal{T},i} \mapsto \boldsymbol{\Sigma}_{\mathcal{T},i}^{1/\sqrt{d_\mathcal{T}}} \tag{3.26}$$

$$\phi(\boldsymbol{\Sigma}_{\mathcal{T},i}, \overline{\boldsymbol{\Sigma}}_\mathcal{T}) = \text{uvec}\left(\log\left((\overline{\boldsymbol{\Sigma}}_\mathcal{T}^{-1/2}\boldsymbol{\Sigma}_{\mathcal{T},i}\overline{\boldsymbol{\Sigma}}_\mathcal{T}^{-1/2})^{1/\sqrt{d_\mathcal{T}}}\right)\right) \in \mathbb{R}^{P(P+1)/2} \tag{3.27}$$

For `No DA`, `Re-center` and `Re-scale`, the regression task was performed using a Ridge regression, which included an intercept:

$$\boldsymbol{\beta}^{\star}_{\mathcal{S}}, \beta^{\star}_{0,\mathcal{S}} = \underset{\substack{\boldsymbol{\beta}\in\mathbb{R}^{P(P+1)/2} \\ \beta_0\in\mathbb{R}}}{\arg\min} \sum_{k=1}^{K}\sum_{i=1}^{N_k} \left(y_{k,i} - \boldsymbol{\beta}^{\top}\boldsymbol{z}_{k,i} - \beta_0\right)^2 + \lambda \left\|\boldsymbol{\beta}\right\|_2^2 \tag{3.28}$$

where $\boldsymbol{z}_{k,i}$ is computed with (3.18), (3.21) or (3.25). The predicted values were computed as:

$$\widehat{y}_{\mathcal{T},i} = (\boldsymbol{\beta}^{\star}_{\mathcal{S}})^{\top}\boldsymbol{z}_{\mathcal{T},i} + \beta^{\star}_{0,\mathcal{S}} \tag{3.29}$$

where $\boldsymbol{z}_{\mathcal{T},i}$ is computed with (3.19), (3.23) or (3.27).

Domain-aware intercept (`DO Intercept`)

This method consists in fitting one intercept β_0 per domain. In practice since we assume to know $\bar{y}_{\mathcal{T}}$, we correct the predicted values so that their mean is equal to $\bar{y}_{\mathcal{T}}$. This approach is in line with defining mixed-effects models on the Riemannian manifold [Li et al., 2022].

In addition to the K labeled source domains, we assume to have access to the mean of the variable to predict of the target domain $\bar{y}_{\mathcal{T}}$. At train-time, we fit a Ridge regression with a specific intercept for each domain

$$\boldsymbol{\beta}^{\star}_{\mathcal{S}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^{P(P+1)/2}}{\arg\min} \sum_{k=1}^{K}\sum_{i=1}^{N_k} \left(y_{k,i} - \boldsymbol{\beta}^{\top}\phi(\boldsymbol{\Sigma}_{k,i}, \overline{\boldsymbol{\Sigma}}_{\mathcal{S}}) - \bar{y}_k\right)^2 + \lambda \left\|\boldsymbol{\beta}\right\|_2^2 \ . \tag{3.30}$$

Then, at test-time, we fit a new intercept $\beta_{0,\mathcal{T}}$ using the target features:

$$\phi\left(\boldsymbol{\Sigma}_{\mathcal{T},i}, \overline{\boldsymbol{\Sigma}}_{\mathcal{S}}\right) = \mathrm{uvec}\left(\log\left(\overline{\boldsymbol{\Sigma}}_{\mathcal{S}}^{-1/2}\boldsymbol{\Sigma}_{\mathcal{T},i}\overline{\boldsymbol{\Sigma}}_{\mathcal{S}}^{-1/2}\right)\right) \in \mathbb{R}^{P(P+1)/2}. \tag{3.31}$$

The fitted intercept is

$$\beta_{0,\mathcal{T}} = \bar{y}_{\mathcal{T}} - \frac{1}{N_{\mathcal{T}}}\sum_{i=1}^{N_{\mathcal{T}}} (\boldsymbol{\beta}^{\star}_{\mathcal{S}})^{\top}\phi\left(\boldsymbol{\Sigma}_{\mathcal{T},i}, \overline{\boldsymbol{\Sigma}}_{\mathcal{S}}\right) \tag{3.32}$$

and the predictions are

$$\widehat{y}_{\mathcal{T},i} = (\boldsymbol{\beta}^{\star}_{\mathcal{S}})^{\top}\phi\left(\boldsymbol{\Sigma}_{\mathcal{T},i}, \overline{\boldsymbol{\Sigma}}_{\mathcal{S}}\right) + \beta_{0,\mathcal{T}} \ . \tag{3.33}$$

Deep learning `GREEN`

For this last method, we used the `GREEN` architecture [Paillard et al., 2024], which is a deep-learning architecture tailored for EEG applications like age prediction. It combines Gabor wavelets and Riemannian geometry to extract useful features

for predicting the desired biomedical outcome. Since the HarMNqEEG dataset consists of covariance matrices, we used the 'G2' variant of `GREEN`: the architecture is cropped to start at the covariance matrices level with pooling layers, so it does not include the Gabor wavelets part. This variant is an SPD network as the input data are SPD matrices [Huang and Van Gool, 2017]. Although `GREEN` has been evaluated on multiple datasets for various predictive tasks, it has not yet been applied in a domain adaptation context and does not include an adaptation layer.

We applied the domain-adaptation methods independently to each of the 49 frequency bins, resulting in 49 geometric means per domain, except for `GREEN`, which processes all frequency bands simultaneously. $\bar{y}_{\mathcal{T}}$ of each domain was estimated on target splits (50% of the data) that do not overlap with the evaluation target splits (50% remaining). Statistical inference for model comparisons was implemented with a corrected t-test following Nadeau and Bengio [1999]. Experiments with 100 repetitions and all site combinations have been run on a standard Slurm cluster for 12 hours with 250 CPU cores.

3.6 Results

Then, we evaluated the effectiveness of `GOPSA` in the context of age prediction from EEG spatial covariance matrices [Sabbagh et al., 2020, Engemann et al., 2022]. Each domain k corresponds to a distinct recording site, while each sample point i represents an EEG recording from a human participant. We compared the performance of `GOPSA` against four baseline methods detailed below. We employed three performance metrics to evaluate the prediction accuracy across different source/target domain splits: Spearman's ρ, quantifying correct ranking and robust to location and scale errors, as well as the commonly used coefficient of determination (R^2) score, and Mean Absolute Error (MAE), which are sensitive to location and scale errors. We present the detailed results on the publicly available dataset HarMNqEEG [Li et al., 2022].

3.6.1 Simulated data

Figure 3.3 presents the results of simulated experiments where shifts are applied on either $\boldsymbol{X}$, y, or both ($\boldsymbol{X}$, y) as presented in Section 3.5.1. All methods were evaluated in three simulation scenarios: shift in $\boldsymbol{X}$ only, shift in y only, and joint shift in $\boldsymbol{X}$ and y. The intensity of the shift was controlled by ξ in all scenarios. If there is no shift in $\boldsymbol{X}$, we observe that `No DA` perfectly estimates the y because the log-linear model is respected across domains even when the y distribution changes (Figure 3.3 B). The performance of `No DA` however drops when a shift in $\boldsymbol{X}$ is introduced (Figure 3.3 A and C). `Re-center` and `Re-scale` led to the same results as no scaling shift was applied in the simulation. Both were able to correct the shift in X, but performed poorly when a shift in y was added (Figure 3.3 B

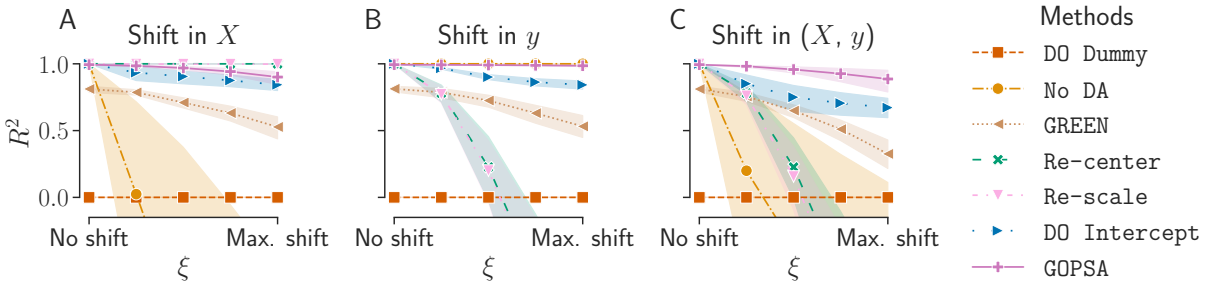

Figure 3.3: R^2 **scores ↑ for different methods on simulated data.** Performance is measured across 5 source domains and 1 target domain, with shifts controlled by ξ (0 to maximum). Data are generated 100 times, with 5 sensors and 300 covariance matrices per domain. The target domain is randomly selected between the 6 domains generated as presented in Section 3.5.1, with the remaining domains used as sources. **(A)** A shift is applied on the covariance matrices following (3.15). **(B)** A shift is applied on the variances following (3.16). **(C)** Both shifts from (3.15) and (3.16) are applied simultaneously.

and C). `GREEN` notably showed consistant performance across all scenarios, and was relatively resistant to both types of shifts given it is not designed for domain adaptation. `DO Intercept` and `GOPSA` showed the best performance across all scenarios, with an advantage for `GOPSA`. The interest of `GOPSA` is to estimate this log-linear model with shifts in $(\boldsymbol{X}, y)$ per domain (Figure 3.3 C) which other methods were not able to do.

These experiments demonstrate the efficiency of the proposed method, in estimating shifts in $\boldsymbol{X}$ between domains, even in the presence of a shift in y, contrary to the baseline methods. Theoretically, based on the generative model of the simulated data, the data and outcome y are linked by a log-linear relationship. This implies that, knowing the shift in $\boldsymbol{X}$ for the target domain, predictions can be made even when y distributions do not overlap between the source and target. Since `GOPSA` estimates the target shift in $\boldsymbol{X}$ by minimizing $(\bar{y} - \hat{y}_i)^2$, it is capable of handling such scenarios.

3.6.2 EEG data

We computed benchmarks for five combinations of source sites and we displayed the results for the three metrics selected for performance evaluation, each colored box representing one method (Figure 3.4). A min-max normalization was applied to each site combinations separately across methods. We first conduced model comparisons in terms of absolute performance across all baselines (**A**). `No DA`, without domain specific re-centering, performed worse than `DO Dummy` in terms of R^2 score and MAE. `Re-center` and `Re-scale` led to lower performances across all metrics, which can be expected as the Riemannian mean is correlated with age

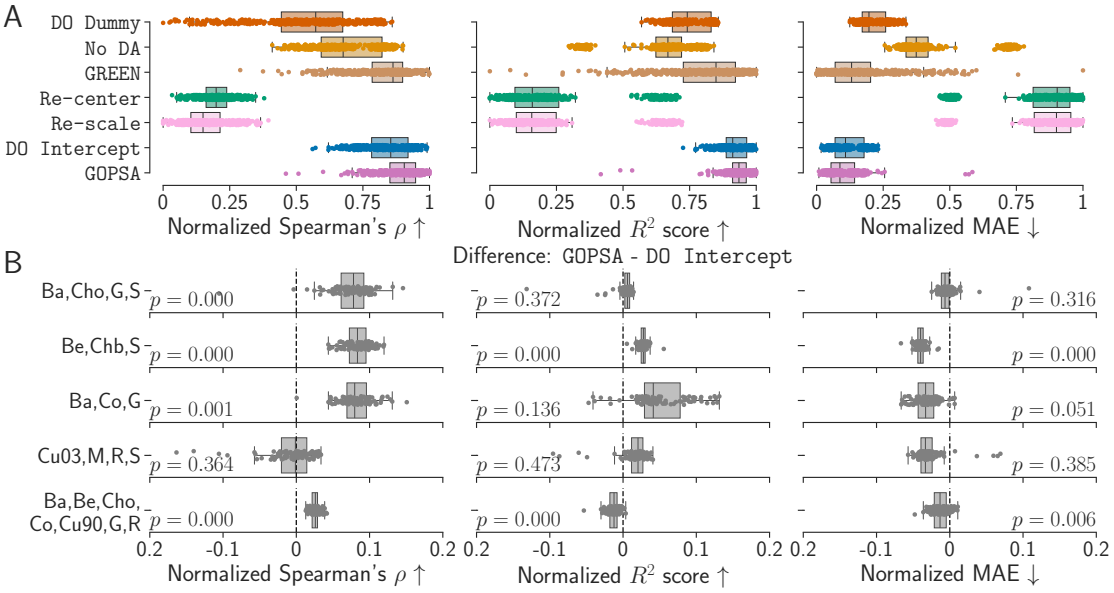

Figure 3.4: **Normalized performance of the different methods on several source-target combinations for three metrics.** Spearman's ρ ↑ (left), R^2 score ↑ (middle) and Mean Absolute Error ↓ (right). As a large variability in the score values was present between the site combinations, we applied a min-max normalization per combination to set the minimum score across all methods to 0 and the maximum score to 1. (**A**) Boxplot of the concatenated results for the three normalized scores. One point corresponds to one split of one site combination. (**B**) Boxplots of the difference between the normalized scores of `GOPSA` and `DO Intercept`. A row corresponds to one site combination, one point corresponds to one split. For each plot, the associated results of Nadeau's & Bengio's corrected t-test [Nadeau and Bengio, 1999] are displayed. A p-value lower than 0.05 indicates a significant difference between the two methods. Ba: Barbados, Be: Bern, Chb: CHBMP (Cuba), Co: Columbia, Cho: Chongqing, Cu03: Cuba2003, Cu90: Cuba90, G: Germany, M: Malaysia, R: Russia, S: Switzerland

Table 3.1: **Performance scores for different combinations of source sites.** The remaining sites were used as target domains.

Spearman's $\rho \uparrow$

Sites source	DO Dummy	No DA	GREEN	Re-center	DO Intercept	GOPSA
Ba,Cho,G,S	0.51 ± 0.04	0.63 ± 0.02	0.69 ± 0.03	0.52 ± 0.02	0.75 ± 0.02	**0.78 ± 0.02**
Be,Chb,S	0.58 ± 0.02	0.73 ± 0.01	**0.75 ± 0.02**	0.43 ± 0.02	0.68 ± 0.02	0.72 ± 0.02
Ba,Co,G	0.62 ± 0.02	0.64 ± 0.02	0.72 ± 0.02	0.42 ± 0.02	0.71 ± 0.02	**0.74 ± 0.02**
Cu03,M,R,S	0.62 ± 0.03	0.63 ± 0.01	0.70 ± 0.05	0.46 ± 0.02	**0.76 ± 0.02**	0.75 ± 0.04
Ba,Be,Cho, Co,Cu90,G,R	0.77 ± 0.02	0.79 ± 0.01	0.82 ± 0.01	0.44 ± 0.03	0.85 ± 0.01	**0.87 ± 0.01**
Mean	0.62 ± 0.03	0.68 ± 0.01	0.74 ± 0.03	0.45 ± 0.02	0.75 ± 0.02	**0.77 ± 0.02**

R^2 score $\uparrow$

Sites source	DO Dummy	No DA	GREEN	Re-center	DO Intercept	GOPSA
Ba,Cho,G,S	0.21 ± 0.02	0.06 ± 0.06	0.26 ± 0.33	-0.32 ± 0.10	0.57 ± 0.03	**0.58 ± 0.05**
Be,Chb,S	0.25 ± 0.02	-0.07 ± 0.08	0.39 ± 0.30	-1.36 ± 0.13	0.43 ± 0.03	**0.49 ± 0.03**
Ba,Co,G	0.48 ± 0.03	0.26 ± 0.03	0.47 ± 0.09	0.10 ± 0.03	0.60 ± 0.03	**0.64 ± 0.03**
Cu03,M,R,S	0.26 ± 0.02	0.26 ± 0.04	0.48 ± 0.13	-0.30 ± 0.07	0.51 ± 0.02	**0.51 ± 0.09**
Ba,Be,Cho, Co,Cu90,G,R	0.60 ± 0.03	0.54 ± 0.02	0.62 ± 0.10	0.14 ± 0.02	**0.76 ± 0.02**	0.75 ± 0.02
Mean	0.36 ± 0.03	0.21 ± 0.05	0.44 ± 0.19	-0.35 ± 0.07	0.57 ± 0.02	**0.59 ± 0.04**

MAE $\downarrow$

Sites source	DO Dummy	No DA	GREEN	Re-center	DO Intercept	GOPSA
Ba,Cho,G,S	9.25 ± 0.16	12.00 ± 0.21	9.08 ± 1.98	14.69 ± 0.24	7.69 ± 0.19	**7.61 ± 0.25**
Be,Chb,S	9.48 ± 0.14	11.83 ± 0.37	8.67 ± 2.30	22.48 ± 0.24	9.28 ± 0.20	**8.61 ± 0.20**
Ba,Co,G	9.42 ± 0.14	13.83 ± 0.46	9.44 ± 0.77	15.25 ± 0.45	8.77 ± 0.15	**8.50 ± 0.20**
Cu03,M,R,S	9.64 ± 0.17	10.98 ± 0.22	**8.53 ± 1.12**	16.50 ± 0.25	8.94 ± 0.18	8.75 ± 0.72
Ba,Be,Cho, Co,Cu90,G,R	10.37 ± 0.23	11.40 ± 0.31	9.53 ± 1.10	15.92 ± 0.45	8.53 ± 0.23	**8.40 ± 0.24**
Mean	9.63 ± 0.17	12.01 ± 0.31	9.05 ± 1.45	16.97 ± 0.33	8.64 ± 0.19	**8.37 ± 0.32**

in our problem setting Figure 3.1. Eventhough its architecture does not include an adaptation layer, `GREEN` reached better performance than the previous methods mentionned, but lacked consistency across site combinations and metrics with large variance especially for the R^2 score and MAE. For all scores, `DO Intercept` and `GOPSA` reached the best average performance with lower variance. A version of Figure 3.4 (**A**) without normalization is presented in Appendix B.3. As `DO intercept` and `GOPSA` showed overlapping performance distributions, we investigated their paired split-wise (non-rescaled) score differences (**B**). The site-specific differences of `GOPSA` scores minus `DO Intercept` are displayed with their associated p-values. For one site combination (Ba,Be,Cho,Co,Cu90,G,R), `DO Intercept` yielded higher R^2 scores, and no significant difference was found between the two methods for Ba,Co,G. Similarly, no significant difference was observed on Spearman's ρ results for Cu03,M,R,S. Overall, `GOPSA` significantly outperformed `DO Intercept` in five site combinations for MAE, four for Spearman's ρ and three for R^2 score. Detailed results for each source-target combination are presented in Appendix 3.1 for Spearman's ρ, R^2 score, and MAE. The bottom rows correspond to the mean performance of each method of all site combinations, and their average standard deviation (see Appendix B.4 for associated boxplots). We expected `GOPSA` to outperform the baseline methods (e.g. `DO Intercept`) whenever joint (X, y) shifts occur. In our experimental benchmark, `GOPSA` significantly outperformed the baseline methods in some site combinations, but not all. This allows us to assume that not all site combinations show joint shifts.

3.7 Model inspection

Next, we investigated the impact of the different re-centering approaches on the data Figure 3.5. Power spectrum densities (PSDs) were computed as the mean across sensors of the diagonals of the covariance matrices Riemannian mean for each site combination after `No DA`, `Re-center` and `GOPSA` (**A**). PSDs for `No DA` display the initial variability between sites without recentering (cf. Figure 3.1). `Re-center` resulted in flat PSDs because all data were re-centered to the identity. PSDs produced by `GOPSA` are flattened and more similar across sites compared to `No DA` without removing too much information, unlike the uneffective `Re-center method` (cf. Figure 3.4). The alpha values are inspected as a function of the site mean age (**B**). `Re-center` leads to alpha values all equal to one as all sites are re-centered to the identity. For `GOPSA`, we observed a linear relationship between alpha and the sites' mean age ($R^2 = 0.99$). This is a direct consequence of the optimization process in `GOPSA`, which thus can be regarded a geodesic intercept in a mixed-effects model. Overall, `GOPSA` effectively re-centered sites with younger participants closer to the identity matrix. Re-centering sites around a common point helped reduce the shift in X, while not placing all sites at the exact same reference point helped manage the shift in y, hence preserving the statistical associations between X and y.

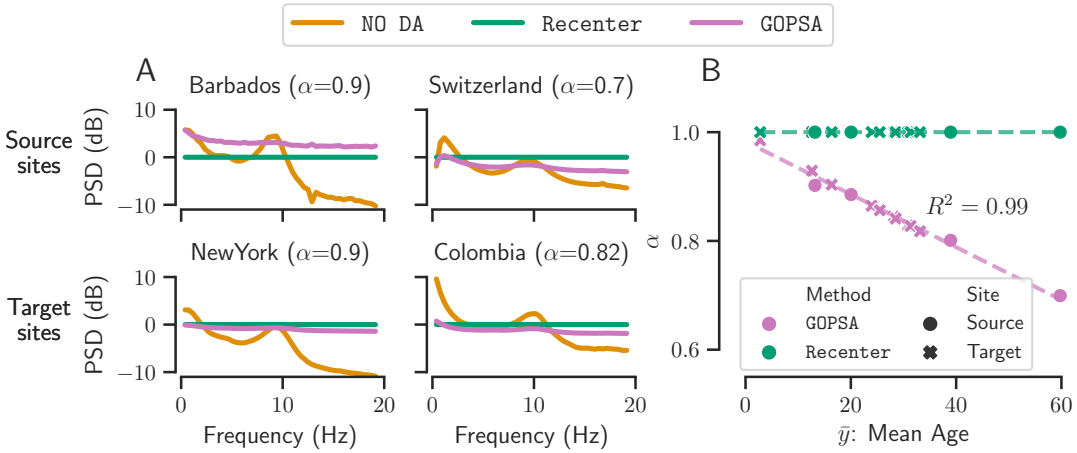

Figure 3.5: **Model inspection of `GOPSA` versus `No DA` and `Re-center`.** Power Spectral Densities (PSDs) and α values were computed on the source sites Barbados, Chongqing, Germany, and Switzerland. The remaining sites were used as target domains. (**A**) Mean PSDs computed across sensors for `No DA`, `Recenter` and `GOPSA` on two source (Barbados and Switzerland) and two target (New York and Columbia) sites. (**B**) α values versus site's mean age for `Re-center` and `GOPSA`. One point corresponds to one site. The coefficient of determination is reported for the `GOPSA` method.

3.8 Conclusion

We proposed a novel multi-source domain adaptation approach that adapts shifts in X and y simultaneously by learning jointly a domain specific re-centering operator and the regression model. `GOPSA` was specifically developed to handle joint shifts in the data distribution and the outcome distribution, as illustrated by the simulations in Figure 3.3.

`GOPSA` is a test-time method that does not require to retrain a model when a new domain is presented. `GOPSA` achieved state-of-the-art performance on the HarMNqEEG [Li et al., 2022] dataset with EEG from 14 recording sites and over 1500 participants. Our benchmarks showed a significant gain in performance for three different metrics in a majority of site combinations compared to baseline methods. `GOPSA` can thus be used by researchers as a decision rule to infer the presence of joint shifts and, hence, serve as a tool for data exploration and model interpretation. While we focused on shallow regression models, the implementation of `GOPSA` using PyTorch readily supports its inclusion in more complex Riemannian deep learning models [Huang and Van Gool, 2017, Wilson et al., 2024, Carrara et al., 2024, Paillard et al., 2024, Kobler et al., 2022]. This direction seems promising given our observation that GREEN – a simple deep net com-

bining Riemannian computation with a fully connected layer - already possessed some intrinsic robustness to data shifts. This may point at the capacity of the fully-connected layer to provide additional non-linear transformations that can accommodate the data-generating scenario in which continuous log-linear generators are modified in a discrete manner by site factors. More generally, it emphasizes the potential of complex nonlinear methods for domain adaptation, in line with a recent study on the same dataset reporting positive generalization results using a kernel method [Jarne et al., 2024]. Furthermore, although this work specifically addresses age prediction, the methodology is applicable to a broader range of regression analyses. While `GOPSA` necessitates knowledge or estimability of the average $\bar{y}$ per domain, this requirement aligns with that of mixed-effects models [Gelman, 2006, Hox, 1998, Yarkoni, 2022], which are extensively employed in biomedical statistics. By combining mixed-effects modeling with Riemannian geometry for EEG, `GOPSA` opens up various applications at the interface between machine learning and biostatistics, such as, biomarker exploration in large multicenter clinical trials [Rossetti et al., 2020, Vassallo et al., 2021, Vespa et al., 2020].

4 Shift due to different recording devices

Contents

In Chapter 2 and Chapter 3, we explored the reasons behind dataset shifts in M/EEG data, identifying that such shifts can manifest as changes in the data distribution, the label distribution, or both. Initially, we focused on how aligning the data distribution alone can enhance the generalization of machine learning models across different tasks and populations, under the assumption that label distributions remain consistent. Subsequently, in Chapter 3, we went further and tackled the joint shift in both data and label distributions. We proposed a novel method to align and combine datasets with significantly different age distributions in order to increase the size of the training set.

However, our analyses so far have been constrained to data recorded with the same EEG sensor configuration. The methods we presented did not address scenarios where features have different dimensions. Although the HarMNqEEG dataset includes a consistent sensor configuration across sites, allowing us to sidestep this issue, the TUAB and LEMON datasets in the first chapter have different sensor numbers and positions. We got around this problem by selecting a subset of common channels across datasets.

In this chapter, we shift our focus to contexts where EEG datasets have differing sensor configurations. Specifically, we address situations where the common channels across datasets are insufficient for performing the desired analytical tasks. To overcome this challenge, we propose using field interpolation to map EEG signals from various sensor locations onto a common template of fixed positions. We then compare the effectiveness of this approach against other state-of-the-art methods in the context of BCI applications. This strategy aims to harmonize datasets with varying electrode configurations, further enhancing the robustness and generalizability of EEG data analysis.

4.1 The problem of EEG data with different sensor configurations

In machine learning, dataset shift occurs when there are changes in the data and/or label distributions between the training and testing datasets, often due to variations in the environment, population, or recording devices. In the previous chapters, we investigated the impact of dataset shift on the performance of machine learning models applied to EEG data. We demonstrated that domain adaptation techniques can mitigate the effects of dataset shift and improve the generalization of models across different datasets. However, using different recording devices can also result in changes in the number and positions of electrodes. Indeed, there are many different devices for recording EEG data, each with its own electrode configuration and labeling. These changes can lead to difficulties in aligning data, extracting features, and building predictive models that are robust and generalizable across different datasets. Even though the recording procedure is different, the underlying phenomenon of interest is not dependent on the setup

and should be invariant to its variability. We can thus expect relevant information to be present in the data for any configuration.

Classical multivariate statistical analysis assumes that the training and testing data share the same feature space and dimensionality. The dimensionality assumption extends to classical domain adaptation approaches, including the methods used in the previous chapters of this thesis. In the case of real-life practical cases, such as EEG recordings with different sensor configurations, this assumption does not hold true. Two datasets recorded with different sensor configurations may not have the same dimensions or features, breaking classical approaches. Even within the same dataset, some electrodes might be malfunctioning or improperly placed for certain participants but not for others. To address this issue, we need to either find a way to match the different EEG configurations, or to extract features with the same dimensions across datasets that are invariant to the sensor configuration and are related to the physical phenomenon under study.

In this chapter, we propose to combine heterogeneous datasets in an unsupervised manner. The pipeline is composed of two steps: the first one to match the EEG positions across datasets, and the second one to align the statistical distributions of the data. The multivariate time series from EEG recordings done with different sensor configurations into signals with a common sensor configuration, matching the positions and dimensions across datasets. Then, the statistical distributions of the data are aligned to mitigate the dataset shift, and harmonized features are extracted. Predictions are finally made by training a model on several labeled datasets (sources) and then applied to an unseen and unlabeled dataset (target). The proposed approach to match both the EEG configurations and the data distributions is a test-time method that does not require access to the training data when applying the model to a new dataset.

We propose to map EEG channels to a template of fixed positions by leveraging the underlying physics of EEG data through field interpolation. This mapping can be applied to the raw data as well as to clean epochs, and before feature extraction. This step addresses the problem of different EEG sensor positions. To benchmark this approach other state-of-the-art methods on BCI applications, we use the same pipeline as in the previous chapters that represents EEG signals as covariance matrices. We leverage the Riemannian re-centering operator to align the statistical distributions of data from different source domains, mitigating dataset shift. In the end, we extract features from compatible statistical distributions and with compatible dimensions.

The rest of this chapter is organized as follows: Section 4.3 presents a literature review on domain adaptation methods that aim at dealing with EEG data with different sensor configurations. Section 4.4 describes the proposed method to match the EEG sensor positions by detailing how interpolation of EEG signals is performed. In Section 4.5, we present the experimental evaluation of the proposed method on six publicly available BCI datasets and report on the advantages of

field interpolation compared to other methods.

4.2 Contribution

The content of this chapter is an extended version of the following publication:

> A. Mellot, A. Collas, S. Chevallier, D. Engemann, A. Gramfort. Physics-informed and Unsupervised Riemannian Domain Adaptation for Machine Learning on Heterogeneous EEG Datasets. 32nd European Signal Processing Conference (EUSIPCO), 2024.

Field interpolation is commonly used to interpolate bad channels in MEG signals. Our contribution is to apply field interpolation in a domain adaptation setup to combine several EEG datasets with different sensor configurations. Spherical spline interpolation has already been investigated in this context, but not field interpolation.

The advantage of our method compared to other existing methods is that it is applied before feature extraction and can therefore be used with any machine learning or domain adaptation models. This approach ensures flexibility and broader applicability across various EEG analysis frameworks.

Python codes for the field interpolation for domain adaptation and for the following benchmark are available in:

`https://github.com/apmellot/EUSIPCO_2024.git`

4.3 Literature review

Heterogeneous domain adaptation is a branch of domain adaptation that deals with domains represented by different types of features. In the context of EEG data, this situation often arises when signals are recorded from different sensor configurations. This chapter focuses on this specific problem.

Few methods have been proposed to tackle the challenge of heterogeneous domain adaptation with EEG data from different sensor configurations. In the following sections, we review the most commonly used and recent methods in this area.

4.3.1 Dimension Reduction Approaches

A group of methods in heterogeneous domain adaptation is based on the idea of projecting data into a common lower-dimensional space where points from all

datasets have the same dimension and can be compared.

There are domain adaptation methods that use this principle to align data distributions across domains. For example, the TCA method [Pan et al., 2011] aims to find a subspace projection that preserves the shared attributes between domains. However, this method is not designed to handle domains with different feature dimensions.

Region Based Pooling (RBP) is a method that reduces data dimensions to match them across domains [Tveitstøl et al., 2024]. To achieve this, one idea is to pool channels from the same region to form a common region representation across datasets. The EEG montage is first split into several regions. Within each region, channels are pooled using various strategies such as average pooling or channel attention weighted average pooling. No clear advantage of this method compared to spherical spline interpolation was found. Therefore, we did not include this method in our benchmark. Additionally, another reason it was not included is that one of our datasets has only three sensors, which limits the number of regions that can be formed and, consequently, the effectiveness of the method.

The current common practice when dealing with EEG data from different sensor configurations is to select a subset of common channels across datasets. This is the method we chose to represent this group of approaches.

Common Channel Selection

Selecting EEG channels that are consistently present across all subjects and datasets is the first strategy that comes to mind, particularly effective when there is a sufficient number of shared channels to perform the desired analytical tasks [Wei et al., 2022, Mellot et al., 2023]. However, as the number of subjects and datasets increases, the variability in sensor positions due to the use of different recording devices becomes more pronounced. This can lead to a situation where there are too few common channels for the task, as was observed in our experimental evaluation, where only one common channel in the middle of the head (Cz) was insufficient for right/left-hand classification.

4.3.2 Imputation Approaches

Another group of methods in heterogeneous domain adaptation is based on imputation. The first step in these methods is to expand the data dimension so that all datasets have the same dimension. Usually, the final expanded dimension corresponds to the union of the feature types from all datasets. This expansion will lead to some missing values in the data. The second step involves using imputation techniques to fill in these missing values.

We investigate two types of imputation: a geometry-based approach linked to the representation of EEG signals as covariance matrices in the predictive pipeline, and a signal-based approach that can be compatible with other predictive modeling techniques.

Geometry-based imputation

Rodrigues et al. [2021] proposed to transform the data points into an expanded common space with an isometric transformation, which preserves the distance of the original data points in the new expanded space. We consider K datasets of different dimensionality (or different numbers of sensors) P_k with $k = 1, \ldots, K$. The expanded space dimension is $P_{\text{exp}} = \bigcup P_k$. The resulting expanded version $\boldsymbol{\Sigma}_i^{\uparrow}$ of a matrix $\boldsymbol{\Sigma}_i$ from the k-th dataset is:

$$\boldsymbol{\Sigma}_i^{\uparrow} = \begin{bmatrix} \boldsymbol{\Sigma}_i & \mathbf{0}_{P_k \times (P_{\text{exp}} - P_k)} \\ \mathbf{0}_{(P_{\text{exp}} - P_k) \times P_k} & \boldsymbol{I}_{(P_{\text{exp}} - P_k)} \end{bmatrix} \in \mathbb{R}^{P_{\text{exp}} \times P_{\text{exp}}} \tag{4.1}$$

This transformation is applied to all datasets to get covariances of size $d_{\text{exp}} \times d_{\text{exp}}$. In addition, a permutation of the expanded matrices rows and columns is performed to make the same channels correspond across datasets. This expansion of covariances was originally designed to be followed by a supervised transfer learning method called Riemannian Procrustes Analysis (RPA) [Rodrigues et al., 2019]. However, since we operate in an unsupervised setting, we only apply the re-centering step introduced in Section 2.4.2 in this framework.

Signal-based imputation

The "ComImp" method [Nguyen et al., 2023], offers a direct machine learning approach to handle missing data in EEG recordings. First, the time series data are expanded to match the union of the set of channels, hence generating missing values in the data. In our case the EEG signal is expanded to be of size $P_{\text{exp}} \times T$. Then, statistical imputation techniques can be employed. Such approaches require fitting the imputation method on training data that include at least one non-missing value per channel. Here, we considered a multivariate imputer that models each feature with missing values as a function of the other features with a regression. We used the IterativeImputer class of the Scikit-Learn software [Pedregosa et al., 2011] with a ridge estimator.

End-to-end deep learning approaches have also been developed to deal with heterogeneous EEG data. For example, recent works propose leveraging spatial attention mechanisms by using the electrodes' coordinates [Truong et al., 2023, Défossez et al., 2023]. These methods dynamically weigh the importance of different electrodes, which helps in handling variability in sensor configurations across datasets. However, we did not investigate these approaches in this work due to their end-to-end nature and computational cost. End-to-end deep learning models require significant computational resources for training and can be less interpretable compared to other methods. Our focus was on methods that could be more easily integrated into existing EEG analysis pipelines and that offered a balance between performance and computational efficiency.

In summary, while different types of approaches exist to address the challenges of heterogeneous domain adaptation in EEG data, our work primarily explores dimension reduction with common channel selection and imputation techniques. These methods provide practical solutions for handling differences in sensor configurations across datasets.

4.4 Proposed approach: EEG channels interpolation

As introduced in Section 1.2.1 Interpolation is a technique used in the context of EEG data processing to reconstruct the signals of malfunctioning or too noisy channels, usually referred to as 'bad' channels. It uses the signals from the functional channels around the bad ones. In this work, we used interpolation to map the different channels of EEG datasets: the EEG signal was reconstructed on fixed final positions based on the existing signal from all sensors of the datasets. Interpolation involves constructing a linear operator $\boldsymbol{A} \in \mathbb{R}^{P \times P_k}$ that maps the P_k existing EEG channels to the P positions of a fixed template: $\hat{\boldsymbol{X}} = \boldsymbol{A}\boldsymbol{X}$. $\boldsymbol{X} \in \mathbb{R}^{P_k \times T}$ are the recorded EEG signals and $\hat{\boldsymbol{X}} \in \mathbb{R}^{P \times T}$ are the reconstructed signals. This operator can be estimated to reconstruct the EEG signal at any desired position, even if there is no corresponding sensor at that location. Depending on the EEG montage, P can either be smaller or larger than P_k. We present two interpolation techniques used with EEG data: the spherical spline interpolation (SSI) and the field interpolation (FI).

Implementations of both spherical spline interpolation and field interpolation are provided in the MNE-python software [Gramfort et al., 2013].

4.4.1 Spherical spline interpolation

The idea behind SSI is to model the data using smooth functions that are defined on the surface of a sphere, and the functions in our case are spherical splines [Perrin et al., 1989]. In practice, the existing sensors' locations and desired final positions are first projected onto a unit sphere. Then, the linear mapping matrix is computed and finally used to interpolate the signal at the desired position based on the existing signal. The following details on the spherical spline interpolation are adapted from the MNE-python documentation.

We approximate the head as a sphere. We measured the potential at the surface of this sphere at P_k points with $r_p, p = 1 \dots P_k$ their locations. We want to estimate the potential at a new point r at the surface of scalp, where we don't have any measurement. Spherical splines estimate this potential as:

$$V(r) = c_0 + \sum_{p=1}^{P} c_p g_m(\cos(r, r_p)) \tag{4.2}$$

with $\boldsymbol{C} = (c_1, \ldots, c_{P_k})$ the unknown spline coefficient, g_m a function of order m defined as:

$$g_m(x) = \frac{1}{4\pi} \sum_{n=1}^{\infty} \frac{2n+1}{(n(n+1))^m} P_n(x) \tag{4.3}$$

where P_n are the Legendre polynomials of degree n.

To determine the potential at any point, we estimate the constants C by imposing two conditions. First, the spherical spline interpolation function must reproduce the potential at the P_k points where we have measurements:

$$\boldsymbol{G_{ss}C} + \boldsymbol{T_s}c_0 = \boldsymbol{X} \tag{4.4}$$

where $\boldsymbol{G_{ss}} \in \mathbb{R}^{P_k \times P_k}$ is the matrix with elements $\boldsymbol{G_{ss}}(p, p') = g_m(\cos(r_p, r_{p'}))$, $\boldsymbol{X}$ is the EEG signal measured at the P_k good sensors' positions, and $\boldsymbol{T_s} = (1, 1, \ldots, 1)^\top$. (4.4) is the matrix formulation of (4.2).

Second, the sum of the spline coefficients $c_1, \ldots, c_{P_k}$ must be equal to zero:

$$\boldsymbol{T_s}^\top \boldsymbol{C} = 0 \tag{4.5}$$

This condition acts as an average refence applied to the data. By solving (4.4) and (4.5) simultaneously, we obtain:

$$\begin{bmatrix} c_0 \\ \boldsymbol{C} \end{bmatrix} = \begin{bmatrix} \boldsymbol{T_s}^\top & 0 \\ \boldsymbol{T_s} & \boldsymbol{G_{ss}} \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ \boldsymbol{X} \end{bmatrix} = \boldsymbol{C'X} \tag{4.6}$$

where $\boldsymbol{C'} \in \mathbb{R}^{(P_k+1) \times P_k}$ is the matrix $\begin{bmatrix} \boldsymbol{T_s}^\top & 0 \\ \boldsymbol{T_s} & \boldsymbol{G_{ss}} \end{bmatrix}$ without the first column.

The potentials $\hat{\boldsymbol{X}}$ we want to estimate at the new positions can then be computed following the same spline interpolation formula as in (4.4):

$$\hat{\boldsymbol{X}} = \boldsymbol{G_{ds}C} + \boldsymbol{T_d}c_0 \tag{4.7}$$

where $\boldsymbol{G_{ds}} \in \mathbb{R}^{P \times P_k}$ is the matrix with elements $\boldsymbol{G_{ds}}(q, p) = g_m(\cos(r_q, r_p))$ beetween the P new positions r_q and the P_k positions where we have measurements, and $\boldsymbol{T_d} = (1, 1, \ldots, 1)^\top$ of dimension P. Using (4.6), we can rewrite (4.7) as:

$$\hat{\boldsymbol{X}} = [\boldsymbol{T_d} \quad \boldsymbol{G_{ds}}] \begin{bmatrix} c_0 \\ \boldsymbol{C} \end{bmatrix} = [\boldsymbol{T_d} \quad \boldsymbol{G_{ds}}]\boldsymbol{C'X} \tag{4.8}$$

The interpolation matrix we estimate to map the EEG signals from the P recorded channels to the P_j final positions is $\boldsymbol{A} = [\boldsymbol{T_d} \quad \boldsymbol{G_{ds}}]\boldsymbol{C'}$.

In practice we use a regularized version of the this interpolation matrix by adding a positive regularization term λ to each elements of the matrix $\boldsymbol{G_{ss}}$ when estimating the spline coefficients:

$$\begin{cases} (\boldsymbol{G_{ss}} + \lambda \mathbf{1})\boldsymbol{C} + \boldsymbol{T_s} c_0 = \boldsymbol{X} \\ \boldsymbol{T_s}^\top \boldsymbol{C} = 0 \end{cases} \tag{4.9}$$

4.4.2 Field interpolation

The field interpolation consists of estimating the generators of activity in the brain from the recorded EEG signals and then mapping these generators to the desired positions using a forward model.

Forward model

The forward model relates the distribution of estimated brain activity to sensor data at any electrode location [Gramfort et al., 2013]:

$$\boldsymbol{X} = \boldsymbol{F}\boldsymbol{G} + \boldsymbol{E} \tag{4.10}$$

where $\boldsymbol{X} \in \mathbb{R}^{P_k \times T}$ is the measured EEG data, $\boldsymbol{F} \in \mathbb{R}^{P_k \times N}$ is the forward model, $\boldsymbol{G} \in \mathbb{R}^{N \times T}$ are the brain generators, and $\boldsymbol{E} \in \mathbb{R}^{P_k \times T}$ is the noise. $\boldsymbol{F}$, also known as the lead field matrix, describes the relationship between the neural generators and the resulting field measured by the sensors.

To estimate this forward model, we model the brain generators as a canonical distribution, and the head as a sphere with several layers. The lead field is computed by solving the Poisson equation for the electric potential, taking into account the conductivity of different tissues. Methods like the finite element method (FEM) or boundary element method (BEM) are commonly used.

Minimum Norm Estimate

To obtain electric potential estimates at a missing electrode location, the data recorded at the available electrodes are mapped to brain space using a Tikhonov regularization [Tikhonov, 1977], which is referred to as Minimum Norm Estimate (MNE) [Hämäläinen and Ilmoniemi, 1994] in the M/EEG community. The solution to the MNE problem is given by:

$$\hat{\boldsymbol{G}} = (\boldsymbol{F}^\top \boldsymbol{F} + \lambda \boldsymbol{I})^{-1} \boldsymbol{F}^\top \boldsymbol{X} \tag{4.11}$$

where λ is the regularization parameter, $\hat{\boldsymbol{G}}$ are the estimated brain generators and $\boldsymbol{I}$ is the identity matrix.

Then the forward model can be applied to the estimated EEG generators to obtain potential values for any electrode location:

$$\hat{\boldsymbol{X}} = \hat{\boldsymbol{F}}\hat{\boldsymbol{G}} \tag{4.12}$$

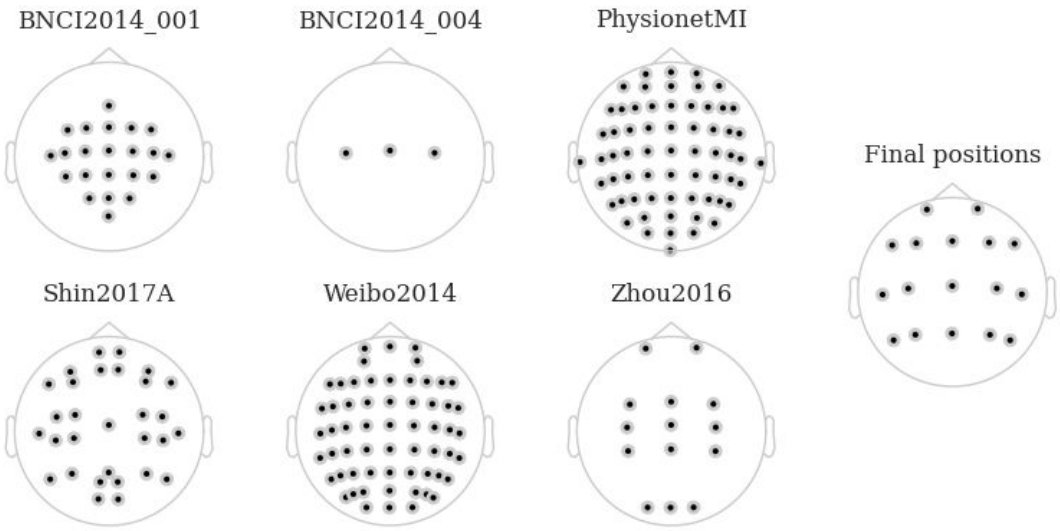

Figure 4.1: **2D projection of sensor positions on the scalp.** On the left, the sensor locations of the 6 datasets studied. On the right, the 17 final positions used for interpolation.

with $\hat{\boldsymbol{F}} \in \mathbb{R}^{P \times T}$ the forward model matrix for the new electrode locations and $\hat{\boldsymbol{X}} \in \mathbb{R}^{P \times T}$ the estimated EEG data at these locations.

4.5 Experimental evaluation

The methods described in Section 4.3 and Section 4.4 were evaluated on six publicly available BCI datasets from the MOABB repository [Aristimunha et al., 2023]. In this section, we present the data and how they were processed for the benchmark. We then describe the classification pipeline and the evaluation protocol used to compare the different methods.

Datasets	Participants	Channels	Sessions	Runs
BNCI2014_001 (B1)	12	22	2	6
BNCI2014_004 (B4)	9	3	5	1
PhysionetMI (P)	109	64	1	1
Shin2017A (S)	29	30	3	1
Weibo2014 (W)	10	60	1	1
Zhou2016 (Z)	4	14	3	2

Table 4.1: **Summary of diverse characteristics of the datasets.**

4.5.1 Data description

The number of subjects, channels, sessions, runs, and trials varies across the datasets, the details of which are listed in Table 4.1. In addition, Figure 4.1 displays the 2D projection of the sensor positions on the scalp for all datasets. All datasets were recorded with the left mastoid as reference, except for Weibo2014 where the reference was taken at the nose. All datasets consist of EEG data recorded while participants performed motor imagery tasks. The subjects were instructed to imagine moving either their right or left hand without actually moving it in response to a visual cue. The classification problem is thus a binary classification task.

Preprocessing

All data were band-pass filtered between 8 and 32 Hz with an Infinite Impulse Response (IIR) forward-backward filter and resampled at 128 Hz. The signals were segmented in epochs at each trial and of duration corresponding to the trial length with no overlap. Subsequently, spatial covariance matrices were computed from the filtered epochs as presented in Section 1.3.3 by regularizing the empirical covariance matrices with the Ledoit-Wolf shrinkage to avoid rank deficiency.

4.5.2 Classification pipeline

Section 4.2 illustrates the classification pipeline depending on the method used for matching the EEG positions across datasets. ComImp, SSI and FI are applied to the clean epochs right after the preprocessing step. The epochs are then transformed into covariance matrices. The DT method is applied to the covariance matrices. The Re-Center step is used to align the data distributions across datasets following the procedure detailed in Section 2.4.2. Subjects are used as domains for re-centering. The geometric mean of the subjects used in the train set is computed by considering all their data points. For the test subject, the geometric mean is computed on the data from its first session, or first run if there is only one session, or the first half of the data if there is only one session and one run.

The six datasets we used have only one channel in common: the Cz channel, located at the top of the head. The common channel selection thus resulted in keeping the signal from the Cz channel, and the associated covariance matrices came down to one value, the variance of this channel. As the union of the channels across all datasets represents $P_{\text{exp}} = 84$ channels, the time series were expanded to 84 channels after ComImp, and the expanded covariance matrices after DT were of size 84×84.

For both interpolation methods, we set the final positions to which all epochs across all datasets were interpolated to the following $P = 17$ channels: [Fp1, Fp2, F7, F3, Fz, F4, F8, C3, Cz, C4, P3, Pz, P4, T3, T4, T5, T6]. The 2D projection of these final positions location on the scalp is shown in Figure 4.1. Even though

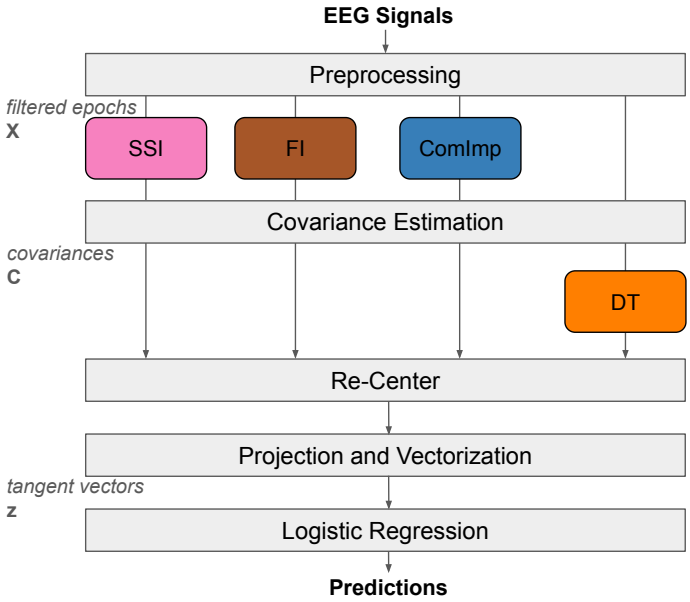

Figure 4.2: **Processing pipeline of EEG data.** Depending on the method, dimensions are matched either when data are represented as epochs for interpolations or as covariances for DT. When there is no alignment performed, the Re-Center step is removed.

we chose $P < P_{\text{exp}}$, all channels of each dataset were used to reconstruct the interpolated signals. We determined this number of final positions not too low to ensure that the task-related data was present in the reconstructed signals, and not too large so the computation time would remain reasonable. The regularization term of the SSI was set to 1e−7 and the one of the FI to 1e−3.

For all methods, we used as classifier a logistic regression from Scikit-Learn with a L2 regularization set to $C = 1$. In addition, we employed a subject-specific calibration procedure to estimate the upper bound of achievable performance. Here we use a fraction of the target subject's data as part of the training set. This was achieved by splitting the target subject's data into two halves, utilizing the first half for training and the second for evaluation. Consequently, this yielded a single calibration accuracy value for each subject.

4.5.3 Leave-one-dataset-out validation

We evaluated the four methods using a leave-one-dataset-out scheme. We first performed this evaluation with an increasing number of target channels seen in the training set by increasing the number of training datasets. For example, when the target dataset was Shin2017A, the first training set consisted of BNCI2014_004 only, the second of BNCI2014_004 and Zhou2016, and so on until the training set included all five remaining datasets. In a second step, five of the six datasets were

combined to form the training set, including all of their subjects. Each subject of the left-out dataset was used as a test set, resulting in one performance value per subject and method. This procedure was repeated six times so that each dataset was left out once. To statistically evaluate the difference in performance between SSI and FI, we performed a Wilcoxon signed-rank test on the classification accuracies obtained from single-trial predictions per method.

4.5.4 Results

Results shown in Figure 4.3 correspond to the first evaluation in which we gradually included datasets in the training set. It displays the classification accuracy difference of each subject of the target dataset as a function of the number of target channels seen during training. The accuracy difference was computed by subtracting the FI accuracy to other methods, subject per subject. The common channel selection method (green lines) systematically led to lower accuracies than FI. For every left-out dataset, except for Shin2017A, ComImp and DT reach lower accuracies than FI with few target channels seen during training. When more target channels are seen in train, source free methods perform similarly to source dependent methods for Weibo2014, Zhou2016 and PhysionetMI, with accuracies comparable to those of the calibration (purple box on Figure 4.4). However, in BNCI datasets, the performance of ComImp and DT outperformed FI when more target channels were seen during training. Interestigly, for BNCI2014_004, the performance of SSI was at the level of the performance of common channel selection. For Shin2017A, FI consistently outperforms all other methods, especially ComImp, DT and common channel selection. The reason is that Shin2017A shares very few channels with the other datasets, and these channels are not located near the motor cortex.

The boxplots displayed in Figure 4.4 correspond to the last point of the learning curves in Figure 4.3, for which five of the datasets were used as training set. The methods are sorted according to whether they require access to target labels (supervised), source data (source dependent), or nothing (source free). For every left-out dataset, except for Shin2017A, interpolation methods performed similarly to source dependent methods and calibration. For Shin2017A, the observations are the same as on the learning curve. Out of all six datasets, the FI led to accuracies significantly higher than the SSI for four datasets with p-values $p \leq 5\text{e}{-2}$. In addition, interpolation methods were faster to compute than ComImp and DT due to the smaller size of the dimension matched data (17 channels for interpolation compared to 84 for ComImp and DT).

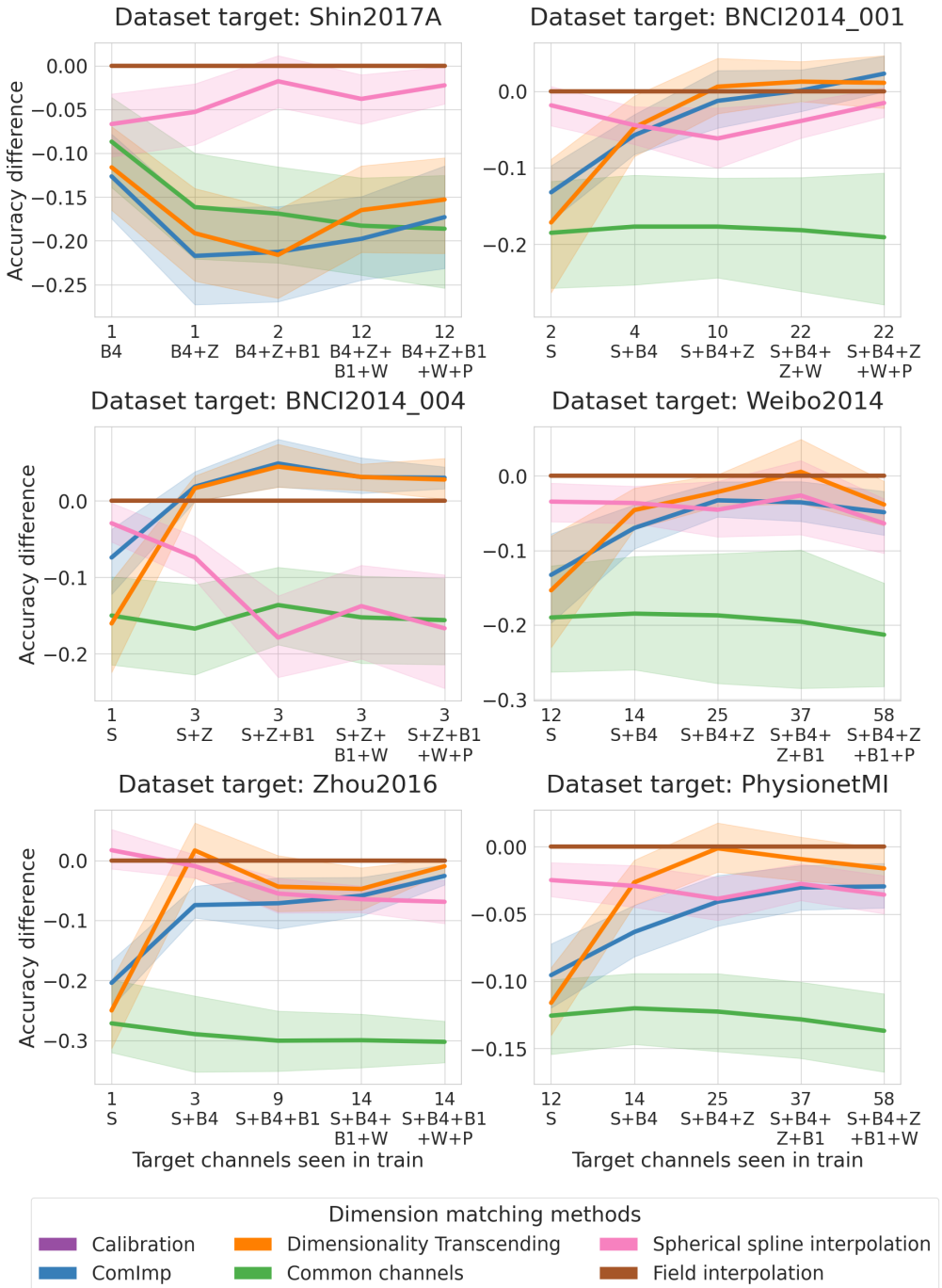

Figure 4.3: **Comparative learning curves for an increasing number of target channels seen during training.** FI performance was used as reference for the accuracy difference. The increasing number of seen target channels is obtained by gradually including datasets in the train set, which is specified in the x-axis. The error bar represents the 95% confidence interval over the target subjects' performance.

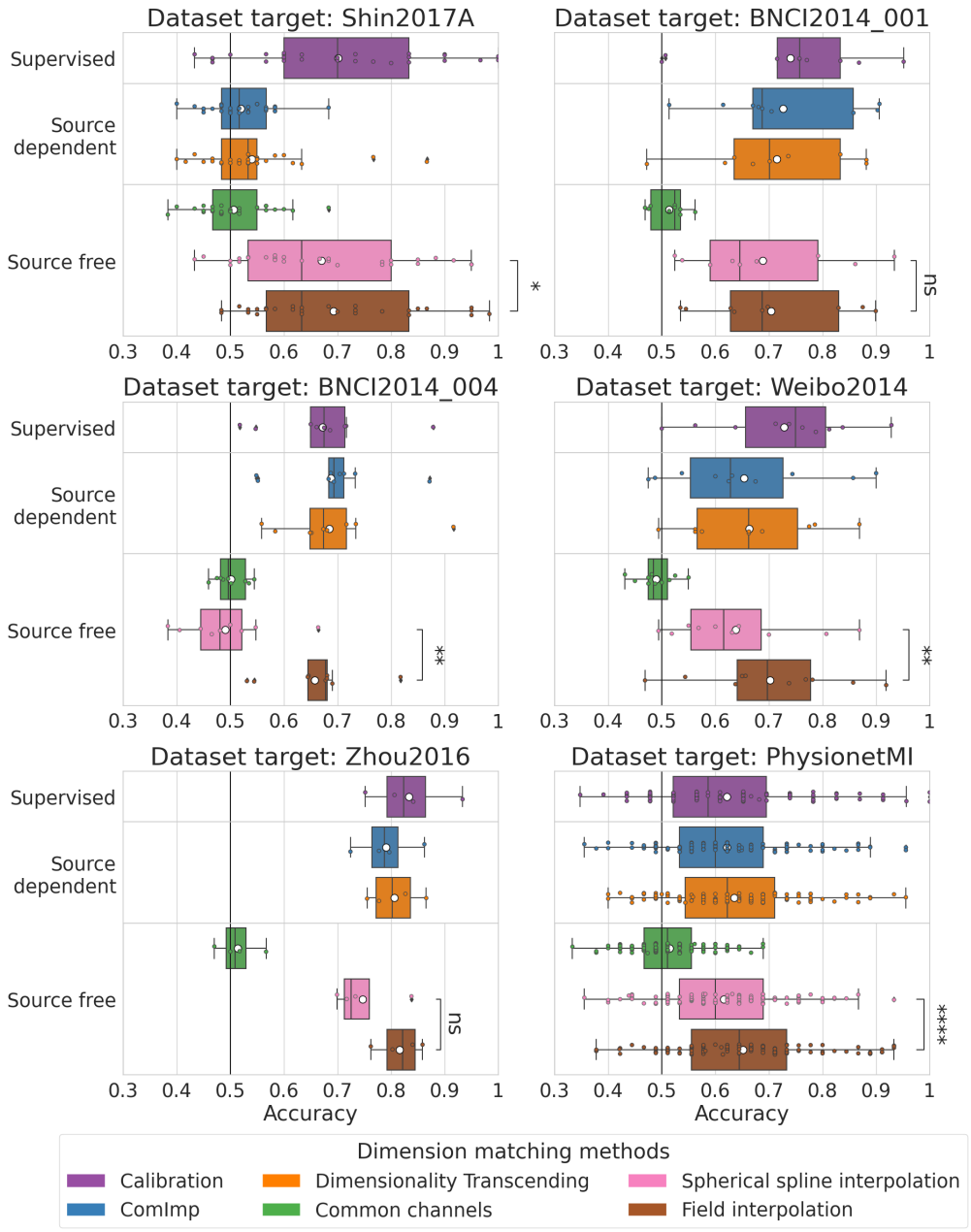

Figure 4.4: **Boxplots of accuracies when the classifier is trained on the five other datasets.** One point corresponds to one subject of the target dataset. A black line represents the median of the box and the mean by a white circle. The black lines indicate the chance level. The stars represent the results of a Wilcoxon test (ns: $p > 5e{-}2$, *: $1e{-}2 < p \leq 5e{-}2$, **: $1e{-}3 < p \leq 1e{-}2$, ****: $p \leq 1e{-}4$).

4.6 Conclusion

In this work, we used a physics-informed, unsupervised, and source-free domain adaptation approach for EEG analysis, specifically addressing challenges arising from varying electrode configurations. This approach leverages the underlying physics of EEG signals through field interpolation, mapping EEG signals to a common sensor configuration As it is applied before any feature extraction, it can be integrated into different EEG analysis pipelines and is not confined to the covariance framework.

The results demonstrated that the field interpolation (FI) approach achieved similar or better accuracies compared to source-dependent approaches, such as signal-based imputation (ComImp) and Dimensionality Transcending (DT), especially when the training data included limited target channels. Additionally, FI systematically outperformed SSI, highlighting its robustness in scenarios with varying electrode configurations.

Despite the promising results, several factors require further investigation. One such factor is the impact of extrapolation, when the new electrode location is outside the surface of the observed sensors' locations. Understanding this impact will help to improve how the interpolation technique is used and ensuring its reliability in diverse recording setups. Another area of interest is the influence of the choice of final positions on the performance of the interpolation methods. Specifically, examining how the number and arrangement of final positions affect the accuracy and generalizability of the model could provide valuable insights for optimizing electrode configurations in future studies.

5 Conclusion

The research presented in this thesis addresses one of the key challenges in M/EEG data analysis: the presence of systematic variability between recordings that arise from varying recording conditions, equipment, and participants. These differences can lead to shifts, commonly referred to as dataset shifts, in the data and the associated labels. Dataset shifts are significant obstacles to the successful application of machine learning models in settings where data is collected across different hospitals, using different equipment, and on varied populations. This thesis has explored the complexities of dataset shifts in M/EEG data and has contributed novel methods to address the difficulty of generalizing predictive models across datasets recorded in different contexts.

Summary of contributions

The focus of this thesis was to better understand and mitigate dataset shifts in M/EEG data. In the course of this work, we considered dataset shifts from multiple perspectives and proposed approaches to address each specific type of shift that can occur in M/EEG data. The contributions made in this work can be summarized as follows:

- **Comprehensive analysis of dataset shifts:** We have provided a detailed investigation of the causes and effects of dataset shifts in M/EEG data. This included a detailed understanding of how the data is generated and of the underlying physiological processes. How to handle and pre-process the data to ensure that the models can learn the relevant patterns was also considered carefully. Additionally, we studied the mechanisms that lead to dataset shifts and examined their consequences on the predictive performance of machine learning models when no domain adaptation is performed.
- **Proposed harmonization and adaptation methods:** Building on existing methods, such as the Riemannian-based alignment techniques and field interpolation, we introduced novel approaches to address different forms of dataset shifts occuring in M/EEG data. We first aimed to harmonize the

data, i.e., to align the data distributions to make them more similar. To begin with, we focused on shifts in the data distribution, and then extended our methods to account for additional shifts in the label distribution. Second, we inspected techniques to match electrode numbers and positions between datasets to ensure comparability across different helmet configurations. The approaches used in this thesis are source-free (or test-time) and require no retraining of the predictive model on the source data when a new target dataset becomes available. We also designed these methods to be as unsupervised as possible, reducing the need for labeled data in the target domain.

- **Evaluation on simulated data and real M/EEG recordings:** The proposed methods were throughly evaluated using both synthetic data and large-scale real-world datasets, including multi-site clinical EEG recordings. These experiments demonstrated the effectiveness of our approaches in mitigating the negative effects of dataset shifts, leading to more reliable and generalizable machine learning models to predict from M/EEG data.
- **Interpretability and clinical relevance:** The framework presented in this thesis has direct implications for clinical EEG analysis, where variability in data collection environments often hampers the success of machine learning models. While addressing these variations is a separate effort in itself, we hope that the methods proposed in this thesis can serve as a catalyst for future work, enabling clinicians and researchers to develop tools that ensure more consistent and accurate assessments of neurological conditions across different patient populations and recording sites.

Futur directions

Following the contributions made in this thesis, several perspectives can be considered for future research. These possible perspectives discussed below are orgamized from relatively easy and short-term to more complex and long-term research directions.

Several aspects of the work presented in this thesis could be further explored in the direct future: the computational efficiency of the methods, the impact of the parameters choices and the evaluation of the proposed methods on other applications. While the current work has demonstrated the effectiveness of these approaches, enhancing the scalability and computational efficiency to accommodate larger datasets would one step further to facilitate the integration of the proposed methods into clinical applications. Moreover, the impact of the "arbitrary" choices made in the proposed methods, such as the choice of the reference electrode, the number of sensors, or the number of components to keep in the covariance matrices, could be more thoroughly assessed. This would provide a better understanding of the robustness of the methods and their sensitivity to these choices. Finally, the proposed methods could be extended beyond the focus of age prediction to

explore the impact of the proposed methods in other applications, such as disease diagnosis or cognitive impairment detection. Doing so would bring us closer to practical clinical implementations, further validating the methods across diverse contexts.

In the medium term, other promising and research directions could be pursued. One perspective would be to combine the different contributions from this thesis to handle shifts in data, labels, and recording devices simultaneously, evaluating this combined framework across multiple applications. The next logical step would be to use the proposed approaches to integrate datasets, potentially achieving better results than traditional supervised learning. By merging complementary information from various datasets, model performance could be enhanced, particularly in clinical scenarios. Additionally, the interpretability of the models should be further investigated. When predicting from M/EEG data, the final results are not the only important aspect; it is also of interest to know which features are driving the predictions and how they relate to the underlying neurophysiological processes. Developing models that are both effective and interpretable will help bringing machine learning techniques in clinical contexts, where understanding of the models are essential.

Finally, several more complex research avenues could be explored. The first one that comes to mind to expand the applicability of the proposed methods beyond covariance-based approaches. For example, integrating these domain adaptation techniques with deep learning architectures could lead to end-to-end learning models capable of directly extracting invariant features from raw M/EEG signals. Such integration could help to be more adaptable to new and unseen datasets while maintaining strong generalization capabilities. Investigating the intersection of domain adaptation with deep learning holds great promise for advancing both theoretical research and practical applications in brain health monitoring and neuroscience. However, in line with the perspectives previously stated, this should be done while finding a way to make it interpretable. Another aspect of M/EEG clinical applications that was not covered in this thesis is the longitudinal analysis. Investigating if and how datasets shifts would manifest in such studies and how they can be mitigated could open new opportunities for longitudinal studies in clinical research and for understanding the aging brain.

Final thoughts

The Riemannian geometry-based methods explored in this thesis, while developed for M/EEG data, are mathematically principled and could have broader utility. For instance, diffusion tensor imaging has already applied Riemannian methods to study white matter structures in the brain. The potential of these techniques could extend to other fields, such as geosciences, where similar challenges of variability, alignment, and feature extraction from complex, high-dimensional data arise. Exploring how these methods could generalize to different scientific

domains may offer new opportunities in fields that require robust data analysis and domain adaptation techniques.

Bibliography

Reza Abiri, Soheil Borhani, Eric W Sellers, Yang Jiang, and Xiaopeng Zhao. A comprehensive review of EEG-based brain–computer interface paradigms. *Journal of Neural Engineering*, 16(1):011001, jan 2019. doi: 10.1088/1741-2552/aaf12e.

Obada Al Zoubi, Chung Ki Wong, Rayus T. Kuplicki, Hung-wen Yeh, Ahmad Mayeli, Hazem Refai, Martin Paulus, and Jerzy Bodurka. Predicting Age From Brain EEG Signals—A Machine Learning Approach. *Frontiers in Aging Neuroscience*, 10, 2018. ISSN 1663-4365. doi: 10.3389/fnagi.2018.00184.

Benyamin Allahgholizadeh Haghi, Spencer Kellis, Sahil Shah, Maitreyi Ashok, Luke Bashford, Daniel Kramer, Brian Lee, Charles Liu, Richard Andersen, and Azita Emami. Deep Multi-State Dynamic Recurrent Neural Networks Operating on Wavelet Based Neural Features for Robust Brain Machine Interfaces. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Cédric Allain, Alexandre Gramfort, and Thomas Moreau. DriPP: Driven Point Processes to Model Stimuli Induced Patterns in M/EEG Signals, 2022.

Javier Andreu-Perez, Carmen C. Y. Poon, Robert D. Merrifield, Stephen T. C. Wong, and Guang-Zhong Yang. Big Data for Health. *IEEE Journal of Biomedical and Health Informatics*, 19(4):1193–1208, 2015. doi: 10.1109/JBHI.2015.2450362.

Gopala K. Anumanchipalli, Josh Chartier, and Edward F. Chang. Speech synthesis from neural decoding of spoken sentences. *Nature*, 568(7753):493–498, 2019. doi: 10.1038/s41586-019-1119-1.

Bruno Aristimunha, Igor Carrara, Pierre Guetschel, Sara Sedlar, Pedro Rodrigues, et al. Mother of all BCI Benchmarks, 2023. URL https://github.com/NeuroTechX/moabb.

Anahit Babayan, Miray Erbey, Deniz Kumral, Janis D Reinelt, Andrea MF Reiter, Josefin Röbbig, H Lina Schaare, Marie Uhlig, Alfred Anwander, Pierre-Louis Bazin, et al. A mind-brain-body dataset of MRI, EEG, cognition, emotion, and peripheral physiology in young and old adults. *Sci Data*, 6(1):1–21, 2019.

Anto Bagić. Look back to leap forward: The emerging new role of magnetoencephalography (MEG) in nonlesional epilepsy. *Clinical Neurophysiology*, 127(1): 60–66, 2016. ISSN 1388-2457. doi: https://doi.org/10.1016/j.clinph.2015.05.009.

Hubert Banville, Omar Chehab, Aapo Hyvärinen, Denis-Alexander Engemann, and Alexandre Gramfort. Uncovering the structure of clinical eeg signals with self-supervised learning. *Journal of Neural Engineering*, 18(4):046020, mar 2021. doi: 10.1088/1741-2552/abca18.

Alexandre Barachant, Stphane Bonnet, Marco Congedo, and Christian Jutten. Common Spatial Pattern revisited by Riemannian geometry. In *2010 IEEE International Workshop on Multimedia Signal Processing*, pages 472–476, 2010. doi: 10.1109/MMSP.2010.5662067.

Alexandre Barachant, Stéphane Bonnet, Marco Congedo, and Christian Jutten. Multiclass Brain–Computer Interface Classification by Riemannian Geometry. *IEEE Transactions on Biomedical Engineering*, 59(4):920–928, 2012. doi: 10.1109/TBME.2011.2172210.

Alexandre Barachant, Stéphane Bonnet, Marco Congedo, and Christian Jutten. Classification of covariance matrices using a Riemannian-based kernel for BCI applications. *Neurocomputing*, 112:172–178, 2013. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2012.12.039. Advances in artificial neural networks, machine learning, and computational intelligence.

Mark Bear, Barry Connors, and Michael A Paradiso. *Neuroscience: exploring the brain, enhanced edition: exploring the brain*. Jones & Bartlett Learning, 2020.

Ramina Behzad and Aida Behzad. The role of EEG in the diagnosis and management of patients with sleep disorders. *Journal of Behavioral and Brain Science*, 11(10):257–266, 2021.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79:151–175, 2010.

Hans Berger. Über das elektroenkephalogramm des menschen. *Archiv für psychiatrie und nervenkrankheiten*, 87(1):527–570, 1929.

Alexandre Bleuzé, Jérémie Mattout, and Marco Congedo. Transfer Learning for the Riemannian Tangent Space: Applications to Brain-Computer Interfaces.

In *2021 International Conference on Engineering and Emerging Technologies (ICEET)*, pages 1–6, 2021. doi: 10.1109/ICEET53442.2021.9659607.

Philipp Bomatter, Joseph Paillard, Pilar Garces, Jörg Hipp, and Denis-Alexander Engemann. Machine learning of brain-specific biomarkers from EEG. *eBioMedicine*, 106, 2024. doi: 10.1016/j.ebiom.2024.105259.

Clément Bonet, Benoît Malézieux, Alain Rakotomamonjy, Lucas Drumetz, Thomas Moreau, Matthieu Kowalski, and Nicolas Courty. Sliced-Wasserstein on symmetric positive definite matrices for M/EEG signals. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 2777–2805. PMLR, 23–29 Jul 2023.

Nicolas Boulant, Franck Mauconduit, Vincent Gras, Alexis Amadon, Caroline Le Ster, Michel Luong, Aurélien Massire, Christophe Pallier, Laure Sabatier, Michel Bottlaender, et al. In vivo imaging of the human brain with the Iseult 11.7-T MRI scanner. *Nature Methods*, pages 1–4, 2024. doi: 10.1038/s41592-024-02472-7.

Nicolas Boumal. *An introduction to optimization on smooth manifolds*. Cambridge University Press, 2023.

Leo Breiman. Random forests. *Machine Learning*, 45:5–32, 2001. doi: 10.1023/A:1010933404324.

György Buzsáki and Andreas Draguhn. Neuronal Oscillations in Cortical Networks. *Science*, 304(5679):1926–1929, 2004. doi: 10.1126/science.1099745.

György Buzsáki and Rodolfo Llinás. Space and time in the brain. *Science*, 358 (6362):482–485, 2017. doi: 10.1126/science.aan8869.

György Buzsáki and Kenji Mizuseki. The log-dynamic brain: how skewed distributions affect network operations. *Nature Reviews Neuroscience*, 15(4):264–278, 2014. doi: doi.org/10.1038/nrn3687.

Igor Carrara, Bruno Aristimunha, Marie-Constance Corsi, Raphael Y. de Camargo, Sylvain Chevallier, and Théodore Papadopoulo. Geometric Neural Network based on Phase Space for BCI-EEG decoding, 2024.

Stanislas Chambon, Mathieu N. Galtier, Pierrick J. Arnal, Gilles Wainrib, and Alexandre Gramfort. A Deep Learning Architecture for Temporal Sleep Stage Classification Using Multivariate and Multimodal Time Series. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 26(4):758–769, 2018. doi: 10.1109/TNSRE.2018.2813138.

Yilun Chen, Ami Wiesel, Yonina C. Eldar, and Alfred O. Hero. Shrinkage Algorithms for MMSE Covariance Estimation. *IEEE Transactions on Signal Processing*, 58(10):5016–5029, 2010. doi: 10.1109/TSP.2010.2053029.

Rasheda Arman Chowdhury, Younes Zerouali, Tanguy Hedrich, Marcel Heers, Eliane Kobayashi, Jean-Marc Lina, and Christophe Grova. MEG–EEG Information Fusion and Electromagnetic Source Imaging: From Theory to Clinical Application in Epilepsy. *Brain topography*, 28:785–812, 2015. doi: 10.1007/s10548-015-0437-3.

David Cohen. Magnetoencephalography: evidence of magnetic fields produced by alpha-rhythm currents. *Science*, 161(3843):784–786, 1968.

Marco Congedo, Alexandre Barachant, and Anton Andreev. A New Generation of Brain-Computer Interface Based on Riemannian Geometry, 2013.

Marco Congedo, Alexandre Barachant, and Rajendra Bhatia. Riemannian geometry for EEG-based brain-computer interfaces; a primer and a review. *Brain-Computer Interfaces*, 4(3):155–174, 2017. doi: 10.1080/2326263X.2017.1297192.

Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(9):1853–1865, 2017. doi: 10.1109/TPAMI.2016.2615921.

Gabriela Csurka et al. *Domain adaptation in computer vision applications*, volume 2. Springer, 2017.

Fernando da Silva Lopes. EEG and MEG: relevance to neuroscience. *Neuron*, 80 (5):1112–1128, 2013.

Sven Dähne, Frank C Meinecke, Stefan Haufe, Johannes Höhne, Michael Tangermann, Klaus-Robert Müller, and Vadim V Nikulin. SPoC: A novel framework for relating the amplitude of neuronal oscillations to behaviorally relevant parameters. *NeuroImage*, 86:111–122, 2014. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2013.07.079.

Bharath Bhushan Damodaran, Benjamin Kellenberger, Remi Flamary, Devis Tuia, and Nicolas Courty. DeepJDOT: Deep Joint Distribution Optimal Transport for Unsupervised Domain Adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

Fernando Soares de Aguiar Neto and João Luís Garcia Rosa. Depression biomarkers using non-invasive EEG: A review. *Neuroscience & Biobehavioral Reviews*, 105:83–93, 2019. ISSN 0149-7634. doi: https://doi.org/10.1016/j.neubiorev.2019.07.021.

Alain de Cheveigné and Lucas C. Parra. Joint decorrelation, a versatile tool for multichannel data analysis. *NeuroImage*, 98:487–505, 2014. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2014.05.068.

Alain de Cheveigné, Giovanni M. Di Liberto, Dorothée Arzounian, Daniel D.E. Wong, Jens Hjortkjær, Søren Fuglsang, and Lucas C. Parra. Multiway canonical correlation analysis of brain data. *NeuroImage*, 186:728–740, 2019. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2018.11.026.

Alexandre Défossez, Charlotte Caucheteux, Jérémy Rapin, Ori Kabeli, and Jean-Rémi King. Decoding speech perception from non-invasive brain recordings. *Nature Machine Intelligence*, 5(10):1097–1107, 2023. doi: https://doi.org/10.1038/s42256-023-00714-5.

Arnaud Delorme. EEG is better left alone. *Scientific reports*, 13(1):2372, 2023. doi: https://doi.org/10.1038/s41598-023-27528-0.

Li Deng and Yang Liu. *Deep Learning in Natural Language Processing*. Springer, 2018.

Jacek P. Dmochowski, Jason J. Ki, Paul DeGuzman, Paul Sajda, and Lucas C. Parra. Extracting multidimensional stimulus-response correlations using hybrid encoding-decoding of neural activity. *NeuroImage*, 180:134–146, 2018. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2017.05.037. New advances in encoding and decoding of brain signals.

Jérôme Dockès, Gaël Varoquaux, and Jean-Baptiste Poline. Preventing dataset shift from breaking machine-learning biomarkers. *GigaScience*, 10(9):giab055, 09 2021. ISSN 2047-217X. doi: 10.1093/gigascience/giab055.

Ulrike Ebert, Matthias Grossmann, Reinhard Oertel, Thomas Gramatté, and Wilhelm Kirch. Pharmacokinetic-pharmacodynamic modeling of the electroencephalogram effects of scopolamine in healthy volunteers. *The Journal of Clinical Pharmacology*, 41(1):51–60, 2001. doi: https://doi.org/10.1177/00912700122009836.

Denis A Engemann, Federico Raimondo, Jean-Rémi King, Benjamin Rohaut, Gilles Louppe, Frédéric Faugeras, Jitka Annen, Helena Cassol, Olivia Gosseries, Diego Fernandez-Slezak, Steven Laureys, Lionel Naccache, Stanislas Dehaene, and Jacobo D Sitt. Robust EEG-based cross-site and cross-protocol classification of states of consciousness. *Brain*, 141(11):3179–3192, 10 2018. ISSN 0006-8950. doi: 10.1093/brain/awy251.

Denis A Engemann, Oleh Kozynets, David Sabbagh, Guillaume Lemaître, Gael Varoquaux, Franziskus Liem, and Alexandre Gramfort. Combining magnetoencephalography with magnetic resonance imaging enhances learning of

surrogate-biomarkers. *eLife*, 9:e54055, may 2020. ISSN 2050-084X. doi: 10.7554/eLife.54055.

Denis A. Engemann, Apolline Mellot, Richard Höchenberger, Hubert Banville, David Sabbagh, Lukas Gemein, Tonio Ball, and Alexandre Gramfort. A reusable benchmark of brain-age prediction from M/EEG resting-state signals. *NeuroImage*, 262:119521, 2022. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2022.119521.

E. Estrada, H. Nazeran, P. Nava, K. Behbehani, J. Burk, and E. Lucas. EEG feature extraction for classification of sleep stages. In *The 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, volume 1, pages 196–199, 2004. doi: 10.1109/IEMBS.2004.1403125.

Abolfazl Farahani, Sahar Voghoei, Khaled Rasheed, and Hamid R Arabnia. A brief review of domain adaptation. *Advances in data science and information engineering: proceedings from ICDATA 2020 and IKE 2020*, pages 877–894, 2021.

Thomas Fletcher. Geodesic Regression on Riemannian Manifolds. In Pennec, Xavier, Joshi, Sarang, Nielsen, and Mads, editors, *Proceedings of the Third International Workshop on Mathematical Foundations of Computational Anatomy - Geometrical and Statistical Methods for Modelling Biological Shape Variability*, pages 75–86, Toronto, Canada, September 2011.

Dylan Forenzo, Hao Zhu, Jenn Shanahan, Jaehyun Lim, and Bin He. Continuous tracking using deep learning-based decoding for noninvasive brain–computer interface. *PNAS Nexus*, 3(4):pgae145, 04 2024. ISSN 2752-6542. doi: 10.1093/pnasnexus/pgae145.

Wolfgang Förstner and Boudewijn Moonen. *A Metric for Covariance Matrices*, pages 299–309. Springer Berlin Heidelberg, Berlin, Heidelberg, 2003. ISBN 978-3-662-05296-9. doi: 10.1007/978-3-662-05296-9_31.

Andrew Gelman. Multilevel (hierarchical) modeling: what it can and cannot do. *Technometrics*, 48(3):432–435, 2006. doi: 10.1198/004017005000000661.

Lukas A.W. Gemein, Robin T. Schirrmeister, Patryk Chrabąszcz, Daniel Wilson, Joschka Boedecker, Andreas Schulze-Bonhage, Frank Hutter, and Tonio Ball. Machine-learning-based diagnostics of EEG pathology. *NeuroImage*, 220:117021, 2020. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2020.117021.

Michael Heath Gene H. Golub and Grace Wahba. Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter. *Technometrics*, 21(2):215–223, 1979. doi: 10.1080/00401706.1979.10489751.

Frederic A Gibbs, EL Gibbs, and WG Lennox. Effect on the electro-encephalogram of certain drugs which influence nervous activity. *Archives of Internal Medicine*, 60(1):154–166, 1937. doi: 10.1001/archinte.1937.00180010159012.

Théo Gnassounou, Rémi Flamary, and Alexandre Gramfort. Convolution Monge Mapping Normalization for learning on sleep data. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 10457–10476. Curran Associates, Inc., 2023.

Théo Gnassounou, Oleksii Kachaiev, Rémi Flamary, Antoine Collas, Yanis Lalou, Antoine de Mathelin, Alexandre Gramfort, Ruben Bueno, Florent Michel, Apolline Mellot, Virginie Loison, Ambroise Odonnat, and Thomas Moreau. Skada : Scikit adaptation, 7 2024. URL https://scikit-adaptation.github.io/.

Alexandre Gramfort. *Mapping, timing and tracking cortical activations with MEG and EEG: Methods and application to human vision*. PhD thesis, Ecole nationale supérieure des telecommunications-ENST, 2009.

Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A. Engemann, Daniel Strohmeier, Christian Brodbeck, Lauri Parkkonen, and Matti S. Hämäläinen. MNE software for processing MEG and EEG data. *NeuroImage*, 86:446–460, 2014. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2013.10.027.

Alexandre Gramfort et al. MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience*, 7(267):1–13, 2013. doi: 10.3389/fnins.2013.00267.

Arthur Gretton, Alex Smola, Jiayuan Huang, Marcel Schmittfull, Karsten Borgwardt, and Bernhard Schölkopf. Covariate shift by kernel mean matching. 2008. doi: https://doi.org/10.7551/mitpress/7921.003.0013.

Dorien Groenendaal, Jan Freijer, Andrea Rosier, Dennis de Mik, Glynis Nicholls, Anne Hersey, Andrew D. Ayrton, Meindert Danhof, and Elizabeth C.M. de Lange. Pharmacokinetic/pharmacodynamic modelling of the EEG effects of opioids: The role of complex biophase distribution kinetics. *European Journal of Pharmaceutical Sciences*, 34(2):149–163, 2008. ISSN 0928-0987. doi: https://doi.org/10.1016/j.ejps.2008.03.003.

Joachim Gross. Magnetoencephalography in cognitive neuroscience: a primer. *Neuron*, 104(2):189–204, 2019.

Haris Hakeem, Wei Feng, Zhibin Chen, Jiun Choong, Martin J. Brodie, Si-Lei Fong, Kheng-Seang Lim, Junhong Wu, Xuefeng Wang, Nicholas Lawn, Guanzhong Ni, Xiang Gao, Mijuan Luo, Ziyi Chen, Zongyuan Ge, and Patrick Kwan. Development and Validation of a Deep Learning Model for Predicting

Treatment Response in Patients With Newly Diagnosed Epilepsy. *JAMA Neurology*, 79(10):986–996, 10 2022. ISSN 2168-6149. doi: 10.1001/jamaneurol.2022.2514.

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. *Reviews of modern Physics*, 65:413–497, Apr 1993. doi: 10.1103/RevModPhys.65.413.

Matti S Hämäläinen and Risto J Ilmoniemi. Interpreting magnetic fields of the brain: minimum norm estimates. *Medical & Biological Engineering & Computing*, 32:35–42, 1994. doi: https://doi.org/10.1007/BF02512476.

A. Harati, S. López, I. Obeid, J. Picone, M. P. Jacobson, and S. Tobochnik. The TUH EEG CORPUS: A big data resource for automated EEG interpretation. In *2014 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, pages 1–5, 2014. doi: 10.1109/SPMB.2014.7002953.

Riitta Hari and Aina Puce. *MEG-EEG Primer*. Oxford University Press, 2023.

Huan He, Owen Queen, Teddy Koker, Consuelo Cuevas, Theodoros Tsiligkaridis, and Marinka Zitnik. Domain Adaptation for Time Series Under Feature and Label Shifts. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 12746–12774. PMLR, 23–29 Jul 2023.

Elisabeth R M Heremans, Huy Phan, Pascal Borzée, Bertien Buyse, Dries Testelmans, and Maarten De Vos. From unsupervised to semi-supervised adversarial domain adaptation in electroencephalography-based sleep staging. *Journal of Neural Engineering*, 19(3):036044, jun 2022. doi: 10.1088/1741-2552/ac6ca8.

JL Hernández, P Valdés, R Biscay, T Virues, S Szava, J Bosch, A Riquenes, and I Clark. A Global Scale Factor in Brain Topography. *International Journal of Neuroscience*, 76(3-4):267–278, 1994. doi: 10.3109/00207459408986009.

Andras Horvath, Anna Szucs, Gabor Csukly, Anna Sakovics, Gabor Stefanics, and Anita Kamondi. EEG and ERP biomarkers of Alzheimer's disease: a critical review. *Frontiers in Bioscience (Landmark edition)*, 23:183–220, 2018.

Behshad Hosseinifard, Mohammad Hassan Moradi, and Reza Rostami. Classifying depression patients and normal subjects using machine learning techniques and nonlinear features from EEG signal. *Computer Methods and Programs in Biomedicine*, 109(3):339–345, 2013. ISSN 0169-2607. doi: https://doi.org/10.1016/j.cmpb.2012.10.008.

Joop Hox. Multilevel modeling: When and why. In *Classification, data analysis, and data highways: proceedings of the 21st Annual Conference of the Gesellschaft für Klassifikation eV, University of Potsdam, March 12–14, 1997*, pages 147–154. Springer, 1998.

Zhiwu Huang and Luc Van Gool. A riemannian network for SPD matrix learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017. doi: https://doi.org/10.1609/aaai.v31i1.10866.

Turan M Itil and Kurt Z Itil. Quantitative EEG brain mapping in psychotropic drug development, drug treatment selection, and monitoring. *American Journal of Therapeutics*, 2(5):359–367, 1995.

Cecilia Jarne, Ben Griffin, and Diego Vidaurre. Predicting subject traits from brain spectral signatures: an application to brain ageing. *bioRxiv*, 2024. doi: 10.1101/2023.11.02.565261.

Mainak Jas, Denis A. Engemann, Yousra Bekhti, Federico Raimondo, and Alexandre Gramfort. Autoreject: Automated artifact rejection for MEG and EEG data. *NeuroImage*, 159:417–429, 2017. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2017.06.030.

Haiteng Jiang, Peiyin Chen, Zhaohong Sun, Chengqian Liang, Rui Xue, Liansheng Zhao, Qiang Wang, Xiaojing Li, Wei Deng, Zhongke Gao, et al. Assisting schizophrenia diagnosis using clinical electroencephalography and interpretable graph neural networks: a real-world and cross-site study. *Neuropsychopharmacology*, 48(13):1920–1930, 2023. doi: https://doi.org/10.1038/s41386-023-01658-5.

Eric R Kandel, James H Schwartz, Thomas M Jessell, Steven Siegelbaum, A James Hudspeth, Sarah Mack, et al. *Principles of Neural Science*, volume 4. McGraw-hill New York, 2000.

Hermann Karcher. Riemannian center of mass and mollifier smoothing. *Communications on Pure and Applied Mathematics*, 30(5):509–541, 1977. doi: https://doi.org/10.1002/cpa.3160300502.

Hyunwoo J. Kim, Nagesh Adluru, Heemanshu Suri, Baba C. Vemuri, Sterling C. Johnson, and Vikas Singh. Riemannian Nonlinear Mixed Effects Models: Analyzing Longitudinal Deformations in Neuroimaging. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5777–5786, 2017. doi: 10.1109/CVPR.2017.612.

Reinmar Kobler, Jun-ichiro Hirayama, Qibin Zhao, and Motoaki Kawanabe. SPD domain-specific batch normalization to crack interpretable unsupervised domain adaptation in EEG. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho,

and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 6219–6235. Curran Associates, Inc., 2022.

Zoltan J Koles, Michael S Lazar, and Steven Z Zhou. Spatial patterns underlying population differences in the background EEG. *Brain topography*, 2(4):275–284, 1990. doi: https://doi.org/10.1007/BF01129656.

Zoltan Joseph Koles. The quantitative extraction and topographic mapping of the abnormal components in the clinical EEG. *Electroencephalography and Clinical Neurophysiology*, 79(6):440–447, 1991. ISSN 0013-4694. doi: https://doi.org/10.1016/0013-4694(91)90163-X.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

Kaisu Lankinen, Jukka Saari, Riitta Hari, and Miika Koskinen. Intersubject consistency of cortical MEG signals during movie viewing. *NeuroImage*, 92:217–224, 2014. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2014.02.004.

Kaisu Lankinen, Jukka Saari, Yevhen Hlushchuk, Pia Tikka, Lauri Parkkonen, Riitta Hari, and Miika Koskinen. Consistency and similarity of MEG- and fMRI-signal time courses during movie viewing. *NeuroImage*, 173:361–369, 2018. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2018.02.045.

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004. ISSN 0047-259X. doi: https://doi.org/10.1016/S0047-259X(03)00096-4.

Agatha Lenartowicz and Sandra K Loo. Use of EEG to diagnose ADHD. *Current psychiatry reports*, 16:1–11, 2014. doi: https://doi.org/10.1007/s11920-014-0498-0.

Dianqi Li, Yizhe Zhang, Zhe Gan, Yu Cheng, Chris Brockett, Bill Dolan, and Ming-Ting Sun. Domain adaptive text style transfer. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3304–3313, Hong Kong, China, November 2019a. Association for Computational Linguistics. doi: 10.18653/v1/D19-1325.

Min Li, Ying Wang, Carlos Lopez-Naranjo, Shiang Hu, Ronaldo César García Reyes, Deirel Paz-Linares, Ariosky Areces-Gonzalez, Aini Ismafairus Abd Hamid, Alan C. Evans, Alexander N. Savostyanov, Ana Calzada-Reyes, Arno

Villringer, Carlos A. Tobon-Quintero, Daysi Garcia-Agustin, Dezhong Yao, Li Dong, Eduardo Aubert-Vazquez, Faruque Reza, Fuleah Abdul Razzaq, Hazim Omar, Jafri Malin Abdullah, Janina R. Galler, John F. Ochoa-Gomez, Leslie S. Prichep, Lidice Galan-Garcia, Lilia Morales-Chacon, Mitchell J. Valdes-Sosa, Marius Tröndle, Mohd Faizal Mohd Zulkifly, Muhammad Riddha Bin Abdul Rahman, Natalya S. Milakhina, Nicolas Langer, Pavel Rudych, Thomas Koenig, Trinidad A. Virues-Alba, Xu Lei, Maria L. Bringas-Vega, Jorge F. Bosch-Bayard, and Pedro Antonio Valdes-Sosa. Harmonized-Multinational qEEG norms (HarMNqEEG). *NeuroImage*, 256:119190, 2022. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2022.119190.

Yan Li, Qingguo Wei, Yuebin Chen, and Xichen Zhou. Transfer Learning Based on Hybrid Riemannian and Euclidean Space Data Alignment and Subject Selection in Brain-Computer Interfaces. *IEEE Access*, 9:6201–6212, 2021. doi: 10.1109/ACCESS.2020.3048683.

Yitong Li, Michael Murias, Samantha Major, Geraldine Dawson, and David Carlson. On Target Shift in Adversarial Domain Adaptation. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 616–625. PMLR, 16–18 Apr 2019b.

Carlos Lopez Naranjo, Fuleah Abdul Razzaq, Min Li, Ying Wang, Jorge F. Bosch-Bayard, Martin A. Lindquist, Anisleidy Gonzalez Mitjans, Ronaldo Garcia, Arielle G. Rabinowitz, Simon G. Anderson, Giuseppe A. Chiarenza, Ana Calzada-Reyes, Trinidad Virues-Alba, Janina R. Galler, Ludovico Minati, Maria L. Bringas Vega, and Pedro A. Valdes-Sosa. EEG functional connectivity as a Riemannian mediator: An application to malnutrition and cognition. *Human Brain Mapping*, 45(7):e26698, 2024. doi: https://doi.org/10.1002/hbm.26698.

Fabien Lotte, Laurent Bougrain, Andrzej Cichocki, Maureen Clerc, Marco Congedo, Alain Rakotomamonjy, and Florian Yger. A review of classification algorithms for EEG-based brain–computer interfaces: a 10 year update. *Journal of neural engineering*, 15(3):031005, 2018. doi: 10.1088/1741-2552/aab2f2.

Gal Maman, Or Yair, Danny Eytan, and Ronen Talmon. Domain Adaptation Using Riemannian Geometry of SPD Matrices. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4464–4468, 2019. doi: 10.1109/ICASSP.2019.8682989.

Apolline Mellot, Antoine Collas, Pedro L. C. Rodrigues, Denis Engemann, and Alexandre Gramfort. Harmonizing and aligning M/EEG datasets with covariance-based techniques to enhance predictive regression modeling. *Imaging Neuroscience*, 1:1–23, 12 2023. ISSN 2837-6056. doi: 10.1162/imag_a_00040.

Harvey Moldofsky. Sleep and pain. *Sleep Medicine Reviews*, 5(5):385–396, 2001. ISSN 1087-0792. doi: https://doi.org/10.1053/smrv.2001.0179.

Claude Nadeau and Yoshua Bengio. Inference for the Generalization Error. In S. Solla, T. Leen, and K. Müller, editors, *Advances in Neural Information Processing Systems*, volume 12. MIT Press, 1999.

Chuong H Nguyen, George K Karavas, and Panagiotis Artemiadis. Inferring imagined speech using EEG signals: a new approach using Riemannian manifold features. *Journal of Neural Engineering*, 15(1):016002, 2017. doi: 10.1088/1741-2552/aa8235.

Thu Nguyen, Rabindra Khadka, Nhan Phan, Anis Yazidi, Pål Halvorsen, and Michael A. Riegler. Combining datasets to increase the number of samples and improve model fitting, 2023.

Jorge Nocedal and Stephen J Wright. *Numerical optimization*. Springer, 1999.

Iyad Obeid and Joseph Picone. The Temple University Hospital EEG Data Corpus. *Frontiers in Neuroscience*, 10, 2016. ISSN 1662-453X. doi: 10.3389/fnins.2016.00196.

Joseph Paillard, Jörg F. Hipp, and Denis A. Engemann. GREEN: a lightweight architecture using learnable wavelets and Riemannian geometry for biomarker exploration. *bioRxiv*, 2024. doi: 10.1101/2024.05.14.594142.

Sinno Jialin Pan, Ivor W. Tsang, James T. Kwok, and Qiang Yang. Domain Adaptation via Transfer Component Analysis. *IEEE Transactions on Neural Networks*, 22(2):199–210, 2011. doi: 10.1109/TNN.2010.2091281.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python . *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Peizhen Peng, Liping Xie, Kanjian Zhang, Jinxia Zhang, Lu Yang, and Haikun Wei. Domain adaptation for epileptic EEG classification using adversarial learning and Riemannian manifold. *Biomedical Signal Processing and Control*, 75:103555, 2022. ISSN 1746-8094. doi: https://doi.org/10.1016/j.bspc.2022.103555.

Xavier Pennec, Pierre Fillard, and Nicholas Ayache. A Riemannian Framework for Tensor Computing. *International Journal of Computer Vision*, 66(1):41–66, January 2006. ISSN 1573-1405. doi: 10.1007/s11263-005-3222-z.

F. Perrin, J. Pernier, O. Bertrand, and J.F. Echallier. Spherical splines for scalp potential and current density mapping. *Electroencephalography and Clinical Neurophysiology*, 72(2):184–187, 1989. ISSN 0013-4694. doi: https://doi.org/10.1016/0013-4694(89)90180-6.

Mathias Perslev, Sune Darkner, Lykke Kempfner, Miki Nikolic, Poul Jørgen Jennum, and Christian Igel. U-Sleep: resilient high-frequency sleep staging. *NPJ digital medicine*, 4(1):72, 2021. doi: https://doi.org/10.1038/s41746-021-00440-5.

Maurice Bertram Priestley. *Spectral Analysis and Time Series*, volume 890. Academic Press London, 1981.

Dale Purves, George J Augustine, David Fitzpatrick, William Hall, Anthony-Samuel LaMantia, and Leonard White. *Neurosciences*. De Boeck Supérieur, 2019.

Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset shift in machine learning*. MIT Press, 2022.

Ievgen Redko, Emilie Morvant, Amaury Habrard, Marc Sebban, and Younes Bennani. *Advances in Domain Adaptation Theory*. Elsevier, 2019.

Pedro L. C. Rodrigues, Marco Congedo, and Christian Jutten. Dimensionality Transcending: A Method for Merging BCI Datasets With Different Dimensionalities. *IEEE Transactions on Biomedical Engineering*, 68(2):673–684, 2021. doi: 10.1109/TBME.2020.3010854.

Pedro Luiz Coelho Rodrigues, Christian Jutten, and Marco Congedo. Riemannian Procrustes Analysis: Transfer Learning for Brain–Computer Interfaces. *IEEE Transactions on Biomedical Engineering*, 66(8):2390–2401, 2019. doi: 10.1109/TBME.2018.2889705.

Benjamin Rohaut, C Calligaris, B Hermann, P Perez, F Faugeras, F Raimondo, J-R King, D Engemann, C Marois, L Le Guennec, et al. Multimodal assessment improves neuroprognosis performance in clinically unresponsive critical-care patients with brain injury. *Nature Medicine*, pages 1–7, 2024. doi: https://doi.org/10.1038/s41591-024-03019-1.

Andrea O. Rossetti, Kaspar Schindler, Raoul Sutter, Stephan Rüegg, Frédéric Zubler, Jan Novy, Mauro Oddo, Loane Warpelin-Decrausaz, and Vincent Alvarez. Continuous vs Routine Electroencephalogram in Critically Ill Adults With Altered Consciousness and No Recent Seizure: A Multicenter Randomized Clinical Trial. *JAMA Neurology*, 77(10):1225–1232, 10 2020. ISSN 2168-6149. doi: 10.1001/jamaneurol.2020.2264.

Yannick Roy, Hubert Banville, Isabela Albuquerque, Alexandre Gramfort, Tiago H Falk, and Jocelyn Faubert. Deep learning-based electroencephalography analysis: a systematic review. *Journal of neural engineering*, 16(5):051001, 2019. doi: 10.1088/1741-2552/ab260c.

David Sabbagh, Pierre Ablin, Gaël Varoquaux, Alexandre Gramfort, and Denis A. Engemann. Manifold-regression to predict from MEG/EEG brain signals without source modeling. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

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. *NeuroImage*, 222:116893, 2020. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2020.116893.

Jean-Baptiste Schiratti, Stéphanie Allassonniere, Olivier Colliot, and Stanley Durrleman. Learning spatiotemporal trajectories from manifold-valued longitudinal data. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

Jean-Baptiste Schiratti, Stéphanie Allassonnière, Olivier Colliot, and Stanley Durrleman. A Bayesian Mixed-Effects Model to Learn Trajectories of Changes from Repeated Manifold-Valued Observations. *Journal of Machine Learning Research*, 18(133):1–33, 2017.

Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. Deep learning with convolutional neural networks for EEG decoding and visualization. *Human Brain Mapping*, 38(11):5391–5420, 2017. doi: https://doi.org/10.1002/hbm.23730.

G. Schneider, A. W. Gelb, B. Schmeller, R. Tschakert, and E. Kochs. Detection of awareness in surgical patients with EEG-based indices—bispectral index and patient state index. *BJA: British Journal of Anaesthesia*, 91(3):329–335, 09 2003. ISSN 0007-0912. doi: 10.1093/bja/aeg188.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000. ISSN 0378-3758. doi: https://doi.org/10.1016/S0378-3758(00)00115-4.

Lene Theil Skovgaard. A Riemannian Geometry of the Multivariate Normal Model. *Scandinavian Journal of Statistics*, 11(4):211–223, 1984. ISSN 03036898,

14679469. Publisher: [Board of the Foundation of the Scandinavian Journal of Statistics, Wiley].

Steven M Stufflebeam, Naoaki Tanaka, and Seppo P Ahlfors. Clinical applications of magnetoencephalography. *Human Brain Mapping*, 30(6):1813–1823, 2009. doi: https://doi.org/10.1002/hbm.20792.

Masashi Sugiyama, Shinichi Nakajima, Hisashi Kashima, Paul Buenau, and Motoaki Kawanabe. Direct Importance Estimation with Model Selection and Its Application to Covariate Shift Adaptation. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007.

Baochen Sun, Jiashi Feng, and Kate Saenko. Return of Frustratingly Easy Domain Adaptation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30 (1), Mar. 2016. doi: 10.1609/aaai.v30i1.10306.

Baochen Sun, Jiashi Feng, and Kate Saenko. *Correlation Alignment for Unsupervised Domain Adaptation*, pages 153–171. Springer International Publishing, Cham, 2017. ISBN 978-3-319-58347-1. doi: 10.1007/978-3-319-58347-1_8.

Haoqi Sun, Luis Paixao, Jefferson T. Oliva, Balaji Goparaju, Diego Z. Carvalho, Kicky G. van Leeuwen, Oluwaseun Akeju, Robert J. Thomas, Sydney S. Cash, Matt T. Bianchi, and M. Brandon Westover. Brain age from the electroencephalogram of sleep. *Neurobiology of Aging*, 74:112–120, 2019. ISSN 0197-4580. doi: https://doi.org/10.1016/j.neurobiolaging.2018.10.016.

Sining Sun, Ching-Feng Yeh, Mei-Yuh Hwang, Mari Ostendorf, and Lei Xie. Domain Adversarial Training for Accented Speech Recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4854–4858, 2018. doi: 10.1109/ICASSP.2018.8462663.

Akara Supratak, Hao Dong, Chao Wu, and Yike Guo. DeepSleepNet: A Model for Automatic Sleep Stage Scoring Based on Raw Single-Channel EEG. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(11):1998–2008, 2017. doi: 10.1109/TNSRE.2017.2721116.

S. Taulu, J. Simola, and M. Kajola. Applications of the signal space separation method. *IEEE Transactions on Signal Processing*, 53(9):3359–3372, 2005. doi: 10.1109/TSP.2005.853302.

Samu Taulu and Matti Kajola. Presentation of electromagnetic multichannel data: the signal space separation method. *Journal of Applied Physics*, 97(12), 2005. doi: https://doi.org/10.1063/1.1935742.

Jason R. Taylor, Nitin Williams, Rhodri Cusack, Tibor Auer, Meredith A. Shafto, Marie Dixon, Lorraine K. Tyler, Cam-CAN, and Richard N. Henson. 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. *NeuroImage*, 144:262–269, 2017. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2015.09.018. Data Sharing Part II.

Andreĭ Nikolayevich Tikhonov. Solutions of Ill-Posed Problems. *VH Winston and Sons*, 1977.

Dung Truong, Muhammad Abdullah Khalid, and Arnaud Delorme. Deep learning applied to EEG data with different montages using spatial attention. In *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2587–2593, 2023. doi: 10.1109/BIBM58861.2023.10385525.

Thomas Tveitstøl, Mats Tveter, Ana Silvina Pérez Teseyra, et al. Introducing Region Based Pooling for handling a varied number of EEG channels for deep learning models. *Frontiers in Neuroinformatics*, 17, 2024. ISSN 1662-5196. doi: 10.3389/fninf.2023.1272791.

Alexandros T. Tzallas, Markos G. Tsipouras, and Dimitrios I. Fotiadis. Epileptic Seizure Detection in EEGs Using Time–Frequency Analysis. *IEEE Transactions on Information Technology in Biomedicine*, 13(5):703–710, 2009. doi: 10.1109/TITB.2009.2017939.

Mikko A Uusitalo and Risto J Ilmoniemi. Signal-space projection method for separating MEG or EEG into components. *Medical and Biological Engineering and Computing*, 35:135–140, 1997. doi: https://doi.org/10.1007/BF02534144.

Paola Vassallo, Jan Novy, Frédéric Zubler, Kaspar Schindler, Vincent Alvarez, Stephan Rüegg, and Andrea O. Rossetti. EEG spindles integrity in critical care adults. Analysis of a randomized trial. *Acta Neurologica Scandinavica*, 144(6):655–662, 2021. doi: https://doi.org/10.1111/ane.13510.

Paul M Vespa, DaiWai M Olson, Sayona John, Kyle S Hobbs, Kapil Gururangan, Kun Nie, Masoom J Desai, Matthew Markert, Josef Parvizi, Thomas P Bleck, et al. Evaluating the clinical impact of rapid response electroencephalography: the DECIDE multicenter prospective observational clinical study. *Critical Care Medicine*, 48(9):1249–1257, 2020. doi: 10.1097/CCM.0000000000004428.

Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153, 2018. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2018.05.083.

Xiaoxi Wei, A. Aldo Faisal, Moritz Grosse-Wentrup, Alexandre Gramfort, Sylvain Chevallier, Vinay Jayaram, Camille Jeunet, Stylianos Bakas, Siegfried Ludwig,

Konstantinos Barmpas, Mehdi Bahri, Yannis Panagakis, Nikolaos Laskaris, Dimitrios A. Adamos, Stefanos Zafeiriou, William C. Duong, Stephen M. Gordon, Vernon J. Lawhern, Maciej Śliwowski, Vincent Rouanne, and Piotr Tempczyk. 2021 BEETL Competition: Advancing Transfer Learning for Subject Independence and Heterogenous EEG Data Sets. In Douwe Kiela, Marco Ciccone, and Barbara Caputo, editors, *Proceedings of the NeurIPS 2021 Competitions and Demonstrations Track*, volume 176 of *Proceedings of Machine Learning Research*, pages 205–219. PMLR, 06–14 Dec 2022.

Andreas Widmann, Erich Schröger, and Burkhard Maess. Digital filter design for electrophysiological data – a practical approach. *Journal of Neuroscience Methods*, 250:34–46, 2015. ISSN 0165-0270. doi: https://doi.org/10.1016/j.jneumeth.2014.08.002. Cutting-edge EEG Methods.

Daniel Wilson, Robin Tibor Schirrmeister, Lukas Alexander Wilhelm Gemein, and Tonio Ball. Deep Riemannian Networks for End-to-End EEG Decoding, 2024.

Jonathan R. Wolpaw, Dennis J. McFarland, Gregory W. Neat, and Catherine A. Forneris. An EEG-based brain-computer interface for cursor control. *Electroencephalography and Clinical Neurophysiology*, 78(3):252–259, 1991. ISSN 0013-4694. doi: https://doi.org/10.1016/0013-4694(91)90040-B.

Wei Wu, Yu Zhang, Jing Jiang, Molly V Lucas, Gregory A Fonzo, Camarin E Rolle, Crystal Cooper, Cherise Chin-Fatt, Noralie Krepel, Carena A Cornelssen, et al. An electroencephalographic signature predicts antidepressant response in major depression. *Nature Biotechnology*, 38(4):439–447, 2020. doi: https://doi.org/10.1038/s41587-019-0397-3.

Alba Xifra-Porxas, Arna Ghosh, Georgios D. Mitsis, and Marie-Hélène Boudrias. Estimating brain age from structural MRI and MEG data: Insights from dimensionality reduction techniques. *NeuroImage*, 231:117822, 2021. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2021.117822.

Or Yair, Mirela Ben-Chen, and Ronen Talmon. Parallel Transport on the Cone Manifold of SPD Matrices for Domain Adaptation. *IEEE Transactions on Signal Processing*, 67(7):1797–1811, 2019. doi: 10.1109/TSP.2019.2894801.

Yuzhe Yang, Yuan Yuan, Guo Zhang, Hao Wang, Ying-Cong Chen, Yingcheng Liu, Christopher G Tarolli, Daniel Crepeau, Jan Bukartyk, Mithri R Junna, et al. Artificial intelligence-enabled detection and assessment of Parkinson's disease using nocturnal breathing signals. *Nature Medicine*, 28(10):2207–2215, 2022. doi: https://doi.org/10.1038/s41591-022-01932-x.

Tal Yarkoni. The generalizability crisis. *Behavioral and Brain Sciences*, 45:e1, 2022. doi: 10.1017/S0140525X20001685.

Florian Yger, Maxime Berar, and Fabien Lotte. Riemannian Approaches in Brain-Computer Interfaces: A Review. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(10):1753–1762, 2017. doi: 10.1109/TNSRE.2016.2627016.

Paolo Zanini, Marco Congedo, Christian Jutten, Salem Said, and Yannick Berthoumieu. Transfer Learning: A Riemannian Geometry Framework With Applications to Brain–Computer Interfaces. *IEEE Transactions on Biomedical Engineering*, 65(5):1107–1116, 2018. doi: 10.1109/TBME.2017.2742541.

Hongyi Zhang and Suvrit Sra. First-order Methods for Geodesically Convex Optimization. In Vitaly Feldman, Alexander Rakhlin, and Ohad Shamir, editors, *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 1617–1638, Columbia University, New York, New York, USA, 23–26 Jun 2016. PMLR.

Appendix

A Chapter 2

A.1 Procrustes unpaired and the Variance- Threshold function

In this appendix, we show that the unpaired Procrustes method should be handled carefully. In particular, we show that the vectorized logarithm maps $\{\boldsymbol{z}_i\}_{i=1}^N$, defined in Equation (2.7), only span a subspace of $\mathbb{R}^{P(P+1)/2}$ regardless of the number of covariance matrices $\boldsymbol{\Sigma}_i$. Indeed, the rank of $\boldsymbol{Z} = \{\boldsymbol{z}_i,\, i = 1 \dots N_{\mathcal{S}}\} \in \mathbb{R}^{P(P+1)/2 \times N}$ is at most P. This implies that, computing the left singular vectors $\boldsymbol{U} \in \mathbb{R}^{P(P+1)/2 \times P(P+1)/2}$ of $\boldsymbol{Z}$, we get $\boldsymbol{U}^\top \boldsymbol{Z} \in \mathbb{R}^{P(P+1)/2 \times N}$ that has at maximum P rows with non zero variances. Thus, the other rows must be discarded using, for example, the class `VarianceThreshold` from the scikit-learn library [Pedregosa et al., 2011]. Otherwise, numerical issues can be encountered using functions like `StandardScaler` from the scikit-learn library. To prove these assertions, we begin by recalling the mixing model of $\boldsymbol{\Sigma}_i$ with no noise in the next hypothesis.

Assumption A.1.1 (Mixing model). We have a set of covariance matrices $\{\boldsymbol{\Sigma}_i\}_{i=1}^N$ that respect the following mixing model

$$\boldsymbol{\Sigma}_i = \boldsymbol{A}\boldsymbol{D}_i\boldsymbol{A}^\top \tag{A.1}$$

with $\boldsymbol{A} \in \mathbb{R}^{P \times P}$ an invertible mixing matrix and with $\boldsymbol{D}_i \in \mathbb{R}^{P \times P}$ diagonal with strictly positive elements.

This assumption induces that the Riemannian mean $\overline{\boldsymbol{\Sigma}}$ of $\{\boldsymbol{\Sigma}_i\}_{i=1}^N$ defined in Equation (1.11) and the associated Riemannian logarithmic mappings $\{\log(\overline{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\boldsymbol{\Sigma}_i\overline{\boldsymbol{\Sigma}}^{-\frac{1}{2}})\}_{i=1}^N$ have particular structures. These structures are computed in the following lemma.

Lemma A.1.2. *Knowing $\boldsymbol{A}$ and $\{\boldsymbol{D}_i\}_{i=1}^N$, the Riemannian mean $\overline{\boldsymbol{\Sigma}}$ has a closed form expression which is*

$$\overline{\boldsymbol{\Sigma}} = \boldsymbol{A}\bar{\boldsymbol{D}}\boldsymbol{A}^\top \tag{A.2}$$

with $\bar{\boldsymbol{D}}$ *diagonal with elements* $\bar{D}_{jj} = \left(\prod_{i=1}^{N}(D_i)_{jj}\right)^{\frac{1}{N}}$. *Furthermore, the Riemannian logarithmic mappings of any* $\overline{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\boldsymbol{\Sigma}_i\overline{\boldsymbol{\Sigma}}^{-\frac{1}{2}}$ *at identity is*

$$\log(\overline{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\boldsymbol{\Sigma}_i\overline{\boldsymbol{\Sigma}}^{-\frac{1}{2}}) = \overline{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\boldsymbol{A}^{-\top}\log(\bar{\boldsymbol{D}}^{-1}\boldsymbol{D}_i)\boldsymbol{A}^{\top}\overline{\boldsymbol{\Sigma}}^{\frac{1}{2}}. \tag{A.3}$$

The log function on the right-hand side of the equation applies the scalar logarithm on the diagonal elements.

Proof: By affine invariance of δ_R, we have $\overline{\boldsymbol{\Sigma}} = \boldsymbol{A}\bar{\boldsymbol{D}}\boldsymbol{A}^{\top}$ with

$$\bar{\boldsymbol{D}} = \underset{\boldsymbol{\Sigma}\in S_P^{++}}{\arg\min}\sum_{i=1}^{N}\delta_R^2(\boldsymbol{\Sigma}, \boldsymbol{D}_i). \tag{A.4}$$

From [**?**], $\bar{\boldsymbol{D}}$ is the unique solution of

$$\sum_{i=1}^{N}\log(\boldsymbol{D}_i^{-1}\bar{\boldsymbol{D}}) = 0_P. \tag{A.5}$$

It is readily checked that $\bar{\boldsymbol{D}}$ diagonal with elements $\bar{D}_{jj} = \left(\prod_{i=1}^{N}(D_i)_{jj}\right)^{\frac{1}{N}}$ satisfies the Equation A.5. Using these results and the matrix logarithm property $\log(\boldsymbol{E}\boldsymbol{B}\boldsymbol{E}^{-1}) = \boldsymbol{E}\log(\boldsymbol{B})\boldsymbol{E}^{-1}$ for any $\boldsymbol{E} \in \mathbb{R}^{P\times P}$ invertible and $\boldsymbol{B} \in \mathbb{R}^{P\times P}$ such that $\log(\boldsymbol{B})$ and $\log(\boldsymbol{E}\boldsymbol{B}\boldsymbol{E}^{-1})$ exist, we have

$$\begin{aligned}\log(\overline{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\boldsymbol{\Sigma}_i\overline{\boldsymbol{\Sigma}}^{-\frac{1}{2}}) &= \overline{\boldsymbol{\Sigma}}^{\frac{1}{2}}\log(\overline{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Sigma}_i)\overline{\boldsymbol{\Sigma}}^{-\frac{1}{2}} \\ &= \overline{\boldsymbol{\Sigma}}^{\frac{1}{2}}\log(\boldsymbol{A}^{-\top}\bar{\boldsymbol{D}}^{-1}\boldsymbol{D}_i\boldsymbol{A}^{\top})\overline{\boldsymbol{\Sigma}}^{-\frac{1}{2}} \\ &= \overline{\boldsymbol{\Sigma}}^{\frac{1}{2}}\boldsymbol{A}^{-\top}\log(\bar{\boldsymbol{D}}^{-1}\boldsymbol{D}_i)\boldsymbol{A}^{\top}\overline{\boldsymbol{\Sigma}}^{-\frac{1}{2}}.\end{aligned} \tag{A.6}$$

■

These structures induce that the vectorized logarithm maps $\{\boldsymbol{z}_i\}_{i=1}^{n}$ only span a subspace of $\mathbb{R}^{\frac{P(P+1)}{2}}$.

Proposition A.1.3. *The matrix* $\boldsymbol{Z} = \{\boldsymbol{z}_i,\, i = 1\dots N\} \in \mathbb{R}^{\frac{P(P+1)}{2}\times N}$ *has a maximum rank of* P.

Proof: We begin by defining the full vectorization counterpart of (2.7)

$$\tilde{\boldsymbol{z}}_i = \mathrm{vec}(\boldsymbol{S}\odot\log(\overline{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\boldsymbol{\Sigma}_i\overline{\boldsymbol{\Sigma}}^{-\frac{1}{2}})) \in \mathbb{R}^{P^2} \tag{A.7}$$

where vec is the operator that concatenates the columns of a given matrix. Then, by denoting $\boldsymbol{s} = \mathrm{vec}(\boldsymbol{S})$, we get

$$\begin{aligned}\tilde{\boldsymbol{z}}_i &= \mathrm{diag}(\boldsymbol{s})\mathrm{vec}(\log(\overline{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\boldsymbol{\Sigma}_i\overline{\boldsymbol{\Sigma}}^{-\frac{1}{2}})) \\ &= \mathrm{diag}(\boldsymbol{s})\mathrm{vec}(\overline{\boldsymbol{\Sigma}}^{\frac{1}{2}}\boldsymbol{A}^{-\top}\log(\bar{\boldsymbol{D}}^{-1}\boldsymbol{D}_i)\boldsymbol{A}^{\top}\overline{\boldsymbol{\Sigma}}^{-\frac{1}{2}}) \\ &= \mathrm{diag}(\boldsymbol{s})(\overline{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\boldsymbol{A}\otimes\overline{\boldsymbol{\Sigma}}^{\frac{1}{2}}\boldsymbol{A}^{-\top})\mathrm{vec}(\log(\bar{\boldsymbol{D}}^{-1}\boldsymbol{D}_i))\end{aligned} \tag{A.8}$$

where $\otimes$ is the Kronecker product. Denoting $\widetilde{\boldsymbol{Z}} = \{\tilde{\boldsymbol{z}}_i,\, i = 1 \dots N_{\mathcal{S}}\} \in \mathbb{R}^{P^2 \times N}$, it follows that

$$\widetilde{\boldsymbol{Z}} = \operatorname{diag}(\boldsymbol{s})(\overline{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\boldsymbol{A} \otimes \overline{\boldsymbol{\Sigma}}^{\frac{1}{2}}\boldsymbol{A}^{-\top}) \times \left[\operatorname{vec}(\log(\bar{\boldsymbol{D}}^{-1}\boldsymbol{D}_1)), \dots, \operatorname{vec}(\log(\bar{\boldsymbol{D}}^{-1}\boldsymbol{D}_n))\right].$$

Since $\operatorname{rank}(\overline{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\boldsymbol{A} \otimes \overline{\boldsymbol{\Sigma}}^{\frac{1}{2}}\boldsymbol{A}^{-\top}) = P^2$, we have that

$$\operatorname{rank}(\widetilde{\boldsymbol{Z}}) = \operatorname{rank}\left(\left[\operatorname{vec}(\log(\bar{\boldsymbol{D}}^{-1}\boldsymbol{D}_1)), \dots, \operatorname{vec}(\log(\bar{\boldsymbol{D}}^{-1}\boldsymbol{D}_n))\right]\right)$$

Since $\log(\bar{\boldsymbol{D}}^{-1}\boldsymbol{D}_i)$ has at most P non-zero elements,

$$\operatorname{rank}(\widetilde{\boldsymbol{Z}}) \leq P.$$

To conclude, the rows of $\boldsymbol{Z}$ are included in those of $\widetilde{\boldsymbol{Z}}$, hence

$$\operatorname{rank}(\boldsymbol{Z}) \leq P.$$

■

B Chapter 3

B.1 Proof of Lemma 3.4.1

First, we recall that the geodesic associated with the affine invariant metric from $\boldsymbol{\Sigma}$ to $\boldsymbol{\Sigma}'$ is

$$\boldsymbol{\Sigma} \sharp_\alpha \boldsymbol{\Sigma}' \triangleq \boldsymbol{\Sigma}^{1/2} \left(\boldsymbol{\Sigma}^{-1/2} \boldsymbol{\Sigma}' \boldsymbol{\Sigma}^{-1/2} \right)^\alpha \boldsymbol{\Sigma}^{1/2} \quad \text{for } \alpha \in [0, 1]. \tag{B.1}$$

Hence, $\boldsymbol{\Sigma} \sharp_\alpha \boldsymbol{I}_P = \boldsymbol{\Sigma}^{1-\alpha}$.

From Yair et al. [2019], the parallel transport of $\boldsymbol{\Sigma}'$ from $\boldsymbol{\Sigma}_1$ to $\boldsymbol{\Sigma}_2$ is

$$\boldsymbol{E} \boldsymbol{\Sigma}' \boldsymbol{E}^\top \quad \text{with} \quad \boldsymbol{E} \triangleq \boldsymbol{\Sigma}_1^{1/2} \left(\boldsymbol{\Sigma}_1^{-1/2} \boldsymbol{\Sigma}_2 \boldsymbol{\Sigma}_1^{-1/2} \right)^{1/2} \boldsymbol{\Sigma}_1^{-1/2} . \tag{B.2}$$

Hence, the parallel transport of $\boldsymbol{\Sigma}'$ from $\boldsymbol{\Sigma}$ to $\boldsymbol{\Sigma} \sharp_\alpha \boldsymbol{I}_P$ is $\boldsymbol{E} \boldsymbol{\Sigma}' \boldsymbol{E}^\top$ with

$$\begin{aligned} \boldsymbol{E} &\triangleq \boldsymbol{\Sigma}^{1/2} \left(\boldsymbol{\Sigma}^{-1/2} \boldsymbol{\Sigma}^{1-\alpha} \boldsymbol{\Sigma}^{-1/2} \right)^{1/2} \boldsymbol{\Sigma}^{-1/2} \\ &= \boldsymbol{\Sigma}^{1/2} \boldsymbol{\Sigma}^{-\alpha/2} \boldsymbol{\Sigma}^{-1/2} = \boldsymbol{\Sigma}^{-\alpha/2} \end{aligned} \tag{B.3}$$

which concludes the proof.

B.2 Cross-spectrum computation and preprocessing

Bartlett estimator

From Li et al. [2022], the features provided in the HarMNqEEG dataset have been computed using the Bartlett's estimator by averaging more than 20 consecutive and non-overlapping segments. Thus, data consist of cross-spectral matrices with a frequency range of $f_{\text{min}} = 1.17\,\text{Hz}$ to $f_{\text{max}} = 19.14\,\text{Hz}$, sampled at a resolution of $\Delta\omega = 0.39\,\text{Hz}$. These cross-spectral matrices are denoted $\boldsymbol{S}_{k,i}(\omega) \in \mathcal{H}_P^{++}$ where k is the site, i the participant and $\omega \in \{f_{\text{min}}, f_{\text{min}} + \Delta\omega, \ldots, f_{\text{max}}\}$.

Common average reference (CAR)

The cross-spectrum matrices $\boldsymbol{S}_i(\omega)$ were re-referenced from their original montages with a CAR:

$$\widetilde{\boldsymbol{S}}_{k,i}(\omega) \triangleq \boldsymbol{H}\boldsymbol{S}_{k,i}(\omega)\boldsymbol{H}^\top \tag{B.4}$$

where $\boldsymbol{H} \triangleq \boldsymbol{I}_P - \mathbf{1}_P\mathbf{1}_P^\top/P$.

Global Scale Factor (GSF)

Co-spectrum matrices were re-scaled with an individual scalar $\widehat{\zeta}_{k,i}$ that is calculated as the geometric mean of their power spectrum across sensors and frequencies:

$$\widehat{\zeta}_{k,i} \triangleq \exp\left(\frac{1}{N_\omega P}\sum_{\ell=0}^{N_\omega-1}\sum_{c=1}^{P}\log\left(\left(\widehat{\mathbf{S}}_{k,i}(f_{\min}+\ell\Delta\omega)\right)_{c,c}\right)\right) \tag{B.5}$$

where $N_\omega \triangleq {}^{f_{\max}-f_{\min}}/_{\Delta\omega}+1$. The GSF correction is then applied to the co-spectrum (the real part of the cross-spectrum) for all frequencies ω:

$$\boldsymbol{\Sigma}_{k,i}(\omega) \triangleq \Re\left(\widetilde{\boldsymbol{S}}_{k,i}(\omega)\right)/\widehat{\zeta}_{k,i}\ . \tag{B.6}$$

The $\boldsymbol{\Sigma}_{k,i}(\omega) \in \mathbb{S}_P^{++}$ are the features used the Section 3.5.

B.3 Figure 3.4 without normalization

Without `Re-center`:

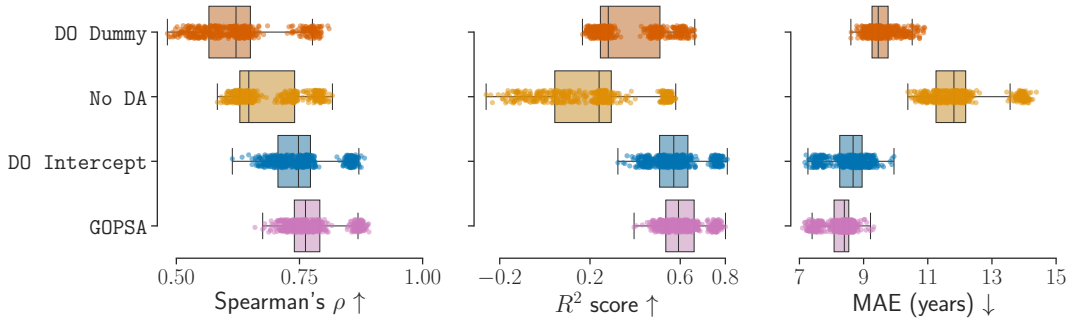

Figure B.1: **Performance of four methods on several source-target combinations for three metrics:** Spearman's $\rho \uparrow$ (left), R^2 score $\uparrow$ (middle) and Mean Absolute Error $\downarrow$ (right). `Re-center` was removed from the plot to better visualize the other methods. A box represents the concatenated results across all site combinations. One point corresponds to one split of one site combination.

With `Re-center`:

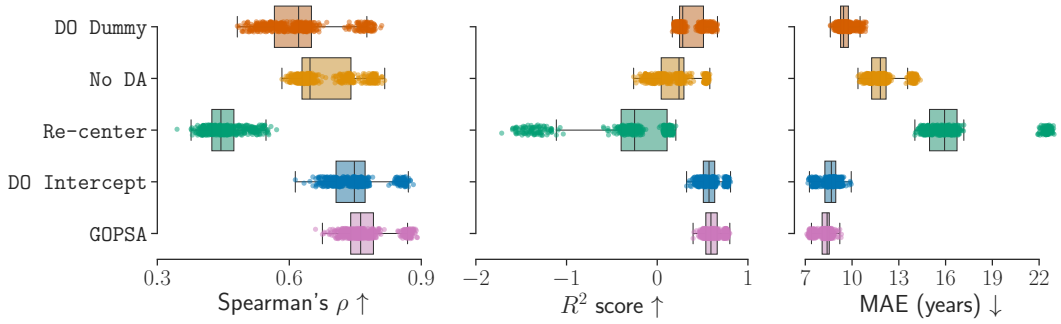

Figure B.2: **Performance of all methods on several source-target combinations for three metrics:** Spearman's $\rho \uparrow$ (left), R^2 score $\uparrow$ (middle) and Mean Absolute Error $\downarrow$ (right). A box represents the concatenated results across all site combinations. One point corresponds to one split of one site combination.

B.4 Boxplots of each source-target sites for the three metrics

The following figures represent the performance scores that are displayed in Table 3.1.

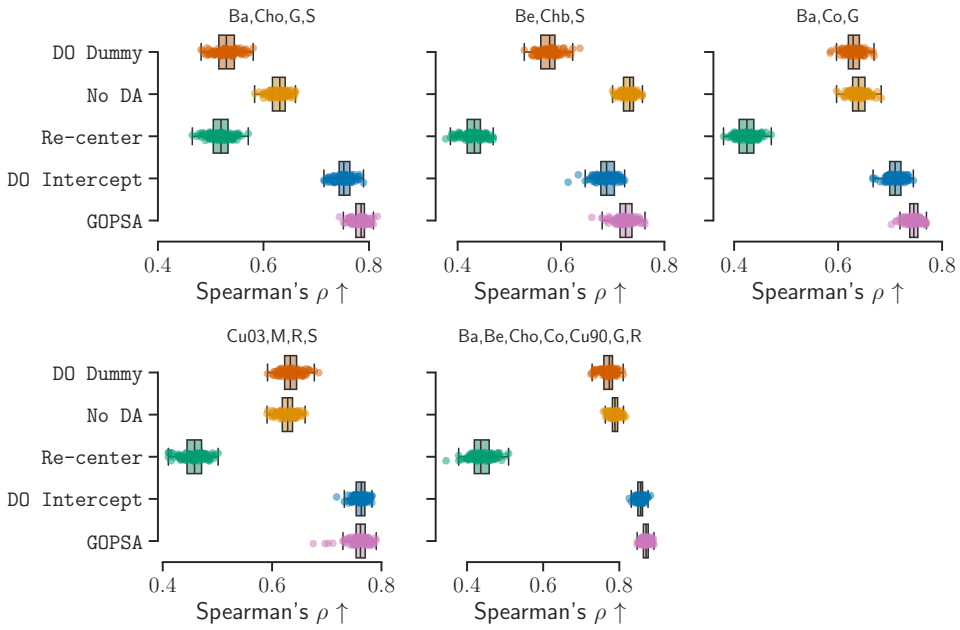

Figure B.3: **Spearman's $\rho \uparrow$ for every site combination.** One panel corresponds to the results of one site combination. One point corresponds to one split.

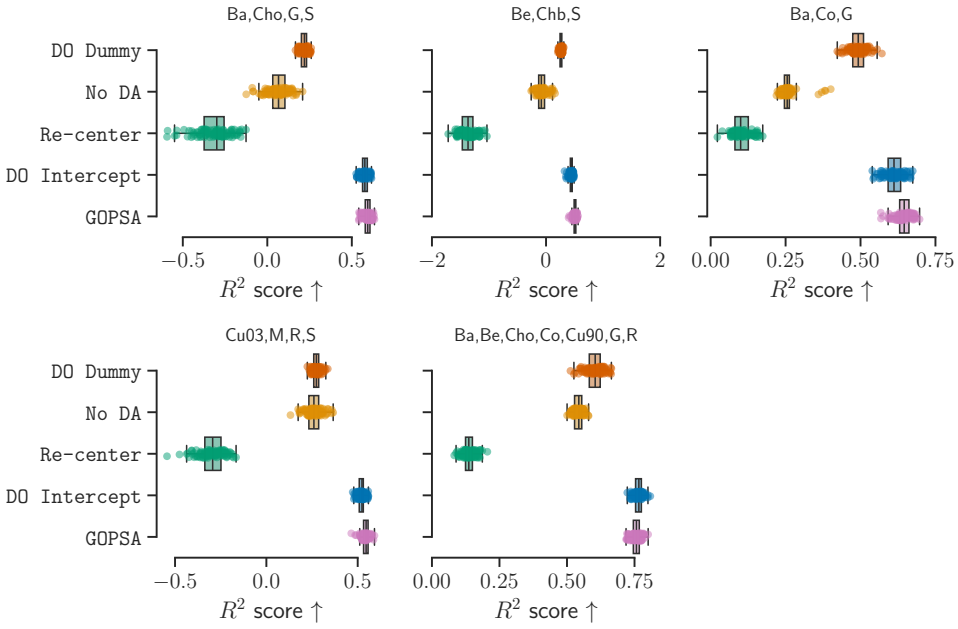

Figure B.4: **R^2 score ↑ for every site combination.** One panel corresponds to the results of one site combination. One point corresponds to one split.

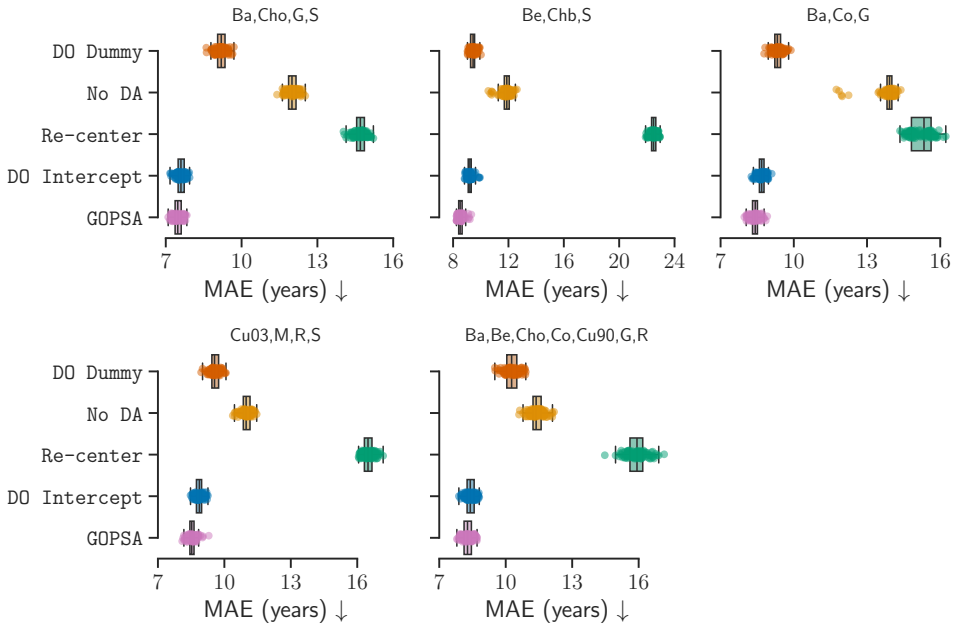

Figure B.5: **Mean Absolute Error ↓ for every site combination.** One panel corresponds to the results of one site combination. One point correspond to one split.

Synthèse en français

Les études en neurosciences rencontrent des défis dans la collecte de grandes bases de données, limitant ainsi l'utilisation de l'apprentissage statistique. L'intégration de données publiques peut être une solution, mais les données recueillies dans différents contextes présentent souvent des différences systématiques, appelées décalages de données (dataset shifts). Ces décalages, causés par des variations dans les sites d'enregistrement, le dispositif d'enregistrement ou les protocoles expérimentaux, compliquent l'application des méthodes d'apprentissage, qui exigent généralement des données d'entraînement et de test similaires. Cette thèse examine ces décalages dans les données M/EEG pour en comprendre les causes, leurs effets sur les modèles d'apprentissage, et propose des solutions adaptées au type de décalage pour améliorer la généralisation des modèles prédictifs.

Dans un premier temps, nous avons analysé les décalages survenant dans les enregistrements M/EEG en lien avec l'activité cérébrale, l'anatomie ou la configuration des capteurs. Pour harmoniser la distribution des données, nous avons utilisé une approche riemannienne d'alignement des données et l'avons adaptée à la régression non supervisée. Pour évaluer l'efficacité de l'alignement, nous avons réalisé des expériences sur des données simulées et réelles. Nous avons montré que la performance des modèles d'apprentissage peut être affectée par ces décalages et qu'elle peut être améliorée en alignant les distributions de données.

Dans la deuxième partie, nous nous sommes concentrés sur les décalages survenant à la fois dans les données M/EEG et la distribution de la variable à prédire y. Cette situation est courante dans les études cliniques où les données sont recueillies dans différents sites et auprès de différentes populations. Dans ce contexte, l'alignement proposé précédemment n'est pas suffisant pour traiter les décalages. Nous avons proposé une nouvelle méthode pour aborder l'adaptation de domaine dans des situations où les domaines sources ont des distributions de y distinctes. Cette méthode exploite la structure géométrique de la variété riemannienne pour apprendre conjointement un opérateur de recentrage spécifique au domaine et le modèle de régression. Nous avons réalisé des comparaisons empiriques sur la généralisation inter-sites des modèles de prédiction de l'âge avec des données EEG provenant d'un grand ensemble de données multinationales. L'approche proposée a significativement amélioré la généralisation des modèles à travers les sites par

rapport aux méthodes de référence.

Finalement, nous avons abordé le problème des différents dispositifs d'enregistrement EEG. Le nombre et les positions variables des capteurs rendent difficile la comparaison des données provenant de différents dispositifs, et rendent même impossible l'utilisation directe des méthodes d'apprentissage. Pour remédier à cela, nous avons proposé une approche non supervisée exploitant la physique des signaux EEG : nous avons interpolé les canaux EEG de diverses configurations sur des positions fixes avec l'interpolation basée sur la physique de la propagation électromagnétique. Une évaluation comparative avec d'autres méthodes a été effectuée sur six bases de données publiques pour la classification d'imagerie motrice main droite/gauche. L'interpolation s'est montrée similaire ou meilleure que les autres méthodes.

Les contributions de cette thèse visent à améliorer la généralisation des modèles d'apprentissage appliqués aux données M/EEG sous différents aspects et situations. Le but était de mieux comprendre les décalages de données M/EEG et de proposer des approches pour atténuer leurs effets dans des scénarios réalistes.