Unsupervised component analysis for neuroimaging data

Hugo Richard

universite PARIS-SACLAY

Unsupervised component analysis for neuroimaging data

Analyse non-supervisée en composante de données d'imageries neuronales

Thèse de doctorat de l'université Paris-Saclay

École doctorale n° 580, Sciences et Technologies de l'Information et de la Communication (ED STIC)
Spécialité de doctorat : Mathématiques et Informatique
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 **Bertrand THIRION**, directeur de recherche.

Thèse soutenue à Paris-Saclay, le 20 décembre 2021, par

Hugo RICHARD

Composition du jury

Tülay ADALI Directrice de recherche, University of Maryland Baltimore County	Présidente & Rapporteur
Moritz GROSSE-WENTRUP Directeur de recherche, University of Vienna	Rapporteur & Examinateur
Christian JUTTEN Directeur de recherche, Université Grenoble Alpes	Examinateur
Aapo HYVÄRINEN Directeur de recherche, University of Helsinki	Examinateur
Matthieu KOWALSKI Maître de conférences, Université Paris Saclay	Examinateur
Bertrand THIRION Directeur de recherche, Université Paris Saclay, INRIA	Directeur de thèse

THESE DE DOCTORAT

NNT : 2021UPASG115

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

Titre : Analyse non-supervisée en composante de données d'imageries neuronales
Mots clés : neuro-imagerie, analyse en composantes, apprentissage automatique

Résumé : Cette thèse d'informatique et de mathématiques s'applique au domaine des neurosciences, et plus particulièrement aux recherches sur la modélisation de l'activité cérébrale humaine par électrophysiologie et imagerie. Dans ce champ, la tendance est actuellement d'expérimenter avec des stimuli naturels, comme le visionnage d'un film ou l'écoute d'une piste audio, et non plus avec des stimuli étroitement contrôlés mais outrageusement simples. L'analyse de ces stimuli « naturels » et de leurs effets demande toutefois de disposer d'une immense quantité d'images, par ailleurs très coûteuses. Sans outils mathématique, identifier l'activité neuronale à partir des données est quasi impossible. Toutefois, ces stimuli sont compliqués à modéliser et à analyser, car l'utilisation de méthodes fondées sur des régressions est limitée par la difficulté de modéliser les stimuli. C'est ce qui motive l'utilisation de méthodes non-supervisées qui ne font pas d'hypothèses sur ce qui déclenche les activations neuronales.

Dans cette thèse, nous considérons d'abord le cas du modèle de réponse partagée (MRP), dans lequel les sujets sont supposés partager une réponse commune. Ce modèle est utile pour réduire la dimension des données, mais son entraînement est coûteux pour les données d'imagerie fonctionnelle (IRMf) dont la dimension peut être immense. Nous présentons une version bien plus rapide et beaucoup plus économe en mémoire. Mais le MRP fait des hypothèses irréalistes sur les données d'imagerie.

Des hypothèses plus réalistes sont utilisées dans l'analyse en composantes indépendantes (ACI) mais cette méthode est difficile à généraliser aux jeux de données qui contiennent plusieurs sujets. Nous proposons alors une extension de l'ACI appelée ACI multi-vue, fondée sur le principe de maximum de vraisemblance et qui convient à des jeux de données multi-sujets. L'ACI multi-vue a une vraisemblance en forme fermée qui peut être maximisée efficacement. Toutefois, cette méthode suppose la même quantité de bruit pour tous les sujets.

Nous présentons donc l'ACI partagée, une généralisation de l'ACI multi-vue qui s'accompagne d'un modèle de bruit plus général. Contrairement à presque tous les modèles fondés sur l'ACI, l'ACI partagée peut séparer des sources gaussiennes et non gaussiennes et propose une estimation optimale des sources communes (au sens des moindres carrés), qui pondère chaque sujet en fonction de son niveau de bruit estimé. En pratique, l'ACI partagée et l'ACI multi-vue permettent d'obtenir, en magnéto-encéphalographie et en IRMf, une estimation plus fiable de la réponse commune que leurs concurrents.

Enfin, nous utilisons l'ACI comme base pour faire de l'augmentation de données. Plus précisément, nous présentons l'ACI conditionnelle, une méthode d'augmentation de données qui exploite la grande quantité de données d'IRMf non étiquetées pour construire un modèle génératif en utilisant seulement un petit nombre de données étiquetées. L'ACI conditionnelle permet d'augmenter de façon appréciable la précision du décodage sur huit grands jeux de données d'IRMf.

Nos principaux apports nous semblent consister dans l'accélération de l'entraînement du MRP ainsi que dans l'introduction de deux modèles plus réalistes pour l'analyse de l'activité cérébrale de sujets exposés à des stimuli naturels : l'ACI multi-vue et l'ACI partagée. Enfin, nos résultats sont prometteurs concernant l'utilisation de l'ACI pour faire de l'augmentation de données.

Nous présentons pour finir quelques pistes qui pourraient guider des travaux ultérieurs. D'un point de vue pratique, des modifications mineures de nos méthodes pourraient permettre l'analyse des données d'imagerie obtenues sur des sujets au repos en faisant l'hypothèse d'une organisation spatiale partagée à la place d'une réponse partagée. D'un point de vue théorique, les travaux futurs pourraient se concentrer sur la compréhension de la façon dont la réduction de dimensions et l'identification de la réponse partagée peuvent être réalisées conjointement.

Title : Unsupervised component analysis for neuroimaging data
Keywords : neuro-imaging, component analysis, machine learning

Abstract :

This thesis in computer science and mathematics is applied to the field of neuroscience, and more particularly to the mapping of brain activity based on imaging electrophysiology. In this field, a rising trend is to experiment with naturalistic stimuli such as movie watching or audio track listening, rather than tightly controlled but outrageously simple stimuli. However, the analysis of these "naturalistic" stimuli and their effects requires a huge amount of images that remain hard and costly to acquire. Without mathematical modeling, the identification of neural signal from the measurements is very hard if not impossible. However, the stimulations that elicit neural activity are challenging to model in this context, and therefore, the statistical analysis of the data using regression-based approaches is difficult. This has motivated the use of unsupervised learning methods that do not make assumptions about what triggers brain activations in the presented stimuli.

In this thesis, we first consider the case of the shared response model (SRM), where subjects are assumed to share a common response. While this algorithm is useful to perform dimension reduction, it is particularly costly on functional magnetic resonance imaging (fMRI) data where the dimension can be very large. We considerably speed up the algorithm and reduce its memory usage. However, SRM relies on assumptions that are not biologically plausible.

In contrast, independent component analysis (ICA) is more realistic but not suited to multi-subject datasets. In this thesis, we present a well-principled method called MultiViewICA that extends ICA to datasets containing multiple subjects. MultiViewICA is a maximum likelihood estimator. It comes with a closed-form likelihood that can be efficiently optimized. However, it assumes the same amount of noise for all subjects.

We therefore introduce ShICA, a generalization of MultiViewICA that comes with a more general noise model. In contrast to almost all ICA-based models, ShICA can separate Gaussian and non-Gaussian sources and comes with a minimum mean square error estimate of the common sources that weights each subject according to its estimated noise level. In practice, MultiViewICA and ShICA yield on magnetoencephalography and functional magnetic resonance imaging a more reliable estimate of the shared response than competitors.

Lastly, we use independent component analysis as a basis to perform data augmentation. More precisely, we introduce CondICA, a data augmentation method that leverages a large amount of unlabeled fMRI data to build a generative model for labeled data using only a few labeled samples. CondICA yields an increase in decoding accuracy on eight large fMRI datasets.

Our main contributions consist in the reduction of SRM's training time as well as in the introduction of two more realistic models for the analysis of brain activity of subjects exposed to naturalistic stimuli : MultiViewICA and ShICA. Lastly, our results showing that ICA can be used for data augmentation are promising.

In conclusion, we present some directions that could guide future work. From a practical point of view, minor modifications of our methods could allow the analysis of resting state data assuming a shared spatial organization instead of a shared response. From a theoretical perspective, future work could focus on understanding how dimension reduction and shared response identification can be achieved jointly.

UNSUPERVISED COMPONENT ANALYSIS FOR NEUROIMAGING DATA

HUGO RICHARD

PhD thesis

SYNTHÈSE

Cette thèse d'informatique et de mathématiques s'applique au domaine des neurosciences, et plus particulièrement aux recherches sur la modélisation de l'activité cérébrale humaine par électrophysiologie et imagerie. Dans ce champ, la tendance est actuellement d'expérimenter avec des stimuli naturels, comme le visionnage d'un film ou l'écoute d'une piste audio, et non plus avec des stimuli étroitement contrôlés mais outrageusement simples. L'analyse de ces stimuli « naturels » et de leurs effets demande toutefois de disposer d'une immense quantité d'images, par ailleurs très coûteuses. Sans outils mathématique, identifier l'activité neuronale à partir des données est quasi impossible. Toutefois, ces stimuli sont compliqués à modéliser et à analyser, car l'utilisation de méthodes fondées sur des régressions est limitée par la difficulté de modéliser les stimuli. C'est ce qui motive l'utilisation de méthodes non-supervisées qui ne font pas d'hypothèses sur ce qui déclenche les activations neuronales.

Dans cette thèse, nous considérons d'abord le cas du modèle de réponse partagée (MRP), dans lequel les sujets sont supposés partager une réponse commune. Ce modèle est utile pour réduire la dimension des données, mais son entraînement est coûteux pour les données d'imagerie fonctionnelle (IRMf) dont la dimension peut être immense. Nous présentons une version bien plus rapide et beaucoup plus économe en mémoire. Mais le MRP fait des hypothèses irréalistes sur les données d'imagerie.

Des hypothèses plus réalistes sont utilisées dans l'analyse en composantes indépendantes (ACI) mais cette méthode est difficile à généraliser aux jeux de données qui contiennent plusieurs sujets. Nous proposons alors une extension de l'ACI appelée ACI multi-vue, fondée sur le principe de maximum de vraisemblance et qui convient à des jeux de données multi-sujets. L'ACI multi-vue a une vraisemblance en forme fermée qui peut être maximisée efficacement. Toutefois, cette méthode suppose la même quantité de bruit pour tous les sujets.

Nous présentons donc l'ACI partagée, une généralisation de l'ACI multi-vue qui s'accompagne d'un modèle de bruit plus général. Contrairement à presque tous les modèles fondés sur l'ACI, l'ACI partagée peut séparer des sources gaussiennes et non gaussiennes et propose une estimation optimale des sources communes (au sens des moindres carrés), qui pondère chaque sujet en fonction de son niveau de bruit estimé. En pratique, l'ACI partagée et l'ACI multi-vue permettent d'obtenir, en magnéto-encéphalographie et en IRMf, une estimation plus fiable de la réponse commune que leurs concurrents.

Enfin, nous utilisons l'ACI comme base pour faire de l'augmentation de données. Plus précisément, nous présentons l'ACI conditionnelle, une méthode d'augmentation de données qui exploite la grande quantité de données d'IRMf non étiquetées pour construire un modèle génératif en utilisant seulement un petit nombre de données étiquetées. L'ACI conditionnelle permet d'augmenter de façon appréciable la précision du décodage sur huit grands jeux de données d'IRMf.

Nos principaux apports nous semblent consister dans l'accélération de l'entraînement du MRP ainsi que dans l'introduction de deux modèles plus réalistes pour l'analyse de l'activité cérébrale de sujets exposés à des stimuli naturels : l'ACI multi-vue et l'ACI partagée. Enfin, nos résultats sont prometteurs concernant l'utilisation de l'ACI pour faire de l'augmentation de données.

Nous présentons pour finir quelques pistes qui pourraient guider des travaux ultérieurs. D'un point de vue pratique, des modifications de nos méthodes pourraient permettre l'analyse des données d'imagerie obtenues sur des sujets au repos en faisant l'hypothèse d'une organisation spatiale partagée à la place d'une réponse partagée. Une autre extension possible serait de prendre en compte le décalage temporel entre sujets que l'on observe en magnéto-encéphalographie. D'un point de vue théorique, les travaux futurs pourraient se concentrer sur la compréhension de la façon dont la réduction de dimensions et l'identification de la réponse partagée peuvent être réalisées conjointement.

ACKNOWLEDGMENTS

Thank you to Pr. Moritz Große-Wentrup and Pr. Tülay Adali for reviewing my thesis and to Pr. Sylvain Chevalier, Pr Mathieu Kowalski, Pr Aapo Hyvärinen and Pr Alexandre Gramfort for being part of the jury. Their insightful remarks and comments have helped me improve the quality of this thesis.

Je remercie chaleuresement mon directeur de thèse Bertrand Thirion pour avoir largement contributé à mon épanouissement durant ces années de thèse. Merci Bertrand de m'avoir appris à écrire des papiers scientifiques, à présenter des résultats et à aborder des problèmes complexes avec méthode. Merci aussi pour ta très grande disponibilité et ton engagement dans mon travail.

Merci du fond du coeur Pierre, pour m'avoir transmis ta grande aisance à formaliser les problèmes, tes connaissances sur l'ICA, sur python et pour avoir investi du temps sur mes problèmes. I would also like to give special thanks to Aapo, who carefully reviewed our proofs and provided deep insights and always well-reasoned comments. It has been an honour to work with you. Merci Alex, la base de code et la pipeline d'intégration continue ne serait pas aussi propre sans ton aide. I deeply thank all other co-authors Luigi, Ana-Luisa, Jonathan, Lucas, Guillaume for your help in writing and great scientific discussions.

Merci à tous mes collègues et amis pour leur soutien. Merci à ceux qui ont partagé avec moi les joies du périple vers Neurospin: Thomas B., Jérome Alexis, Hamza, Binh, Antonia, Zac, Jerome, Arthur, Kamalaker. Merci aux volleyeurs Alexis T., Thomas M., Raphael, Louis, Maeliss, Demian qui m'ont permis de pratiquer au labo le sport que j'aime le plus au monde. Merci aussi à Olivier et Gaël pour ces discussions stimulantes. Merci à toute l'équipe. Merci à Corinne et Stephanie pour m'avoir guidé dans le labyrinthe administratif.

Merci à ma famille. Merci à ma mère, à mon père et à mon frère que j'aime fort et qui m'ont permi d'arriver jusque-là. Un grand merci à mon grand-père maternel Papou, qui malgré son grand âge a contribué significativement à l'écriture du résumé de cette thèse. Merci aussi à mon grand-père paternel Papet pour ses encouragements et son soutient. Enfin, merci à Laurie ma fiancée que j'admire et que j'aime qui m'a toujours accompagné, inspiré et soutenu dans les moments faciles comme dans les plus durs.

CONTENTS

ACRONYMS

BOLD	Blood oxygenated level dependent
EEG	Electroencephalography
fMRI	Functional magnetic resonance imaging
ICA	Independent component analysis
MEG	Magnetoencephalography
ML	Maximum likelihood
MMSE	Minimum mean squared error
MNI template	Montreal Neuroscience Institute template
MVICA	MultiViewICA
PCA	Principal component analysis
RAM	Random access memory
ShICA	Shared ICA
SRM	Shared response model
SVD	Singular value decomposition

NOTATIONS

We write vectors in bold letter $\mathbf{v}$ and scalars in lower case a. Upper case letters M are used to denote matrices. We denote $|W|$ the absolute value of the determinant of W. $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ means that $\mathbf{x} \in \mathbb{R}^k$ follows a multivariate normal distribution of mean $\boldsymbol{\mu} \in \mathbb{R}^k$ and covariance $\Sigma \in \mathbb{R}^{k \times k}$. When $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, its density is given by $\mathbf{x} \to \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \Sigma)$. The j, j entry of a diagonal matrix Σ_i is denoted Σ_{ij}, the j entry of $\mathbf{y}_i$ is denoted y_{ij}. δ is the Kronecker delta. We use the usual scalar product for matrices $\langle A, B \rangle = \mathrm{tr}(A^\top B)$ and the associated norm is denoted $\|A\| = \sqrt{\langle A, A \rangle}$. Vectors can be seen as tall matrices and therefore the scalar product and the norm are the same as for matrices. The gradient of a real function $f(\mathbf{x}) \in \mathbb{R}$ is denoted $\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}}$ and is seen as a column vector. The Jacobian of a vector valued function $\mathbf{f}(\mathbf{x})$ is denoted $\frac{\partial \mathbf{f}(\mathbf{x})}{\partial \mathbf{x}}$ and is a matrix such that the line j is given by $\frac{\partial \mathbf{f}(x_j)}{\partial \mathbf{x}}^\top$ where x_j is the j-th coordinate of $\mathbf{x}$.

1 OVERVIEW

1.1 INTRODUCTION

1.1.1 *Controlled experiments in cognitive brain imaging*

When a subject is reading a sentence or when she is listening to a sentence, her brain activity is expected to differ. In order to measure where it differs and by which amount, one can perform a controlled experiment. The controlled experiment comes with a design matrix that describes the features driving her brain activation across time. In the above case, the occurrences of the subject listening to a sentence or reading a sentence are encoded in a design matrix. Then, a model is introduced to explain how the design matrix relates to brain activity. A simple model can consider that brain activity only differs when the task differs and therefore have a stereotypical representation of brain activity when she is reading a sentence and another one when she is listening to a sentence. While this model may simplify the reality (two repetitions of the same task would never yield exactly the same pattern or amount of brain activity), it is easy to interpret: the two stereotypical representations of each task can easily be compared and analyzed.

This procedure naturally extends to multiple subjects. Indeed, a model can be fit independently for each subject using the same design matrix.

1.1.2 *Naturalistic stimuli*

While controlled experiments give some insights about brain functionality, the subjects' experience is far from their every-day life. Naturalistic stimuli are meant to overcome this issue. Example of naturalistic stimuli include movie watching, music listening or resting (subjects are just asked to lie still in the scanner without further instruction). While there is a broad interest in understanding how the brain reacts in such ecological conditions, the recorded brain activity is difficult to analyze. In particular, design matrices are notoriously difficult to construct for naturalistic stimuli.

1.1.3 *Component analysis*

A possible solution is to learn the design matrix as part of the model. The widely used independent component analysis applied on the data of one subject extracts a set of components that are maximally independent. Such components give a plausible design matrix as each component may be seen as a different set of features driving brain activity. However, many questions remain. How do we efficiently generalize such methods to multi-subject data ? How to measure their performance ? Why is it useful to construct a generative model for neuroimaging data ? In this thesis, we develop well principled unsupervised methods for component analysis of neuroimaging data.

1.2 ORGANIZATION OF THE MANUSCRIPT

In part I, we give some background on statistical learning and optimization (chapter 2), neuroscience (chapter 3) and unsupervised methods popular in neuroimaging (chapter 4).

The parts II, III, IV and V highlight four different contributions that led to different publications.

1.2.1 *Fast shared response model for fMRI data*

When subjects are exposed to the same stimuli, their brain activity likely exhibits some common, or shared response. Our goal is to recover this shared response. The *shared response model* is one possible solution to this problem. However, the algorithm used in the shared response model does not scale well with the size of input . This is problematic because fMRI data have a very large size. Indeed, the fMRI data of each subject have a dimension on the order of $p = 10^5$ and a number of samples on the order of $n = 10^3$. Therefore, there is a need for faster algorithms. This will be presented in part II.

1.2.2 *MultiViewICA for neuroimaging data*

In order to obtain a meaningful shared response, we need to impose constraints on the model. In the shared response model, the linear combination of the shared sources is done under orthogonality constraints which are often deemed not biologically plausible. A more biologically plausible constraint is to assume independence of the recovered responses. However, most popular methods using this assumption are partially heuristic when multiple subjects are involved. In part III, we introduce MultiViewICA: a method based on the maximum likelihood principle that can be efficiently optimized. In practice our method is able to recover more reliable estimate of the shared response on fMRI and MEG data than competitors.

PUBLISHED WORK (Spotlight) H. Richard et al. “Modeling Shared responses in Neuroimaging Studies through MultiView ICA.” In: *Advances in Neural Information Processing Systems 33*. Dec. 2020

1.2.3 *Shared ICA for neuroimaging data*

The model used in MultiViewICA allows the response of each subject to differ from the stereotypical response. However, it is assumed that the deviation is on the same order of magnitude for all subjects and all components. In Shared ICA, we use a more general model that allows to model different deviations from the stereotypical response depending on the subject considered and/or on the response. Importantly, the difference between subjects can be used as an additional source of information to recover an even better estimate of the shared response. In practice, we observe that Shared ICA improves upon MultiViewICA. This work will be presented in part IV.

PUBLISHED WORK H. Richard et al. "Shared Independent Component Analysis for Multi-Subject Neuroimaging." In: *Advances in Neural Information Processing Systems 33*. Dec. 2021

1.2.4 *Conditional ICA*

Our results suggest that ICA is a good generative model for fMRI data. Can it be used to perform data augmentation ? We design a data augmentation method, that leverages the large amount of (unlabeled) resting state data to generate realistic fake task data from a small set of images. Our data augmentation method yields an improvement in classification accuracy on eight large datasets. This will be discussed in part V.

PUBLISHED WORK Badr Tajini, Hugo Richard, and Bertrand Thirion. "Functional Magnetic Resonance Imaging data augmentation through conditional ICA." In: *MICCAI* (2021) (Oral, co-first authorship with equal contribution)

1.3 CHAPTER ORDERING

The appendices can be skipped at first read. Chapters 2, 3 and 4 present the necessary background. The reader more interested in the theory should first read chapters 5, 7, 9 and 11. Chapters 6, 8, 10 and 12 focus on practical results.

Part I

BACKGROUND CONCEPTS

2 STATISTICAL LEARNING AND OPTIMIZATION

In this chapter, we present the probabilistic modeling and optimization background needed for the present thesis. The material presented in section 2.1 is inspired from the section 4 of [87], the section 1.1 of [1], the chapter 17 of [54] and [140]. The material presented in section 2.2 comes from the chapter 9 of [24] and from [106].

2.1 PROBABILISTIC GENERATIVE MODELS

In a probabilistic generative modeling framework, learning from experiments means identifying the underlying process that generates the data we observe.

More formally, consider $\mathcal{X} = \mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)} \in \mathbb{R}^k$, n random vectors with joint density $\nu^*_{\mathcal{X}}$ and consider the available data $X \in \mathbb{R}^{k,n}$ as an observation of $\mathcal{X}$. We also say that data X are *generated from* $\nu^*_{\mathcal{X}}$. The broad goal of probabilistic generative modeling is to recover $\nu^*_{\mathcal{X}}$ from X. When samples $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}$ are independent and identically distributed, $\nu^*_{\mathcal{X}} = \otimes_{i=1}^{n} \nu^*_{\mathbf{x}}$ so that X can be seen as n observations of the random variable $\mathbf{x}$ with density $\nu^*_{\mathbf{x}}$. In the remaining of the thesis, we use the same notation $\mathbf{x}^{(i)}$ for the random variable associated to the i-th sample and its observation (the i-th column of X). In addition, we use ν^* to denote both $\nu^*_{\mathbf{x}}$ and $\nu^*_{\mathcal{X}}$ depending on whether samples are independent and identically distributed or not.

In practice, we assume a model for the true density ν^* meaning that we assume that ν^* belongs to a family of densities $\mathcal{F}$. When it is indeed true that $\nu^* \in \mathcal{F}$, we say that *the model holds*.

Often, we assume $\mathcal{F}$ to be a set of parametric densities so that any density in $\mathcal{F}$ can be written as ν_θ where $\theta \in \Theta$ is a set of parameters and Θ is the set of all possible θ. When the model holds, there exists an optimal set of parameters θ_* such that $\nu^* = \nu_{\theta_*}$. The goal is then to find θ_*.

Before we even try to find θ_*, it is instructive to wonder whether the problem is well defined. Ideally we would like our model to be such that if $\nu_{\theta_1} = \nu_{\theta_2}$ then $\theta_1 = \theta_2$. When this is the case, we say that the model is *identifiable*.

In all the proofs in this section, we assume that the integration and differentiation operators can always be exchanged and that all quantities introduced are well defined. The set of parameters θ is viewed as a vector.

Note that our definition of a model includes the latent variable model (which specifies the dependencies across the different random variables) and the densities of each random variable. However it does not include the estimation procedure. Including the densities in the model allows us to state claims like : "this model is identifiable" which would not always be possible if the densities were not included. Yet, a model is most useful when it comes with an efficient estimation procedure so we often present together the model and the algorithm used to estimate its parameters.

2.1.1 *Desirable properties of estimators*

An *estimator* $\hat{\theta}_n$ of θ_* is a function of the observations $\mathcal{X} = \mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}$ that aims at finding θ_*. An estimator is a random variable, as it depends on $\mathcal{X}$. Therefore, it can almost never be perfectly accurate and that is why we need a criterion to measure its inaccuracy.

A common choice is the mean squared error criterion given by:

$$\mathbb{E}[\|\hat{\theta}_n - \theta_*\|^2] = \mathbb{E}[\|\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n] + \mathbb{E}[\hat{\theta}_n] - \theta_*\|^2] \tag{2.1}$$

$$= \mathbb{E}[\|\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n]\|^2] + \|\mathbb{E}[\hat{\theta}_n] - \theta_*\|^2 \tag{2.2}$$

$$= \mathrm{tr}(\mathbb{V}(\hat{\theta}_n)) + \|\mathbb{E}[\hat{\theta}_n] - \theta_*\|^2 \tag{2.3}$$

where the right hand side in equation 2.3 gives the *bias-variance decomposition*. The left term is the trace of the *covariance* $\mathbb{V}(\hat{\theta}_n) = \mathrm{Cov}(\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n], \hat{\theta}_n - \mathbb{E}[\hat{\theta}_n])$ where $\mathrm{Cov}(\mathbf{a}, \mathbf{b}) = \mathbb{E}[(\mathbf{a} - \mathbb{E}[\mathbf{a}])(\mathbf{b} - \mathbb{E}[\mathbf{b}])^\top]$ and the right term is the squared norm of the *bias* given by $\mathbb{E}[\hat{\theta}_n] - \theta_*$.

The norm of the bias can be minimized exactly and such estimators that achieve $\mathbb{E}[\hat{\theta}_n] - \theta_* = 0$ are called *unbiased*.

In Example 1, we study the bias of the sample mean and sample variance. The sample mean is shown to be unbiased. In contrast, the sample variance is biased: we show how to correct the estimator of the variance so that it becomes unbiased.

Example 1 (Biased and unbiased estimate of the parameters of a 1D Gaussian)**.** *Consider* n *observations* $x^{(1)}, \dots x^{(n)}$ *of* x *with mean* μ_* *and variance* σ_*^2. *Consider the sample mean:* $\hat{\mu}_e = \frac{1}{n}\sum_{i=1}^n x^{(i)}$. *This estimate is unbiased as* $\mathbb{E}[\hat{\mu}_e] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[x^{(i)}] = \mu_*$.

Consider the sample variance: $\hat{\sigma}_e^2 = \frac{1}{n}\sum_{i=1}^n (x^{(i)} - \sum_{z=1}^n \frac{1}{n} x^{(z)})^2$.

We have

$$\mathbb{E}[\hat{\sigma}_e^2] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[(x^{(i)} - \sum_{z=1}^{n}\frac{1}{n}x^{(z)})^2] \tag{2.4}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[(x^{(i)})^2 - 2x^{(i)}(\sum_{z=1}^{n}\frac{1}{n}x^{(z)}) + (\sum_{z=1}^{n}\frac{1}{n}x^{(z)})^2] \tag{2.5}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[(x^{(i)})^2 - 2x^{(i)}(\sum_{z=1}^{n}\frac{1}{n}x^{(z)}) + \frac{1}{n^2}\sum_{y=1,z=1}^{n}x^{(y)}x^{(z)}] \tag{2.6}$$

$$= \frac{1}{n}\sum_{i=1}^{n}(\sigma_*^2 + \mu_*^2 - \frac{2}{n}\sigma_*^2 - 2\mu_*^2 + \frac{1}{n^2}(\sum_{z=1}^{n}\sigma_*^2 + n^2\mu_*^2)) \tag{2.7}$$

$$= \sigma_*^2 - \frac{1}{n}\sigma_*^2 \tag{2.8}$$

$$= \frac{n-1}{n}\sigma_*^2 \tag{2.9}$$

so the sample covariance is biased. In contrast, if we consider the estimator

$$\hat{\sigma}_u^2 = \frac{n}{n-1}\hat{\sigma_e}^2 \tag{2.10}$$

we can see that it is unbiased.

However the variance of an estimator cannot be arbitrarily low as shown by Proposition 1 (Cramer-Rao bound).

Proposition 1 (Cramer-Rao bound)**.** *Let* $\hat{\theta}_n$ *be an estimator of* θ_**. Then,* $\mathbb{V}(\hat{\theta}_n) \succeq \frac{\partial \mathbb{E}[\hat{\theta}_n]}{\partial \theta} I_n(\theta_*)^{-1} (\frac{\partial \mathbb{E}[\hat{\theta}_n]}{\partial \theta})^\top$

where $A \succeq B$ is understood as $A - B$ is positive semi-definite. We introduced $I_n(\theta)$, the Fisher information matrix given by

$$I_n(\theta) = \mathbb{E}[\frac{\partial \log(\nu_\theta(\mathcal{X}))}{\partial \theta}\frac{\partial \log(\nu_\theta(\mathcal{X}))}{\partial \theta}^\top] \tag{2.11}$$

Lastly, the quantity $l(\mathbf{x}, \theta) = \log(\nu_\theta(\mathbf{x}))$ is called *log-likelihood of* $\mathbf{x}$ and its derivative $\boldsymbol{\psi}(\mathbf{x}, \theta) = \frac{\partial \log(\nu_\theta(\mathbf{x}))}{\partial \theta}$ is called the *score function of* $\mathbf{x}$.

Proof of Cramer-Rao bound. First let us show that when the optimal parameter is used, the expected score function cancels:

$$
\begin{aligned}
\mathbb{E}_{\mathcal{X}}[\boldsymbol{\psi}(\theta_*)] &= \mathbb{E}_{\mathcal{X}}[\frac{\partial \log(\nu_{\theta_*}(\mathcal{X}))}{\partial \theta}] && (2.12)\\
&= \mathbb{E}_{\mathcal{X}}[\frac{1}{\nu_{\theta_*}(\mathcal{X})}\frac{\partial \nu_{\theta_*}(\mathcal{X})}{\partial \theta}] && (2.13)\\
&= \int_{\mathcal{X}} \frac{1}{\nu_{\theta_*}(\mathcal{X})}\frac{\partial \nu_{\theta_*}(\mathcal{X})}{\partial \theta}\nu_{\theta_*}(\mathcal{X})d\mathcal{X} && (2.14)\\
&= \int_{\mathcal{X}} \frac{\partial \nu_{\theta_*}(\mathcal{X})}{\partial \theta}d\mathcal{X} && (2.15)\\
&= \frac{\partial \int_{\mathcal{X}} \nu_{\theta_*}(\mathcal{X})d\mathcal{X}}{\partial \theta} && (2.16)\\
&= \frac{\partial 1}{\partial \theta} && (2.17)\\
&= 0 && (2.18)
\end{aligned}
$$

so that $\mathbb{V}[\boldsymbol{\psi}(\theta_*)] = I_n(\theta_*)$.

We also have that:

$$
\begin{aligned}
\operatorname{Cov}(\hat{\theta}_n, \boldsymbol{\psi}(\theta_*)) &= \mathbb{E}[(\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n])(\boldsymbol{\psi}(\theta_*) - \mathbb{E}[\boldsymbol{\psi}(\theta_*)])^\top] && (2.19)\\
&= \mathbb{E}[\hat{\theta}_n \boldsymbol{\psi}(\theta_*)^\top] && (2.20)\\
&= \int_{\mathcal{X}} \hat{\theta}_n \frac{\partial \nu_{\theta_*}(\mathcal{X})}{\partial \theta}^\top d\mathcal{X} && (2.21)\\
&= \frac{\partial \int_{\mathcal{X}} \hat{\theta}_n \nu_{\theta_*}(\mathcal{X})d\mathcal{X}}{\partial \theta} && (2.22)\\
&= \frac{\partial \mathbb{E}[\hat{\theta}_n]}{\partial \theta} && (2.23)
\end{aligned}
$$

and similarly $\operatorname{Cov}(\boldsymbol{\psi}(\theta_*), \hat{\theta}_n) = (\frac{\partial \mathbb{E}[\hat{\theta}_n]}{\partial \theta})^\top$.

Then we apply the following Cauchy Schwartz inequality for random vectors:

$$\forall \mathbf{x}, \mathbf{y} \quad \operatorname{Var}(\mathbf{y}) \succeq \operatorname{Cov}(\mathbf{y}, \mathbf{x}) \operatorname{Var}(\mathbf{x})^{-1} \operatorname{Cov}(\mathbf{x}, \mathbf{y}) \qquad (2.24)$$

(see a proof in [148]) and therefore get the expected result:

$$\operatorname{Var}(\hat{\theta}_n) \succeq \frac{\partial \mathbb{E}[\hat{\theta}_n]}{\partial \theta} I_n(\theta_*)^{-1} (\frac{\partial \mathbb{E}[\hat{\theta}_n]}{\partial \theta})^\top \qquad (2.25)$$

□

When an estimator is both unbiased and reaches the Cramer-Rao bound we call the estimator *efficient*. If we assume that samples are independent and identically distributed we have: $\psi(\mathcal{X}, \theta) = \sum_{i=1}^n \psi(\mathbf{x}^{(i)}, \theta)$ and therefore

$$
\begin{aligned}
I_n(\theta_*) &= \mathbb{E}[\psi(\mathcal{X}, \theta_*)\psi(\mathcal{X}, \theta_*)^\top] && (2.26)\\
&= \sum_{i=1}^n \mathbb{E}[\psi(\mathbf{x}^{(i)}, \theta_*)\psi(\mathbf{x}^{(i)}, \theta_*)^\top] && (2.27)\\
&= nI(\theta_*) && (2.28)
\end{aligned}
$$

where

$$I(\theta_*) = \mathbb{E}_{\mathbf{x}\sim\nu_{\theta_*}}[\psi(\mathbf{x},\theta_*)\psi(\mathbf{x},\theta_*)^\top] \tag{2.29}$$

When samples are independent and identically distributed and $\hat{\theta}_n$ is unbiased the Cramer-Rao bound is given by:

$$\mathrm{Var}(\hat{\theta}_n) \succeq \frac{1}{n} I(\theta_*)^{-1} \tag{2.30}$$

The Example 2 hows that the sample mean is efficient.

Example 2 (Sample mean is efficient)**.** *Consider n observations $x^{(1)}, \dots x^{(n)}$ of x generated from $\nu_{\mu_*}(x) = \mathcal{N}(x; \mu_*, \sigma_*^2)$ and consider the sample mean $\hat{\mu}_e = \frac{1}{n}\sum_{i=1}^n x^{(i)}$. The variance is given by:*

$$\mathbb{V}[\hat{\mu}_e] = \mathbb{V}[\frac{1}{n}\sum_{i=1}^n x^{(i)}] \tag{2.31}$$

$$= \frac{\sigma_*^2}{n} \tag{2.32}$$

The Fisher information matrix is given by

$$I = \mathbb{E}[(\frac{\partial \log(\nu_{\mu_*}(x))}{\partial \mu})^2] \tag{2.33}$$

$$= \mathbb{E}[(\frac{\partial - \frac{1}{2\sigma_*^2}(x-\mu_*)^2 - \frac{1}{2}\log(2\pi\sigma_*^2)}{\partial \mu})^2] \tag{2.34}$$

$$= \mathbb{E}[(-\frac{1}{\sigma_*^2}(\mu_* - x))^2] \tag{2.35}$$

$$= \frac{1}{\sigma_*^4}\mathbb{E}[(\mu_* - x)^2] \tag{2.36}$$

$$= \frac{1}{\sigma_*^2} \tag{2.37}$$

and therefore we have

$$\mathbb{V}[\hat{\mu}_e] = \frac{1}{n} I^{-1} \tag{2.38}$$

so the sample mean is efficient.

In practice, efficient estimators are extremely rare. In the next section, we introduce the maximum likelihood estimator which is not always unbiased nor efficient in the finite sample case, but satisfies these properties asymptotically.

2.1.2 *Maximum likelihood*

The maximum likelihood estimates the parameters $\hat{\theta}$ such that the density $\nu_{\hat{\theta}}$ at $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}$ is the highest among all possible values for $\theta \in \Theta$:

$$\hat{\theta}_n = \underset{\theta\in\Theta}{\mathrm{argmax}}\, \nu_\theta(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}) \tag{2.39}$$

The quantity $\nu_\theta(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)})$ is called the *likelihood* and we call $\hat{\theta}_n$ the *maximum likelihood estimator*.

It is often assumed that samples are independent and identically distributed so that the joint density can be written: $\nu_\theta(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}) = \prod_{i=1}^n \nu_\theta(\mathbf{x}^{(i)})$

Let us define the *empirical expected log-likelihood*:

$$l_n(\theta) = \frac{1}{n}\sum_{i=1}^n l(\mathbf{x}^{(i)}, \theta) \tag{2.40}$$

$$= \mathbb{E}_n l(\mathbf{x}, \theta) \tag{2.41}$$

where $\mathbb{E}_n$ is the empirical expectation operator defined by

$$\mathbb{E}_n f(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^n f(\mathbf{x}^{(i)}) \tag{2.42}$$

The next lines show that finding the maximum likelihood is done by optimizing the empirical expected log-likelihood.

$$\hat{\theta}_n = \operatorname*{argmax}_{\theta\in\Theta} \nu_\theta(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}) \tag{2.43}$$

$$= \operatorname*{argmax}_{\theta\in\Theta} \prod_{i=1}^n \nu_\theta(\mathbf{x}^{(i)}) \tag{2.44}$$

$$= \operatorname*{argmax}_{\theta\in\Theta} \sum_{i=1}^n \log(\nu_\theta(\mathbf{x}^{(i)})) \tag{2.45}$$

$$= \operatorname*{argmax}_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^n l(\mathbf{x}^{(i)}, \theta) \tag{2.46}$$

$$= \operatorname*{argmax}_{\theta\in\Theta} l_n(\theta) \tag{2.47}$$

$$\tag{2.48}$$

By the law of large numbers, as the number of samples increases, the empirical expected likelihood converges almost surely to the *expected log-likelihood*: $l(\theta) = \mathbb{E}_{\mathbf{x}}[l(\mathbf{x}, \theta)]$.

Example 3 shows that, under a Gaussian model, the maximum likelihood estimator of the mean and the variance is the sample mean and sample variance respectively.

Example 3. *Consider* n *observations* $x^{(1)}, \dots, x^{(n)}$ *of an unknown random variable* x *and consider the model* $x \sim \mathcal{N}(\mu, \sigma^2)$. *The empirical expected log-likelihood is:*

$$l_n(\mu, \sigma^2) = -\frac{1}{n}\sum_{i=1}^n \frac{1}{2\sigma^2}(x^{(i)} - \mu)^2 - \frac{1}{2}\log(2\pi\sigma^2) \tag{2.49}$$

First order conditions yield:

$$\sum_{i=1}^{n}(\mu - x^{(i)}) = 0 \iff \mu = \frac{1}{n}\sum_{i=1}^{n} x^{(i)} \tag{2.50}$$

$$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{(\sigma^2)^2}(x^{(i)}-\mu)^2 - \frac{1}{\sigma^2} = 0 \iff \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x^{(i)}-\mu)^2 \tag{2.51}$$

So the maximum likelihood estimators of the mean and the variance are the sample mean and sample variance respectively.

The maximum likelihood estimator $\hat{\theta}_n$ is a random variable as it depends on $\mathcal{X} = \mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}$. As the number of samples n increases, we expect $\hat{\theta}_n$ to get closer to the optimal set of parameters θ_*. In Proposition 2, it is shown that the maximum likelihood estimator is *consistent* meaning $\hat{\theta}_n \xrightarrow[n\to\infty]{P} \theta_*$ where $\hat{\theta}_n \xrightarrow[n\to\infty]{P} \theta_*$ denotes the convergence in probability defined by $\forall \epsilon > 0,\ p(\|\hat{\theta}_n - \theta_*\| < \epsilon) \xrightarrow[n\to\infty]{} 1$.

Proposition 2 (Consistency of the maximum likelihood estimator)**.** *Assume Θ is compact, Assume l_n converge uniformly in probability to l and assume l is continuous. Lastly, assume that the model is identifiable. Then, the maximum likelihood estimator $\hat{\theta}_n$ is consistent meaning that $\hat{\theta}_n \xrightarrow[n\to\infty]{P} \theta_*$.*

where uniform convergence in probability means $\sup_\theta \|l_n(\theta) - l(\theta)\| \xrightarrow[n\to\infty]{P} 0$.

Proof. We first show that l is maximum at θ_*:

$$l(\theta_*) - l(\theta) = \mathbb{E}_{\mathbf{x}\sim\nu_{\theta_*}}[\log \nu_{\theta_*}(\mathbf{x}) - \log(\nu_\theta(\mathbf{x}))] \tag{2.52}$$

$$= D_{KL}(\nu_{\theta_*}, \nu_\theta) \tag{2.53}$$

$$\geqslant 0 \tag{2.54}$$

where D_{KL} is the Kullback-Leibler divergence that is always positive. The maximum is unique. This comes from the identifiability of the model that implies

$$l(\theta_*) = l(\theta) \implies \theta_* = \theta \tag{2.55}$$

Let $\varepsilon > 0$ and define $V_\varepsilon = \{\theta, \|\theta - \theta_*\| < \varepsilon\}$ an open neighborhood of θ_*. Because Θ is compact and V_ε open, $\Theta \cap V_\varepsilon^C$ is compact and since l is continuous, $\max_{\theta \in \Theta \cap V_\varepsilon^C} l(\theta)$ is reached for a value $\theta_0 \in \Theta \cap V_\varepsilon^C$.

Let us define $\delta = l(\theta_*) - l(\theta_0) > 0$ and consider the events

$$A_n = \sup_{\theta \in \Theta \cap V_\varepsilon^C} \|l_n(\theta) - l(\theta)\| < \frac{\delta}{2} \tag{2.56}$$

$$B_n = \sup_{\theta \in V_\varepsilon} \|l_n(\theta) - l(\theta)\| < \frac{\delta}{2} \tag{2.57}$$

We have

$$
\begin{aligned}
A_n &\implies \forall \theta \in \Theta \cap V_\varepsilon^C,\ l_n(\theta) - l(\theta) < \frac{\delta}{2} && (2.58)\\
&\implies \forall \theta \in \Theta \cap V_\varepsilon^C,\ l_n(\theta) < \frac{\delta}{2} + l(\theta_0) && (2.59)\\
&\implies \forall \theta \in \Theta \cap V_\varepsilon^C,\ l_n(\theta) < -\frac{\delta}{2} + l(\theta_*) && (2.60)
\end{aligned}
$$

and

$$
\begin{aligned}
B_n &\implies \forall \theta \in V_\varepsilon,\ l_n(\theta) > -\frac{\delta}{2} + l(\theta) && (2.61)\\
&\implies l_n(\theta_*) > -\frac{\delta}{2} + l(\theta_*) && (2.62)\\
& && (2.63)
\end{aligned}
$$

Then consider $\hat{\theta}_n = \text{argmax}_{\theta \in \Theta}(l_n(\theta))$ and assume $\hat{\theta}_n \in V_\varepsilon^C \cap \Theta$. We have

$$
A_n \cap B_n \implies l_n(\hat{\theta}_n) < -\frac{\delta}{2} + l(\theta_*) < l_n(\theta_*) \tag{2.64}
$$

So if $\hat{\theta}_n \in V_\varepsilon^C \cap \Theta$, $A_n \cap B_n$ contradicts the fact that $\hat{\theta}_n = \text{argmax}_{\theta \in \Theta}(l_n(\theta))$.

Therefore $A_n \cap B_n \implies \hat{\theta}_n \in V_\varepsilon$. From the uniform convergence in probability of l_n to l we have $p(A_n \cap B_n) \xrightarrow[n \to \infty]{} 1$ and therefore $p(\hat{\theta}_n \in V_\varepsilon) \xrightarrow[n \to \infty]{} 1$ which means $\hat{\theta}_n \xrightarrow[n \to \infty]{P} \theta_*$. □

As we have ween in Example 1, the sample variance is biased. However, in the large sample limit the bias disappears. Is this a consequence of consistency ? Intuitively, consistency seems to be a stronger condition than asymptotic unbiasedness since consistency implies convergence of the random variable $\hat{\theta}_n$ whereas asymptotic unbiasedness only implies convergence of the mean $\mathbb{E}[\hat{\theta}_n]$. Proposition 3 shows that this intuition is correct as long as the variance is bounded.

Proposition 3 (Consistency implies asymptotic unbiasedness)**.** *Assume* $\theta_n \xrightarrow[n \to \infty]{P} \theta_*$ *and assume* $\exists M \in \mathbb{R},\ \mathbb{E}[\|\theta_n - \theta_*\|^2] < M$.
Then $\mathbb{E}[\|\theta_n - \theta_*\|] \xrightarrow[n \to \infty]{} 0$

Proof. Set $\varepsilon > 0$, we have

$$
\begin{aligned}
\theta_n \xrightarrow[n \to \infty]{P} \theta_* &\implies p(\|\theta_n - \theta_*\| > \frac{\varepsilon}{2}) \xrightarrow[n \to \infty]{} 0 && (2.65)\\
&\implies \exists N \in \mathbb{N}, \forall n > N,\ p(\|\theta_n - \theta_*\| > \frac{\varepsilon}{2}) < \frac{\varepsilon}{M} && (2.66)
\end{aligned}
$$

So $\forall n > N$ we get:

$$
\begin{aligned}
\mathbb{E}[\|\theta_n - \theta_*\|] &= \mathbb{E}[\|\theta_n - \theta_*\| \mathbf{1}_{\|\theta_n - \theta_*\| \leqslant \frac{\varepsilon}{2}}] + \mathbb{E}[\|\theta_n - \theta_*\| \mathbf{1}_{\|\theta_n - \theta_*\| > \frac{\varepsilon}{2}}] && (2.67)\\
&\leqslant \frac{\varepsilon}{2} + \mathbb{E}[\|\theta_n - \theta_*\|^2] p(\|\theta_n - \theta_*\| > \frac{\varepsilon}{2}) && (2.68)\\
&< \varepsilon && (2.69)
\end{aligned}
$$

Where equation (2.68) follows from Cauchy-Schwarz, for the scalar product $a, b \to \mathbb{E}[ab^\top]$. □

While Proposition 2 states that $\hat{\theta}_n \xrightarrow[n\to\infty]{P} \theta_*$, Proposition 4 goes further in the analysis. It states that $\sqrt{n}\|\hat{\theta}_n - \theta_*\|$ approaches a Gaussian density as n gets large. This property is called *asymptotic normality*.

Proposition 4 (Asymptotic normality of maximum likelihood estimators)**.** *We assume the same hypothesis as in Proposition 2. We further assume that θ_* is in the interior of Θ. Lastly, denoting $J_n(\theta_n) = \frac{\partial^2 l_n}{\partial\theta^2}(\theta_n)$ and $J(\theta_*) = \frac{\partial^2 l}{\partial\theta^2}(\theta_*)$ we assume that*
$\theta_n \xrightarrow[n\to\infty]{P} \theta_* \implies J_n(\theta_n) \xrightarrow[n\to\infty]{P} J(\theta_*)$
Then,
$\sqrt{n}\|\hat{\theta}_n - \theta_*\| \xrightarrow[n\to\infty]{D} \mathcal{N}(0, I(\theta_*)^{-1})$

where the convergence in distribution $x^{(n)} \xrightarrow[n\to\infty]{D} x$ means $\forall t,\ F_{x^{(n)}}(t) \xrightarrow[n\to\infty]{} F_x(t)$ where F_x is the distribution function of x.

Proof. First order conditions give $\frac{\partial l_n(\hat{\theta}_n)}{\partial\theta} = 0$.
From the mean value theorem there exists θ_0 such that

$$\frac{\partial l_n(\hat{\theta}_n)}{\partial\theta} - \frac{\partial l_n(\theta_*)}{\partial\theta} = J_n(\theta_0)(\hat{\theta}_n - \theta_*) \tag{2.70}$$

$$\iff -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_{\theta_*}(\mathbf{x}^{(i)}) = J_n(\theta_0)\sqrt{n}(\hat{\theta}_n - \theta_*) \tag{2.71}$$

We know from the proof of the Cramer-Rao bound (Proposition 1) that $\mathbb{E}_{\mathbf{x}}[\psi_{\theta_*}(\mathbf{x})] = 0$. We can use the central limit theorem and write:

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_{\theta_*}(\mathbf{x}^{(i)}) \xrightarrow{D} \mathcal{N}(0, I(\theta_*)) \tag{2.72}$$

Then, we have:

$$J(\theta_*) = \mathbb{E}_{\mathbf{x}}[\frac{\partial\psi_{\theta_*}(\mathbf{x})}{\partial\theta}] \tag{2.73}$$

$$= \mathbb{E}_{\mathbf{x}}[\frac{\partial \frac{1}{\nu_{\theta_*}(\mathbf{x})}\frac{\partial\nu_{\theta_*}(\mathbf{x})}{\partial\theta}}{\partial\theta}] \tag{2.74}$$

$$= \mathbb{E}_{\mathbf{x}}[\frac{1}{\nu_{\theta_*}(\mathbf{x})}\frac{\partial^2\nu_{\theta_*}(\mathbf{x})}{\partial\theta^2} - \frac{\frac{\partial\nu_{\theta_*}(\mathbf{x})}{\partial\theta}}{\nu_{\theta_*}(\mathbf{x})}\frac{\frac{\partial\nu_{\theta_*}(\mathbf{x})}{\partial\theta}}{\nu_{\theta_*}(\mathbf{x})}^\top] \tag{2.75}$$

$$= -I(\theta_*) \tag{2.76}$$

$$\tag{2.77}$$

Then since $\hat{\theta}_n \xrightarrow[n\to\infty]{P} \theta_*$, $\theta_0 \xrightarrow[n\to\infty]{P} \theta_*$ and from our assumption about J_n we get $J_n(\theta_0) \xrightarrow[n\to\infty]{P} J(\theta_*) = -I(\theta_*)$.

From Slutsky's theorem we have

$$I(\theta_*)\sqrt{n}(\hat{\theta}_n - \theta_*) \xrightarrow{D} \mathcal{N}(0, I(\theta_*)) \tag{2.78}$$

$$\iff \sqrt{n}(\hat{\theta}_n - \theta_*) \xrightarrow{D} \mathcal{N}(0, I(\theta_*)^{-1}) \tag{2.79}$$

□

Looking at the variance of $\hat{\theta}_n - \theta_*$, we see that it behaves like $\frac{1}{n}I(\theta_*)^{-1}$ which is the quantity in the Cramer-Rao bound for unbiased estimators (see Proposition 1). Because of this, the maximum likelihood estimator is called *asymptotically efficient*.

2.1.3 *Some shortcomings*

It may look like an efficient and unbiased estimator should always yield the best possible mean squared error. However, this is not the case. A striking example is given by the Stein paradox [140]. As can be seen in Example 4, there exists an estimator of the mean that achieves a strictly better mean squared error than the sample mean. However, we have shown that the sample mean (which is also the maximum likelihood estimator) is unbiased and efficient. Stein's estimate shows that biased estimators can sometimes achieve lower mean squared error than unbiased ones.

Example 4 (Stein's estimate of the mean)**.** *Let us consider a single observation $\mathbf{x}$ generated from a multivariate Gaussian of dimension $\nu \geqslant 3$: $\mathcal{N}(\boldsymbol{\mu}, I)$ where I is the identity matrix and $\boldsymbol{\mu} \in \mathbb{R}^\nu$ is the unknown mean that we want to estimate. As in the one-dimensional case, the sample mean $e(\mathbf{x}) = \mathbf{x}$ is unbiased and efficient. The mean squared error is given by:*

$$\mathbb{E}[(e(\mathbf{x}) - \boldsymbol{\mu})^2] = \mathrm{tr}(V(e(\mathbf{x}))) = \nu \tag{2.80}$$

Now consider the estimator $s(\mathbf{x}) = \mathbf{x} - (\nu - 2)\frac{\mathbf{x}}{\|\mathbf{x}\|^2}$. The mean squared error is given by:

$$\mathbb{E}[(s(\mathbf{x}) - \boldsymbol{\mu})^2] \tag{2.81}$$

$$= \mathbb{E}[(\mathbf{x} - (\nu - 2)\frac{\mathbf{x}}{\|\mathbf{x}\|^2} - \boldsymbol{\mu})^2] \tag{2.82}$$

$$= \mathbb{E}[(\mathbf{x} - \boldsymbol{\mu})^2] - 2(\nu - 2)\mathbb{E}[\frac{(\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{x}}{\|\mathbf{x}\|^2}] + (\nu - 2)^2\mathbb{E}[\frac{1}{\|\mathbf{x}\|^2}] \tag{2.83}$$

$$= \mathbb{E}[(\mathbf{x} - \boldsymbol{\mu})^2] - 2(\nu - 2)\sum_{j=1}^{\nu}\mathbb{E}[\frac{(x_j - \mu_j)x_j}{\|\mathbf{x}\|^2}] + (\nu - 2)^2\mathbb{E}[\frac{1}{\|\mathbf{x}\|^2}] \tag{2.84}$$

where x_j is the j-th coordinate of $\mathbf{x}$

By integration by part

$$\mathbb{E}[\frac{(x_j-\mu_j)x_j}{\|\mathbf{x}\|^2}] = \mathbb{E}[\frac{\partial \frac{x_j}{\|\mathbf{x}\|^2}}{\partial x_j}] \tag{2.85}$$

$$= \mathbb{E}[\frac{1}{\|\mathbf{x}\|^2} - \frac{2x_j^2}{\|\mathbf{x}\|^4}] \tag{2.86}$$

so that

$$\mathbb{E}[(s(\mathbf{x})-\mu)^2] \tag{2.87}$$

$$= \mathbb{E}[(\mathbf{x}-\mu)^2] - 2(\nu-2)\sum_{j=1}^{\nu}\mathbb{E}[\frac{1}{\|\mathbf{x}\|^2} - \frac{2x_j^2}{\|\mathbf{x}\|^4}] + (\nu-2)^2\mathbb{E}[\frac{1}{\|\mathbf{x}\|^2}] \tag{2.88}$$

$$= \mathbb{E}[(\mathbf{x}-\mu)^2] - 2(\nu-2)\mathbb{E}[\frac{\nu}{\|\mathbf{x}\|^2} - \frac{2\sum_{j=1}^{\nu} x_j^2}{\|\mathbf{x}\|^4}] + (\nu-2)^2\mathbb{E}[\frac{1}{\|\mathbf{x}\|^2}] \tag{2.89}$$

$$= \mathbb{E}[(\mathbf{x}-\mu)^2] - 2(\nu-2)^2\mathbb{E}[\frac{1}{\|\mathbf{x}\|^2}] + (\nu-2)^2\mathbb{E}[\frac{1}{\|\mathbf{x}\|^2}] \tag{2.90}$$

$$= \nu - (\nu-2)^2\mathbb{E}[\frac{1}{\|\mathbf{x}\|^2}] \tag{2.91}$$

Therefore $s(\mathbf{x})$ *always yields lower expected mean squared error than* $e(\mathbf{x})$*.*

Example 4 shows that shrinking the sample mean towards the origin decreases the mean squared error of the estimate in high dimensions. A similar technique can also be used for estimating the covariance of variables. For instance, Ledoit and Wolf showed in [84] that shrinking the sample covariance towards identity decreases lower expected mean squared error in high dimensions.

2.2 OPTIMIZATION

The maximum likelihood estimator gives a natural way to obtain estimators that are consistent and asymptotically efficient. However, it requires finding the maximum of the empirical expected log-likelihood. This can rarely be done using a closed form formula and one almost always have to resort to iterative methods.

2.2.1 *Some iterative optimization algorithms*

Let us consider a function $f : \mathbb{R}^\nu \to \mathbb{R}$ that we want to minimize. $f(\theta)$, where $\theta \in \mathbb{R}^\nu$, can be seen as the negative expected likelihood.

Optimization algorithms that only use first order derivatives to make a step are called *first order methods*. We will begin by presenting the famous *gradient descent*. Then, we move on to second order methods with *Newton* and *quasi-Newton methods*. This section closely follows the chapter 9 of [24].

2.2.1.1 *Gradient descent*

We assume that f is differentiable everywhere. From a point θ_0, assuming a small *step size* α and a direction $\mathbf{d}$ such that $\|\mathbf{d}\| = 1$, a Taylor decomposition at first order gives:

$$f(\theta_0 + \alpha\mathbf{d}) = f(\theta_0) + \langle\frac{\partial f(\theta_0)}{\partial\theta}, \alpha\mathbf{d}\rangle + o(\alpha) \tag{2.92}$$

The best direction is the one that minimizes $f(\theta_0 + \alpha\mathbf{d})$. If we neglect the terms in $o(\alpha)$, it is given by

$$\underset{\mathbf{d}, \|\mathbf{d}\|=1}{\operatorname{argmin}} f(\theta_0) + \langle\frac{\partial f(\theta_0)}{\partial\theta}, \alpha\mathbf{d}\rangle = \frac{-\frac{\partial f(\theta_0)}{\partial\theta}}{\|\frac{\partial f(\theta_0)}{\partial\theta}\|} \tag{2.93}$$

Therefore, gradient descent updates are given by:

$$\theta_{k+1} = \theta_k - \alpha\frac{\partial f(\theta_k)}{\partial\theta} \tag{2.94}$$

where α is a small quantity.

Under certain conditions that we specify in Proposition 5, gradient descent converges to the minimum.

Proposition 5 (Convergence of gradient descent). *Assume that f is twice differentiable and μ-strongly convex:*

$$\frac{\partial^2 f(\theta)}{\partial\theta^2} \succeq \mu I \tag{2.95}$$

In addition, assume that f is ℓ-smooth:

$$\frac{\partial^2 f(\theta)}{\partial\theta^2} \preceq \ell I \tag{2.96}$$

where I is the identity matrix. From $\theta_0 \in \mathbb{R}^\nu$ and given $\alpha \in \mathbb{R}$ such that $\frac{1}{\ell} \geqslant \alpha > 0$, the iterates $\theta_{k+1} = \theta_k - \alpha\frac{\partial f(\theta_k)}{\partial\theta}$ converge to the minimum θ_* according to

$$\|\theta^{k+1} - \theta_*\|^2 \leqslant (1 - \alpha\mu)^{k+1}\|\theta_0 - \theta_*\|^2 \tag{2.97}$$

Proof. We have

$$\|\theta^{k+1} - \theta_*\|^2 = \|\theta_k - \alpha\frac{\partial f(\theta_k)}{\partial\theta} - \theta_*\|^2 \tag{2.98}$$

$$= \|\theta_k - \theta_*\|^2 - 2\langle\theta_k - \theta_*, \alpha\frac{\partial f(\theta_k)}{\partial\theta}\rangle + \|\alpha\frac{\partial f(\theta_k)}{\partial\theta}\|^2 \tag{2.99}$$

From μ-strong convexity and Lagrange inequality we get:

$$f(\theta_*) \geqslant f(\theta_k) + \langle\frac{\partial f(\theta_k)}{\partial\theta_k}, \theta_* - \theta_k\rangle + \frac{\mu}{2}\|\theta_* - \theta_k\|^2 \tag{2.100}$$

so that

$$-2\alpha\langle\frac{\partial f(\theta_k)}{\partial\theta_k},\theta_k-\theta_*\rangle \leqslant -2\alpha(f(\theta_k)-f(\theta_*))-\mu\alpha\|\theta_*-\theta_k\|^2 \tag{2.101}$$

and using the fact that f is ℓ-smooth we have:

$$\forall\theta,y,\ \ f(y)\leqslant f(\theta)+\frac{\partial f(\theta)}{\partial\theta}(y-\theta)+\frac{\ell}{2}\|y-\theta\|^2 \tag{2.102}$$

$$\Longrightarrow \forall\theta,\ \ f(\theta-\frac{1}{\ell}\frac{\partial f(\theta)}{\partial\theta})-f(\theta)\leqslant-\frac{1}{2\ell}\|\frac{\partial f(\theta)}{\partial\theta}\|^2 \tag{2.103}$$

$$\Longrightarrow \forall\theta,\ \ f(\theta_*)-f(\theta)\leqslant-\frac{1}{2\ell}\|\frac{\partial f(\theta)}{\partial\theta}\|^2 \tag{2.104}$$

$$\Longrightarrow \|\alpha\frac{\partial f(\theta_k)}{\partial\theta}\|^2\leqslant 2\alpha^2 l(f(\theta_k)-f(\theta_*)) \tag{2.105}$$

So from (2.99) using the inequalities (2.101) and (2.105) we get

$$\|\theta^{k+1}-\theta_*\|^2\leqslant(1-\alpha\mu)\|\theta_k-\theta_*\|^2-2\alpha(1-\alpha\ell)(f(\theta_k)-f(\theta_*)) \tag{2.106}$$

and since $0<\alpha<\frac{1}{\ell}$ we get that

$$\|\theta^{k+1}-\theta_*\|^2\leqslant(1-\alpha\mu)\|\theta_k-\theta_*\|^2 \tag{2.107}$$

which by induction yields the desired result. □

The type of convergence we get in Proposition 5 is called a *linear convergence* (because $\|\theta^{k+1}-\theta_*\|$ is bounded by a linear function of $\|\theta^k-\theta_*\|$) .

2.2.1.2 *Newton method and quasi-Newton methods*

We assume that f is twice differentiable everywhere. As in gradient descent, we depart from a point θ_0, assume a small *step size* α and consider a direction $\mathbf{d}$ such that $\|\mathbf{d}\|=1$. A Taylor decomposition at the second order gives:

$$f(\theta_0+\alpha\mathbf{d})=f(\theta_0)+\langle\frac{\partial f(\theta_0)}{\partial\theta},\alpha\mathbf{d}\rangle+\langle\alpha\mathbf{d},\frac{\partial^2 f(\theta_0)}{\partial\theta^2}\alpha\mathbf{d}\rangle+o(\alpha^2) \tag{2.108}$$

If we neglect the terms in $o(\alpha^2)$, the best direction is given by

$$\underset{\mathbf{d},\|\mathbf{d}\|=1}{\operatorname{argmin}}\, f(\theta_0)+\langle\frac{\partial f(\theta_0)}{\partial\theta},\alpha\mathbf{d}\rangle+\langle\alpha\mathbf{d},\frac{\partial^2 f(\theta_0)}{\partial\theta^2}\alpha\mathbf{d}\rangle \tag{2.109}$$

$$=-\frac{(\frac{\partial^2 f(\theta_0)}{\partial\theta^2})^{-1}\frac{\partial f(\theta_0)}{\partial\theta}}{\|(\frac{\partial^2 f(\theta_0)}{\partial\theta^2})^{-1}\frac{\partial f(\theta_0)}{\partial\theta}\|} \tag{2.110}$$

Therefore, Newton updates are given by:

$$\theta_{k+1} = \theta_k - \alpha(\frac{\partial^2 f(\theta_k)}{\partial \theta^2})^{-1} \frac{\partial f(\theta_k)}{\partial \theta} \tag{2.111}$$

where α is a small quantity.

Unfortunately, Newton's method is not guaranteed to converge. This contrasts with the gradient descent method that is guaranteed to converge when the step-size is small enough. In order to converge, Newton's method needs that each step yields a sufficient decrease of the loss. This is done by a line search.

The *exact line search* sets $\alpha = \text{argmin}_{t \in [0,1]} f(\theta_k - t(\frac{\partial^2 f(\theta_0)}{\partial \theta^2})^{-1} \frac{\partial f(\theta_k)}{\partial \theta})$ while the *backtracking line search* is an iterative procedure where one starts with $\alpha = 1$ and repeatedly halves α until $f(\theta_{k+1}) < f(\theta_k)$.

As shown by Proposition 6, Newton's method is guaranteed to converge when an exact line search is used.

Proposition 6 (Convergence of Newton's method)**.** *Assume that* f *is twice differentiable,* μ*-strongly convex and* ℓ*-smooth:*

$$\mu I \preceq \frac{\partial^2 f(\theta)}{\partial \theta^2} \preceq \ell I \tag{2.112}$$

where I *is the identity matrix. Assume that the Hessian is Lipschitz with constant* h*:*

$$\|\frac{\partial^2 f(\theta)}{\partial \theta^2} - \frac{\partial^2 f(\theta')}{\partial \theta^2}\| \leqslant h\|\theta - \theta'\| \tag{2.113}$$

From $\theta_0 \in \mathbb{R}^v$ *the iterates* $\theta_{k+1} = \theta_k - \alpha(\frac{\partial^2 f(\theta_0)}{\partial \theta^2})^{-1} \frac{\partial f(\theta_k)}{\partial \theta}$ *where* α *is chosen using an exact line search converge to the minimum* θ_**. Depending on the norm of the gradient two phases exist:*

- *The damped phase where* $\|\frac{\partial f(\theta)}{\partial \theta}\| \geqslant \frac{\mu^2}{h}$*. In this phase we have* $f(\theta_{k+1}) - f(\theta_k) \leqslant \frac{\mu^5}{2h^2\ell^2}$*. Therefore the number of iterations in this phase* i *cannot be larger than* $\frac{2h^2\ell^2}{\mu^5}(f(\theta_0) - f(\theta_*))$*.*
- *The pure Newton phase where* $\|\frac{\partial f(\theta)}{\partial \theta}\| < \frac{\mu^2}{h}$*. In this phase* $\|f(\theta_k) - f(\theta_*)\| \leqslant \frac{2\mu^3}{h^2}(\frac{1}{2})^{2^{k-i+1}}$

Once the Newton phase is reached, it continues until convergence of the algorithm.

Proof. In the damped phase we have $\|\frac{\partial f(\theta)}{\partial \theta}\| \geqslant \frac{\mu^2}{h}$. From Lagrange inequality and ℓ-smoothness, denoting $\mathbf{d}_k = -(\frac{\partial^2 f(\theta)}{\partial \theta^2})^{-1} \frac{\partial f(\theta)}{\partial \theta}$ we have:

$$f(\theta_k + t\mathbf{d}_k) \leqslant f(\theta) + t\langle \frac{\partial f(\theta)}{\partial \theta}, \mathbf{d}_k \rangle + \frac{t^2 \ell}{2}\|\mathbf{d}_k\|^2 \tag{2.114}$$

$$\leqslant f(\theta) + t\langle \frac{\partial f(\theta)}{\partial \theta}, \mathbf{d}_k \rangle - \frac{t^2 \ell}{2\mu}\langle \frac{\partial f(\theta)}{\partial \theta}, \mathbf{d}_k \rangle \tag{2.115}$$

$$\tag{2.116}$$

We then have

$$f(\theta_k + \alpha \mathbf{d}_k) \leqslant f(\theta) + t\langle \frac{\partial f(\theta)}{\partial \theta}, \mathbf{d}_k \rangle - \frac{t^2 \ell}{2\mu} \langle \frac{\partial f(\theta)}{\partial \theta}, \mathbf{d}_k \rangle \tag{2.117}$$

$$\tag{2.118}$$

setting $t = \frac{\mu}{\ell}$ gives

$$f(\theta_k + \alpha \mathbf{d}_k) \leqslant f(\theta) + \frac{\mu}{2\ell} \langle \frac{\partial f(\theta)}{\partial \theta}, \mathbf{d}_k \rangle \tag{2.119}$$

$$\leqslant f(\theta) - \frac{\mu}{2\ell^2} \| \frac{\partial f(\theta)}{\partial \theta} \|^2 \tag{2.120}$$

$$\leqslant f(\theta) - \frac{\mu^5}{2h\ell^2} \tag{2.121}$$

$$\tag{2.122}$$

which gives the desired result.

In the pure Newton phase, we assume $\alpha = 1$ (the exact line search can only be better than this). We have

$$\frac{\partial f(\theta_{k+1})}{\partial \theta} = \frac{\partial f(\theta_k + \mathbf{d}_k)}{\partial \theta} - \frac{\partial f(\theta_k)}{\partial \theta} - \frac{\partial^2 f(\theta_k)}{\partial \theta^2} \mathbf{d}_k \tag{2.123}$$

$$= \int_{t \in [0,1]} \frac{\partial^2 f(\theta_k + t\mathbf{d}_k)}{\partial \theta^2} \mathbf{d}_k dt - \frac{\partial^2 f(\theta_k)}{\partial \theta^2} \mathbf{d}_k \tag{2.124}$$

$$= \int_{t \in [0,1]} (\frac{\partial^2 f(\theta_k + t\mathbf{d}_k)}{\partial \theta^2} - \frac{\partial^2 f(\theta_k)}{\partial \theta^2}) \mathbf{d}_k dt \tag{2.125}$$

$$\tag{2.126}$$

Then

$$\| \frac{\partial f(\theta_{k+1})}{\partial \theta} \| \leqslant \int_{t \in [0,1]} \| (\frac{\partial^2 f(\theta_k + t\mathbf{d}_k)}{\partial \theta^2} - \frac{\partial^2 f(\theta_k)}{\partial \theta^2}) \mathbf{d}_k \| dt \tag{2.127}$$

$$\leqslant \int_{t \in [0,1]} \| (\frac{\partial^2 f(\theta_k + t\mathbf{d}_k)}{\partial \theta^2} - \frac{\partial^2 f(\theta_k)}{\partial \theta^2}) \| \| \mathbf{d}_k \| dt \tag{2.128}$$

$$\leqslant \int_{t \in [0,1]} th \| \mathbf{d}_k \|^2 dt \tag{2.129}$$

$$\leqslant \frac{h}{2\mu^2} \| \frac{\partial f(\theta_k)}{\partial \theta} \|^2 \tag{2.130}$$

$$\tag{2.131}$$

Since $\| \frac{\partial f(\theta_k)}{\partial \theta} \| \leqslant \frac{\mu^2}{h}$ the sequence of gradients converges to zero. The convergence is quadratic. At iteration k we therefore have:

$$\frac{h}{2\mu^2} \| \frac{\partial f(\theta_k)}{\partial \theta} \| \leqslant (\frac{1}{2})^{2^{k-i}} \tag{2.132}$$

$$\tag{2.133}$$

Using strong convexity we conclude that

$$f(\theta_k) - f(\theta_*) \leqslant \frac{1}{2\mu} \| \frac{\partial f(\theta_k)}{\partial \theta} \|^2 \leqslant 2\frac{\mu^3}{h^2} (\frac{1}{2})^{2^{k-i}} \tag{2.134}$$

□

In the pure Newton phase, the convergence is quadratic. This is a very strong advantage of Newton methods. A second advantage of Newton method is its *equivariance* as demonstrated in Proposition 7. Equivariance means that for any invertible matrix A, working with parameters $A\theta$ instead of parameters θ has no impact on the result given by the algorithm.

Proposition 7 (Newton algorithms are equivariant)**.** *Let us consider* θ_k *the current estimate of the minimum of* f *obtained by the Newton method after* k *iterations starting from* θ_0*. For any invertible matrix* A*, consider* y_k *the current estimate of the minimum of* $g(y) = f(Ay)$ *obtained by the Newton method after* k *iterations starting from* $y_0 = A^{-1}\theta_0$*. Then,* $y_k = A^{-1}\theta_k$*.*

Proof. The relation holds for $k = 0$ and assuming $y_k = A^{-1}\theta_k$ we have:

$$y_{k+1} = y_k - \alpha(\frac{\partial^2 g(y_k)}{\partial y^2})^{-1} \frac{\partial g(y_k)}{\partial y} \tag{2.135}$$

$$= y_k - \alpha A^{-2}(\frac{\partial^2 f(\theta_k)}{\partial \theta^2})^{-1} A \frac{\partial f(\theta_k)}{\partial \theta} \tag{2.136}$$

$$= A^{-1}(\theta_k - \alpha(\frac{\partial^2 f(\theta_k)}{\partial \theta^2})^{-1} \frac{\partial f(\theta_k)}{\partial \theta}) \tag{2.137}$$

$$= A^{-1}\theta_{k+1} \tag{2.138}$$

$$\tag{2.139}$$

□

In contrast, gradient descent is not equivariant. However, despite its attractive properties, Newton's method is rarely used in practice because it is often intractable. Indeed inverting the Hessian $\frac{\partial^2 f}{\partial \theta^2}$ is difficult as it is an operator in dimension $\mathbb{R}^{\nu \times \nu \times \nu \times \nu}$. In some cases, it is possible to construct an approximation of the inverse of the Hessian in a reasonable time. This can be done iteratively by building a sequence of matrices B_k that approaches the inverse of the Hessian as k grows. Methods that use an approximation of the inverse instead of the true inverse are called *quasi-Newton* methods [56].

2.2.2 *EM and generalized EM*

In the maximum likelihood framework, our goal is to find the parameters θ_* that maximize the expected likelihood $l(\theta)$ of observations $\mathcal{X} = x^{(1)}, \dots, x^{(n)}$ assuming a model $\nu_\theta(\mathcal{X})$. In the previous section, we have presented some iterative optimization methods that can be used to directly maximize the log-likelihood. In this section, we present an alternative technique suited to *latent variable models*. Latent variable models include, in addition to observed data x, unobserved data z called *latent variables*. The observed data x is assumed to depend on the unobserved data z. With some abuse of notation we denote $\nu_\theta(z)$ the likelihood of z, $\nu_\theta(x|z)$ the density of $x|z$ evaluated at x, z and $\nu_\theta(x,z) = \nu_\theta(z)\nu_\theta(x|z)$ the joint likelihood of x, z called the *complete likelihood*.

We can relate the likelihood and the completed likelihood by:

$$\nu_\theta(x) = \int_z \nu_\theta(x,z)dz \tag{2.140}$$

The EM algorithm maximizes an expression that depends on the complete log-likelihood rather than the log-likelihood so it is useful when the former is much simpler to maximize than the later. This section follows the work in [106].

In order to make this more concrete, let us focus on the Gaussian mixture model in Example 5.

Example 5 (Gaussian mixture models). *Let us consider a latent variable* z *sampled from a Bernoulli distribution with parameter* ϕ*, and* x *is given by* $x|z = 0 \sim \mathcal{N}(\mu^0, 1)$ *and and* $x|z = 1 \sim \mathcal{N}(\mu^1, 1)$*.*

We call $x^{(1)}, \dots, x^{(n)}$ *the observed samples and* $z^{(1)}, \dots, z^{(n)}$ *the corresponding unobserved latent variables.*

The log-likelihood of $x^{(i)}$ *is given by:*

$$l(x^{(i)}, (\mu^0, \mu^1, \phi)) \tag{2.141}$$

$$= \log(\phi\mathcal{N}(x^{(i)}, \mu^1, 1) + (1-\phi)\mathcal{N}(x^{(i)}, \mu^0, 1)) \tag{2.142}$$

$$= \log\left(\phi\frac{\exp(-\frac{(x^{(i)}-\mu^1)^2}{2})}{\sqrt{2\pi}} + (1-\phi)\frac{\exp(-\frac{(x^{(i)}-\mu^0)^2}{2})}{\sqrt{2\pi}}\right) \tag{2.143}$$

Whereas the completed log-likelihood of $(x^{(i)}, z^{(i)})$ *is given by*

$$l((x^{(i)}, z^{(i)}), (\mu^0, \mu^1, \phi)) \tag{2.144}$$

$$= \log((\phi\mathcal{N}(x^{(i)}, \mu^1, 1))^{z^{(i)}}((1-\phi)\mathcal{N}(x^{(i)}, \mu^0, 1))^{1-z^{(i)}}) \tag{2.145}$$

$$= z^{(i)}(-\frac{(x^{(i)}-\mu^1)^2}{2} + \log(\phi)) \tag{2.146}$$

$$+ (1-z^{(i)})(-\frac{(x^{(i)}-\mu^0)^2}{2} + \log(1-\phi)) + c \tag{2.147}$$

where c *is a constant that does not depend on* μ^1 *or* μ_2*.*

It is easy to see that the expected completed log-likelihood will be much easier to maximize than the expected log-likelihood.

This example shows that unsurprisingly, it would be easier to find the parameters of a Gaussian mixture model if we knew the component from which are generated each sample.

Instead of optimizing the expected log-likelihood $l(\theta)$ directly, the EM algorithm optimizes the following function:

$$F(q,\theta) = \mathbb{E}_q[\log(\nu_\theta(x,z))] + H_q \tag{2.148}$$

where q is a density and

$$\mathbb{E}_q[\log(\nu_\theta(x,z))] = \int_z \log(\nu_\theta(x,z))q(z)dz \tag{2.149}$$

$$H_q = \int_z -\log(q(z))q(z)dz \tag{2.150}$$

H_q is called the entropy of q and is always positive. Let us introduce $\nu_\theta(z|x)$ the density of $z|x$, we have: $\nu_\theta(x,z) = \nu_\theta(z|x)\nu_\theta(x)$ and therefore

$$\begin{aligned} F(q,\theta) &= \mathbb{E}_q[\log(\nu_\theta(z|x)) + \log(\nu_\theta(x))] + \mathbb{E}_q[-\log(q)] && (2.151)\\ &= \log(\nu_\theta(x)) - \mathbb{E}_q[\log(q) - \log(\nu_\theta(z|x))] && (2.152)\\ &= l(x,\theta) - D_{KL}(q,\nu_\theta(z|x)) && (2.153)\end{aligned}$$

From the fact that D_{KL} is positive and $D_{KL}(a,b) = 0 \iff a = b$ we have that

$$F(q,\theta) \leqslant l(x,\theta) \tag{2.154}$$

$$F(\nu_\theta(z|x),\theta) = l(x,\theta) \tag{2.155}$$

and therefore

$$\max_{q,\theta} F(q,\theta) = \max_\theta(\max_q F(q,\theta)) = \max_\theta l(x,\theta) \tag{2.156}$$

Any EM algorithm maximizes l by maximizing F. The most common practice is to maximize alternatively F with respect to q and θ. At iteration k, let us call θ_k the current estimate of θ. According to equation (2.155), the maximum with respect to q is given by $q = \nu_{\theta_k}(z|x)$. Then, we have to maximize equation (2.148) with respect to θ. As the entropy H does not depend on θ the function to maximize is given by:

$$Q(\theta) = \mathbb{E}_{z\sim\nu_{\theta_k}(z|x)}[\log(\nu_\theta(x,z))] \tag{2.157}$$

Computing Q is called the *E-step*. Then we maximize Q and set

$$\theta_{k+1} = \operatorname*{argmax}_\theta Q(\theta) \tag{2.158}$$

This step is called the *M-step*.

In Example 6 we us use the EM algorithm to optimize the Gaussian mixture model introduced in Example 5.

Example 6 (Optimizing the Gaussian mixture via EM)**.** *Let us take the same data as in Example 5. We call μ_k^0, μ_k^1, ϕ_k the estimates of μ^0, μ^1, ϕ at iteration k and define:*

$$\gamma_k^{(i)} = p(z^{(i)} = 1 | x^{(i)}) = \frac{\phi_k \mathcal{N}(x^{(i)}; \mu_k^0, 1)}{\phi_k \mathcal{N}(x^{(i)}; \mu_k^0, 1) + (1 - \phi_k)\mathcal{N}(x^{(i)}; \mu_k^1, 1)} \tag{2.159}$$

The E-step is given by:

$$Q(\mu^0, \mu^1, \phi) = \sum_{i=1}^{n} \Big(\gamma_k^{(i)} \log(\nu_{\mu^0,\mu^1,\phi}(x^{(i)}, 1)) \tag{2.160}$$

$$+ (1 - \gamma_k^{(i)}) \log(\nu_{\mu^0,\mu^1,\phi}(x^{(i)}, 0))\Big) \tag{2.161}$$

$$= \sum_{i=1}^{n} \Big(\gamma_k^{(i)}(-\frac{(x^{(i)} - \mu^1)^2}{2} + \log(\phi)) \tag{2.162}$$

$$+ (1 - \gamma_k^{(i)})(-\frac{(x^{(i)} - \mu^0)^2}{2} + \log(1 - \phi)) + c\Big) \tag{2.163}$$

The M-step is given by:

$$\frac{\partial Q}{\partial \phi}(\phi^{k+1}) = 0 \iff \phi^{k+1} = \frac{\sum_{i=1}^{n} \gamma_k^{(i)}}{n} \tag{2.164}$$

$$\frac{\partial Q}{\partial \phi}(\mu_0^{k+1}) = 0 \iff \mu_0^{k+1} = \frac{\sum_{i=1}^{n} (1 - \gamma_k^{(i)}) x^{(i)}}{n} \tag{2.165}$$

$$\frac{\partial Q}{\partial \phi}(\mu_1^{k+1}) = 0 \iff \mu_1^{k+1} = \frac{\sum_{i=1}^{n} \gamma_k^{(i)} x^{(i)}}{n} \tag{2.166}$$

The maximization of F via the *E-step* and *M-step*, like in Example 6, is the historical version of the EM algorithm. However, other optimization techniques referred to as *EM variants* are possible. For instance, one could replace the maximization in equation (2.158) by just one step of an iterative optimization algorithm giving a set of parameters θ_{k+1} that verifies:

$$Q(\theta_{k+1}) > Q(\theta_k) \tag{2.167}$$

When the M-step is performed in such an approximated way, the EM algorithm is renamed *generalized EM* [44]. Similarly, one could replace the exact *E-step* by just improving the current density estimate with respect to the value function $q \to F(q, \theta)$ [67].

2.3 CONCLUSION

In this chapter, we have introduced the maximum-likelihood estimator and have shown that it is consistent and asymptotically efficient. This estimator is the solution of an optimization problem. We have presented gradient descent and Newton's method, two popular iterative optimization methods that can be used to maximize the likelihood directly. Lastly, we explored the special case of latent variable models, where the maximum likelihood estimator can be estimated via the EM algorithm.

In the next chapter, we introduce the necessary background on neuro-imaging.

3 NEUROSCIENCE BACKGROUND

In the previous chapter, we have introduced some background on statistics and optimization. In this chapter, we present the necessary background on the data used in this thesis. In our experiments, we use functional magnetic resonance imaging (fMRI) or magnetoencephalograpy (MEG) data. As will be seen in this chapter, these two modalities have very different characteristics.

3.1 FUNCTIONAL MAGNETIC RESONANCE IMAGING (FMRI)

This section relies on the handbook of fMRI data analysis [128].

3.1.1 *The BOLD signal and hemodynamic response*

The story of fMRI begins with the discovery of the blood-oxygen-level-dependent (BOLD) imaging contrast in 1990 by Ogawa [110]. When neurons fire, they consume oxygen and therefore, the level of oxygen in blood changes (it is actually over-compensated by blood supply). Since oxygenated and deoxygenated blood (oxyhemoglobin and deoxyhemoglobin) do not have the same magnetic susceptibility, changes in the oxygen level in blood can be tracked by a magnetic resonance imaging (MRI) scanner.

When a short stimuli occurs, the relative change in the MRI signal (BOLD signal) is not instantaneous. The typical response to a short stimuli follows the curve displayed in figure 3.1 and is called the *hemodynamic response*. The BOLD signal can be seen as the convolution of neural activation with the hemodynamic response.

3.1.2 *Spatial and temporal resolution of fMRI data*

The success of fMRI is due to its advantages compared to positron emission topography (PET) that uses radioactive tracers to follow glucose or water levels. PET and fMRI both have a spatial resolution of a few millimeters (which still encompasses several hundred thousands of neurons). However, while taking an image with PET takes about a minute, it takes on the order of 2s to produce with fMRI. In addition, fMRI is non-invasive and, unlike PET, it does not involve being exposed to radioactive tracers.

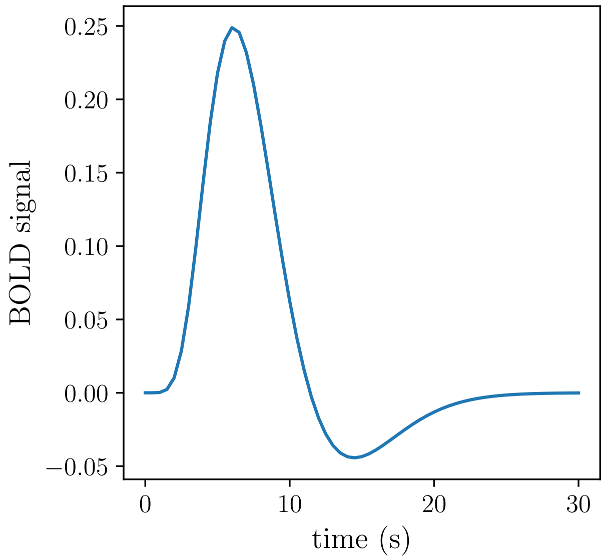


Figure 3.1: A model of the hemodynamic response function

3.1.3 *Experimental designs*

In *resting state* fMRI, subjects are instructed to lie still in the scanner without further instruction. This kind of data can be used to extract networks that highlight regions that tend to co-activate. This contrasts with *task* fMRI, in which subjects have to comply with specific instructions. In task fMRI, two types of experimental designs are widespread. *Block designs* are suited to functional imaging modalities with low temporal resolutions. In block designs, subjects are exposed to a continuous stimulus that lasts of a rather long period of time before it switches to another. This contrasts with *event related designs* that are only suited to modalities with a finer time resolution. In event related designs, a sequence of short events are presented and separated by a time window (inter-stimulus interval). When the inter-stimulus interval is shorter than the length of the hemodynamic response, we talk about "fast" even related designs. While block designs or event related designs are controlled designs, a rising trend is to use more naturalistic paradigms that are unconstrained from behavioral manipulations and thus, more ecological with respect to real-world conditions. In such naturalistic paradigms, referred to as *naturalistic task fMRI* in this thesis, subjects are exposed to *naturalistic stimuli* such as movie watching or audio track listening. These kinds of stimuli constitute a middle ground between resting state and task fMRI data as activations are time-locked but the environment is not as controlled as in a typical task fMRI setting. In this thesis, we use mostly *naturalistic task fMRI*, as one of our motivations is to propose a suitable framework to analyse such data.

3.1.4 *Preprocessing*

The fMRI signal is particularly noisy. Several techniques are used to reduce the noise and enhance the quality of the data.

3.1.4.1 *Distortion correction*

In the MRI scanner, a constant magnetic field is applied. However, in areas where there is an air/tissue interface, the magnetic field is distorted leading to errors in the localization of voxels near the sinus or ears. However, most scanners come with a map quantifying the distance that each voxel has been shifted. By inverting the map, one can try and recover the correct location of voxels.

3.1.4.2 *Slice timing correction*

Images are acquired in slices, a few slices at a time, so that different voxels are acquired at different times. In order to correct for this effect, a reference slice is chosen and other slices are interpolated so that we can consider that all voxels are acquired at the same time.

3.1.4.3 *Motion correction*

Head motion causes massive distortion of the signal. Such effects are corrected by chosing a reference image and applying a rigid body transformations so that other images match the orientation and localization of the reference image. One also records estimated head movements parameters so that they can be regressed out. However, if the task is correlated with head motion, this regression can lead to a loss of information.

3.1.4.4 *Spatial smoothing*

Spatial smoothing blurs the image through the application of a Gaussian kernel with a given width (in mm). When the signal of interest has a large spatial extend, smoothing increases the signal to noise ratio. Smoothing is also used as a way to decrease between-subject variability. In our experiments, we usually don't apply smoothing in order to measure how our methods are able to handle fined grained details. While there is currently no consensus on smoothing, we observe that a smoothing of 3 mm generally improves most analysis on fMRI data.

3.1.4.5 *Frequency filtering*

Heating of the scanner causes a low temporal frequency noise to appear. In order to remove such noise, a high-pass filtering is applied with a cut-off frequency of $0.01\mathrm{Hz}$. Sometimes a low-pass filter with a cut-off frequency of 0.1 Hz is also used since artifacts induced by motion are usually of higher frequency than the signal of interest.

3.1.4.6 *Spatial normalization*

Each subject has a brain of different size and shape. The goal of spatial normalization is to align brain images of different subjects in order to reduce anatomical variability.

The most widespread technique is to register all images into the Montreal Neurological Institute template (MNI template). The MNI template is built from 305 anatomical images that are aligned based on anatomical features via a non-linear registration model. The non-linear registration model uses rigid body transformations followed by non-linear diffeomorphic deformations to better match brain shape.

The projection of fMRI data to the MNI space proceeds in two steps. First, a high dimensional anatomical image of each subject is aligned on the MNI template. Then the fMRI images of each subjects are aligned on the anatomical image of the same subject (this step is called *co-registration*). The fMRI images are mapped to the MNI template by composing the two transformations.

3.1.4.7 *Masking*

It is often the case that only a subpart of the brain is of interest. When this is the case, we use a mask to only keep the set of voxels corresponding to particular locations defined by the mask. In the case of our experiments, we use a gray-matter mask, since this is the only part that matters when one wants to capture functional activations in the brain that reflect behavioral responses.

3.1.4.8 *Runs*

For the comfort of participants and to ensure their active participation, it is in general not possible to use the scanner without interruptions for more than 15 minutes. When long acquisitions must be performed, they are split into short runs of approximately 15 minutes.

3.1.5 *Example of fMRI Datasets*

In this subsection we provide examples of fMRI datasets. These datasets will be used to evaluate the methods developed in this thesis. Datasets are preprocessed with FSL `http://fsl.fmrib.ox.ac.uk/fsl` and SPM `https://www.fil.ion.ucl.ac.uk/spm/software` using slice timing correction, distortion correction spatial realignment, co-registration to the T1 image and affine transformation of the functional volumes to a template brain (MNI). Using Nilearn [4], preprocessed data are masked (using a full brain mask available at http://cogspaces.github.io/assets/data/hcp_mask.nii.gz), detrended and standardized (so that any voxel's timecourse has zero mean and unit variance). We also apply a high-pass filter and a low-pass filter with cut-off frequencies of 0.01 Hz and 0.1 Hz respectively).

3.1.5.1 *Sherlock*

In the *SHERLOCK* dataset, 17 participants are watching "Sherlock" BBC TV show (episode 1). These data are downloaded from http://arks.princeton.edu/ark:/88435/dsp01nz8062179. Data were acquired using a 3T scanner with an isotropic spatial resolution of 3 mm. More information including the preprocessing pipeline is available in [35]. Subject 5 is removed because of missing data, leaving us with 16 participants. Although SHERLOCK data contains originally only 1 run, we split it into 4 blocks of 395 timeframes and one block of 396 timeframes for the needs of our experiments.

3.1.5.2 *Forrest*

In the *FORREST* dataset, 20 participants are listening to an audio version of the movie Forrest Gump. FORREST data are downloaded from OpenNeuro [126]. Data were acquired using a 7T scanner with an isotropic spatial resolution of 1 mm (see more details in [65]. More information about the forrest project can be found at http://studyforrest.org. Subject 10 and run 8 are discarded because of missing data. We therefore use full brain data of 19 subjects split in 7 runs of respectively 451, 441, 438, 488, 462, 439 and 542 timeframes.

3.1.5.3 *CamCAN*

In the fMRI *CamCAN* dataset, 647 participants aged from 18 to 88 years are watching Alfred Hitchcock's "Bang! You're Dead" (edited so that it lasts only 8 minutes). CamCAN consists of data obtained from the CamCAN repository (available at http://www.mrc-cbu.cam.ac.uk/datasets/camcan/) (see [146] and [137]). We use all available subjects and runs yielding 647 participants and 1 run of 193 timeframes.

3.1.5.4 *Raiders*

The *RAIDERS* dataset reproduces the protocol described in [69]. 10 participants are watching the movie "Raiders of the Lost Ark". This dataset pertains to the Individual Brain Charting dataset [124, 125]. We use full brain data of 10 subjects split in 9 runs of respectively 374, 297, 314, 379, 347, 346, 350, 353 and 211 timeframes. Note that the Raiders dataset is different from the one used in [36], as it involves different subjects, and because data were acquired at NeuroSpin using a 3T scanner with an isotropic spatial resolution of 1.5 mm. The *raiders-full*

Dataset	**Subjects** m	**Runs**	**Average run length (in timeframes)**	**Voxels (per subject)** v
CLIPS	10	17	325	212445
SHERLOCK	16	5	395	212445
RAIDERS	10	9	330	212445
FORREST	19	7	465	212445
CamCAN	647	1	193	212445

Table 3.1: **Datasets description**

dataset [124, 125] is an extension of the *raiders* dataset where the first two scenes of the movie are shown twice (130 mins).

3.1.5.5 *CLIPS*

The CLIPS dataset reproduces the protocol of original studies described in [108] and [71]. 10 participants are exposed to short clips. The data were acquired in 17 runs of 325 timeframes. The CLIPS dataset also pertains to the Individual Brain Charting dataset ([124, 125]). The CLIPS and RAIDERS data are available in OpenNeuro under the identification number: ds002685. Protocols on the visual stimuli presented are available in a dedicated repository on Github: `https://github.com/hbp-brain-charting/public_protocols`.

Unless stated otherwise we use spatially unsmoothed data, except for the *sherlock* dataset, for which the available data are already preprocessed with a 6 mm spatial smoothing. The temporal resolution or repetition time (TR) is 2s for all datasets except for the Sherlock dataset where the TR is 1.5s.

A summary about the size of each dataset is available in Table 3.1. Note in particular that all datasets have been resampled to 2mm isotropic resolution, leading to 212,445 voxels in the brain mask.

3.1.6 *An example of univariate analysis: analyzing task fMRI data via the general linear model*

In this section, we describe a framework successfully applied to identify brain regions involved in specific tasks: the general linear model (GLM) [55]. We refer the reader to Poline [129] for a more detailed description of the model and only cover here what we consider to be the most important parts.

We associate to each image t a set of numbers $\mathbf{s}(t)$ that describe the experiment and nuisance parameters. For instance, in an experiment where the subject is either asked to listening a sentence or reading a sentence we could have the association: $\mathbf{s}(t) = (\phi(t), \tau(t), x(t), y(t), z(t))$ where $\phi(t) = 1$ if the subject is reading a sentence and 0 otherwise, $\tau(t) = 1$ if the subject is listening to a sentence and 0 otherwise and $x(t), y(t), z(t)$ describe the position of the head. While ϕ and τ describe the experiment, x, y, z describe nuisance parameters. Note that in practice we use many more nuisance regressors, such as parameters describing the rotation of the head, heartbeats, respiration rhythm and really any other parameter that might introduce artifacts. We also usually convolve the regressors describing the experiment with the hemodynamic function so that they are closer to the actual brain response and sometimes also its derivative to account for small temporal deviations from our model of hemodynamic response. We call $S \in \mathbb{R}^{k,n}$ the *design matrix* such that column t of S is given by $\mathbf{s}(t)$. Then, the general linear model sees the data of one subject $X \in \mathbb{R}^{\nu \times n}$ as a linear combination of the design matrix S with additive noise N.

$$X = AS + N \tag{3.1}$$

where $A \in \mathbb{R}^{\nu \times k}$ contains k spatial maps in the sense that column j of A can be plotted as a brain image representing the localization of activity described by the row j of S. The noise is often assumed to be Gaussian with unit variance so that A can be obtained as the result of a linear regression. Another common hypothesis is to assume temporal correlation of the noise in the form of an AR-1 process. In this case, we consider instead the data XB where $B \in \mathbb{R}^{n \times n}$ is chosen so that the noise BN has no temporal correlation allowing us to perform linear regression.

The result of the GLM procedure is often given by *contrasts maps* where the spatial maps corresponding to conditions of interest are subtracted. In our reading versus listening example, we would typically display the first column of A (that corresponds to ϕ which describes the activation related to reading a sentence) minus the second column of A (that corresponds to τ which describes the activation related to listening to a sentence). An example of "reading versus listening contrast is displayed in Figure 3.2.

In this thesis, we always work with unthresholded contrast maps like the one in Figure 3.2. For interpretation purposes, contrast maps are usually z-scored and thresholded to keep only activity that is significantly different from zero. The GLM model is called univariate since interactions between voxels are not modeled. Univariate methods are often criticized for their inability to capture well correlations and interactions between brain-wide measurements. This makes it difficult to precisely locate brain functions from brain maps. An approach to overcome this issue is to train classifiers to *decode* brain maps, i.e to discriminate between different stimulus of task types [95, 138,

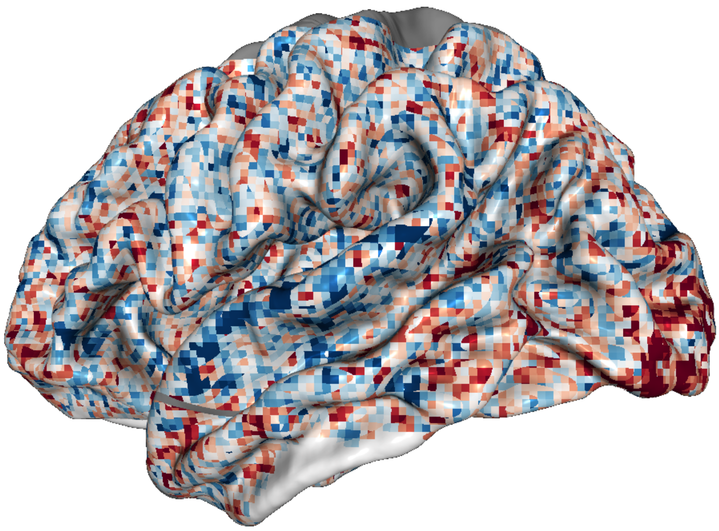


Figure 3.2: Contrast "sentence listening versus sentence reading" computed from the fMRI data of one subject in the IBC dataset [124, 125]. This contrast was downloaded from neurovault [59] (collection 2138 subject 1 session 0) .

151]. When linear classifiers are used, the weights of this classifiers localize the brain functions. This approach can be more powerful, as classifiers take correlations between voxels into account. Decoding is also popular for individual imaging-based diagnosis.

3.2 MAGNETO ELECTRO ENCEPHALOGRAPHY (MEG)

Most material in this section is inspired from the book of Riitta Hari and Aina Puce [66].

3.2.1 *Principle of MEG*

When a neuron fires, it emits a current that generates a magnetic field. By recording the magnetic field close to the skull, we gain insight on neural activity.

A limiting factor is that the magnetic fields induced by neuronal currents are extremely weak (on the order of $10\mathrm{fT}$) which is much lower than the ambient magnetic noise ($10^8\mathrm{fT}$). Therefore recordings are performed in a shielded room and extremely sensitive magnetometers are used. The best current tools can measure the magnetic field generated by approximately 50 000 neurons oriented in the same direction.

MEG recording device also include gradiometers in addition to magnetometers. These are sensitive to the local variations of the magnetic field. However in our experiments, we only use the magnetometers data.

MEG has a temporal resolution on the order of the ms which is essentially the same as electroencephalography (EEG). This is much better than fMRI. The spatial resolution is similar to that of EEG and is on the order of the centimeter. In general the brain sources are slightly better localized in MEG since the magnetic field is not affected by changes in conductivity in the head.

3.2.2 *Solving the inverse problem*

In order to locate activation from the brain measurements, we need to solve an inverse problem: what kind of sources can generate the magnetic field we observe ? This section gives an overview of the sLORETA method [113] and is strongly inspired from the tutorial of the authors [112].

From a vector describing the 3D current density inside the brain $\mathbf{j} \in \mathbb{R}^{3\times\nu}$ (ν is the number of voxels) we can predict the magnetic field recorded by the sensors $\mathbf{b} \in \mathbb{R}^{p}$ (p is the number of sensors) by solving Maxwell's equations. This is called the *forward model*. Models describing the geometry of the head (head models) are used to provide an approximate solution. We have therefore

$$\mathbf{b} = K\mathbf{j} \tag{3.2}$$

where $K \in \mathbb{R}^{p\times 3\nu}$ is the solution to the forward model. K can be seen as $K = [K_1 \dots K_\nu]$ where $K_i \in \mathbb{R}^{p\times 3}$ describes how voxel i contributes to the measured magnetic field $\mathbf{b}$.

Note that if $\mathbf{b}$ and K are given there is an infinite number of possible $\mathbf{j}$ that satisfy equation (3.2), so the inverse problem must be regularized in order to select a solution among all possible ones. A common parametrization is to assume that the current at voxel i, $\mathbf{j}_i \in \mathbb{R}^3$ verifies:

$$\mathbf{j}_i = (K_i^\top C K_i)^{-\frac{1}{2}} K_i^\top C \mathbf{b} \tag{3.3}$$

for a symmetric matrix $C \in \mathbb{R}^{p\times p}$. In sLORETA, we have $C = (KK^\top + \alpha I)^\dagger$ where $\dagger$ represents Moore's pseudo inverse, I is the identity matrix and α is an hyperparameter.

Taking a point wise source $\mathbf{a} \in \mathbb{R}^3$ at voxel i_* equation (3.2) yields $\mathbf{b} = K_{i_*}\mathbf{a}$. Then, $\mathbf{j}_i = (K_i^\top C K_i) K_{i_*}\mathbf{a}$ is such that $\|\mathbf{j}_i\|^2$ is maximum at $i = i_*$ showing that the method can correctly recover point sources.

3.2.3 *Preprocessing steps*

This section is inspired from the preprocessing recommandations in [77] and the book of Riitta Hari and Aina Puce [66].

3.2.3.1 *Temporal filtering*

Temporal filtering is performed in two steps. First an analogous pass band filter selects a large band of frequencies between 0.01Hz and 200Hz. Then, a digital filtering is applied. Typically, a low-pass filtering with cut-off frequency at 40Hz is used as it is removes the line frequency at 50Hz while keeping most of the brain's signals.

3.2.3.2 *Independent component analysis*

Independent component analysis (ICA) can be used to isolate artifacts due to heart beating or muscle contraction. As will be seen in section 4.1. ICA extracts independent components from the data. The data can be cleaned by removing components corresponding to artifacts [78].

3.2.3.3 *Maxwell filtering*

Maxwell filtering also called signal space separation (SSS) [145] identifies the contributions to the magnetic field from brain sources outside the brain and removes it. This decreases the effects of outside noise and even allows to scan subjects with implanted stimulators.

3.2.4 *Experimental designs*

In magneto-encephalography, we measure *event related fields* (ERF) related to an *evoked response* in the brain. Experiments are often repeated several times yielding a number of *trials*. By averaging the event related fields over trials, the signal to noise ratio increases.

3.2.5 *MEG datasets*

In this section, we give examples of MEG datasets. These examples are used in the rest of the thesis to evaluate the methods we develop.

The *Sinusoidal Phantom MEG* dataset uses data collected with a realistic head phantom, which is a plastic device mimicking real electrical brain components. Eight current dipoles positioned at different locations can be switched on or off. We only consider the 102 magnetometers. An epoch corresponds to 3 s of MEG signals where a dipole is switched on for 0.4 s with an oscillation at 20 Hz and a peak-to-peak amplitude of 200 nAm. We have access to 100 epochs per dipole. Maxwell filtering is applied on the data as well as low-pass filtering with a cut-off frequency at 40 Hz.

The *Elekta Phantom MEG* dataset also uses data collected with a realistic head phantom and is available as part of the Brainstorm application [143]. This dataset uses 32 dipoles positioned at different locations. Like in the *Sinusoidal Phantom MEG* dataset, we only consider the 102 magnetometers and apply Maxwell filtering and low-pass filtering with a cut-off frequency at 40 Hz. The dipoles emit a signal at either a strong, medium or low amplitude yielding 3 different datasets. We use the dataset where the emitted signal is very strong to recover the true signal by performing a PCA with 1 component. Then we work with the dataset where the emitted signal is the weakest. Each epoch corresponds to 301 samples and 20 epochs are available in total.

The *CamCAN* dataset [137] contains the MEG data of 647 different subjects exposed to an audio-visual stimuli. More precisely, subjects are presented simultaneously an auditory stimuli lasting 300ms at frequency 300, 600 or 1200 Hz and a checkerboard pattern lasting 34ms. 120 trials are available.

3.3 CONCLUSION

In this chapter, we described the principles on which fMRI and MEG imaging are based as well as the standard preprocessing pipelines and the datasets used in this thesis.

In the next chapter, we review several methods to perform unsupervised analysis of neuroimaging data.

4 REVIEW OF SELECTED UNSUPERVISED METHODS POPULAR IN NEUROIMAGING STUDIES

When exposed to naturalistic stimuli (e.g. movie watching or simulated driving), subjects' experience is closer to their every-day life than with classical psychological experiments. This makes naturalistic paradigms an attractive class of stimulation protocols for brain imaging. However, such stimulations are difficult to model, therefore the statistical analysis of the data using supervised regression-based approaches is challenging. This has motivated the use of unsupervised learning methods that do not make assumptions about what triggers brain activations in the presented stimuli.

In this chapter, we first present *independent component analysis* (ICA), a widely used unsupervised method for neuroimaging studies routinely applied on individual subject electroencephalography (EEG) [97], magnetoencephalography (MEG) [154] or functional MRI (fMRI) [100] data. Then, we review *multiview* unsupervised techniques that leverage the availability of data from multiple subjects performing the same experiments.

4.1 INDEPENDENT COMPONENT ANALYSIS

Independent component analysis (ICA) models a set of signals as the product of a *mixing matrix* and a *component* matrix containing independent components. As will be seen in this section, the required assumptions on the independent components to guarantee identifiability are rather weak, making ICA a method of choice to analyze the data of subjects exposed to a stimulus that is difficult to quantify.

ICA is applied to fMRI data to analyze resting state data [16] or when subjects are exposed to natural [98] [14] or complex stimuli [30]. In M/EEG processing, it is widely used to isolate acquisitions artifacts from neural signal [78], and to identify brain components of interest [43, 155].

Mainly for computational reasons, it is often assumed that the number of components k is much lower than the dimensionality of the data v. However in ICA, the dimensionality of the data must be equal to the number of components. We therefore first present principal component analysis, a standard method to perform dimension reduction, then present non-Gaussian ICA and non-stationary ICA.

4.1.1 *Principal component analysis (PCA)*

Let us assume our data are given by n observations of a random vector $x \in \mathbb{R}^v$ that we stack into a matrix $X \in \mathbb{R}^{v \times n}$. The data are centered ($\mathbb{E}[x] = 0$). Principal component analysis (PCA) yields an orthonormal familly of p vectors $R \in \mathbb{R}^{v \times p}$, $R^\top R = I_p$, such that the projected data $RR^\top x$ does not yield a large mean reconstruction error in Frobenius norm. The corresponding optimization problem is given by:

$$\underset{R, R^\top R = I}{\operatorname{argmin}} \mathbb{E}_x \|x - RR^\top x\|^2 \tag{4.1}$$

$$= \underset{R, R^\top R = I}{\operatorname{argmin}} \mathbb{E}_x \operatorname{trace}(x^\top (I - RR^\top)(I - RR^\top)x) \tag{4.2}$$

$$= \underset{R, R^\top R = I}{\operatorname{argmin}} \mathbb{E}_x (\|x\|^2 - \|R^\top x\|^2) \tag{4.3}$$

$$= \underset{R, R^\top R = I}{\operatorname{argmax}} \mathbb{E}_x \|R^\top x\|^2 \tag{4.4}$$

$$= \underset{R, R^\top R = I}{\operatorname{argmax}} \operatorname{trace}(R^\top \mathbb{V}[x] R) \tag{4.5}$$

Therefore the matrix R is just given by the first p eigenvectors of $\mathbb{V}[x]$. In practice, we can use the sample variance $\frac{1}{n} XX^\top$ to estimate $\mathbb{V}[x]$. However, such an estimate is not practical in high dimension: when v is very large $XX^\top$ is a prohibitively large matrix.

In such case, assuming n is small, we rely instead of the *singular value decomposition* (SVD) of X:

$$X = UDV \tag{4.6}$$

where $U \in \mathbb{R}^{v \times n}$ is an orthogonal matrix ($U^\top U = I_t$) of *left-singular vectors*, $D \in \mathbb{R}^n$ is a positive diagonal matrix of *singular values* and $V \in \mathbb{R}^{n \times n}$ is an orthogonal matrix of *right-singular vectors*. Such a decomposition can be computed in $\tilde{O}(vn^2)$ operations which is not prohibitive when n is small enough. Every matrix has a singular value decomposition and, provided that all singular values are distinct, this decomposition is unique up to a permutation and sign indeterminacy. More precisely, if UDV and $U'D'V'$ are two singular decomposition of X and if D has only distinct values, then $U' = U\Pi\Xi$, $D' = \Pi^\top D \Pi$ and $V' = \Xi \Pi^\top V$ where Π is a permutation matrix and Ξ a diagonal matrix with diagonal values in $\{-1, 1\}$.

Let $X = UDV$ be a singular value decomposition of X, we have $XX^\top = UD^2U^\top$ and therefore the left-singular vectors U_p corresponding to the largest p singular values of X yield the p largest eigenvectors of $XX^\top$. Therefore R is given by the left-singular vectors corresponding to the largest p singular values: $R = U_p$.

Sometimes the PCA includes whitening $R = U_p D_p^{-1}$ where D_p contains the p largest singular values of X so that the components of $R^\top x$ are uncorrelated (but R is no longer orthonormal). In this thesis what we call PCA does not include signal whitening.

After the data are reduced, we can perform independent component analysis (ICA). ICA can exploit a number of different properties of the signal. We focus in the next section on non-Gaussian ICA.

4.1.2 *Non Gaussian ICA*

ICA sees the data $x \in \mathbb{R}^p$ as a linear mixtures of components $\mathbf{s} \in \mathbb{R}^p$. Data are therefore modeled as:

$$x = A\mathbf{s} \tag{4.7}$$

where A is the unmixing matrix and $\mathbf{s}$ is a vector with independent components, of which at most one is Gaussian and whose densities are not reduced to a point-like mass. Without loss of generality, the data x are assumed centered ($\mathbb{E}[x] = 0$). This section relies heavily on the ICA review [76] and on [32].

4.1.2.1 *Identifiability*

A matrix P that can be written $P = \Xi\Pi$, where Ξ is a diagonal matrix and Π is a permutation matrix, is called a *scale and permutation matrix*. If furthermore Ξ only has diagonal values in $\{-1, 1\}$, then P is a *sign and permutation matrix*.

It is easily seen that if $\mathbf{s}$ is a vector with independent components, of which at most one is Gaussian and whose densities are not reduced to a point-like mass, so is $P\mathbf{s}$ where P is a scale and permutation matrix. Therefore, if $x = A\mathbf{s}$, take $A' = AP^\top$ and $\mathbf{s}' = P\mathbf{s}$ and we have $x = A'\mathbf{s}'$. In other words, there exists a permutation and scaling indeterminacy.

To fix the scaling indeterminacy, we assume that $\mathbf{s}$ has unit variance ($\mathbb{V}[\mathbf{s}] = I$). In addition, we assume that the data are whitened ($\mathbb{V}[x] = I$). Note that the whitening assumption does not imply any loss of generality. If a matrix H is used to whiten the data, the mixing matrix of the unwhitened data is given by $H^{-1}A$ where A is the unmixing matrix obtained on the whitened data. With these assumptions, A is orthogonal. Indeed, $\mathbb{V}[x] = A^\top A = I$.

Are there any other indeterminacies ? Assume that there exists two mixing matrices A_1 and A_2 such that $x = A_1\mathbf{s}_1$ and $x = A_2\mathbf{s}_2$ with A_1 and A_2 orthogonal. Then $\mathbf{s}_1 = O\mathbf{s}_2$ where $O = A_2A_1^\top$ is an orthogonal matrix. The following theorem in [38] shows that O is necessarily a sign and permutation matrix.

Theorem 8. *Let $\mathbf{s}_2$ be a vector with independent components, of which at most one is Gaussian, and whose densities are not reduced to a point-like mass. Let O be an orthogonal matrix and $\mathbf{s}$ such that $\mathbf{s} = O\mathbf{s}_2$. Then, $\mathbf{s}$ has independent component if and only if O is a sign and permutation matrix.*

As a result, if we assume that at most one component is Gaussian and that components densities are not reduced to a point-like mass, ICA is identifiable up to a scale and permutation indeterminacy.

4.1.2.2 *Infomax: Maximum likelihood estimation*

Infomax is introduced in [19] and although initially formulated as a maximum entropy problem, it has been shown in [31] to be equivalent to maximum likelihood.

Assume for simplicity that the components have the same density δ, denoting $f(x) = -\log(\delta(x))$ and $f(\mathbf{x}) = \sum_{i=1}^m f(x_i)$ the expected negative log-likelihood is given by:

$$\begin{aligned}
\mathcal{L}(W) &= \mathbb{E}[-l(\mathbf{x}, \theta)] = \mathbb{E}[-\log(p(\mathbf{x}))] && (4.8)\\
&= -\log(|W|) - \mathbb{E}[\log(\delta(W\mathbf{x}))] && (4.9)\\
&= -\log(|W|) - \mathbb{E}[\log(\delta(\mathbf{y}))] && (4.10)\\
&= -\log(|W|) + \mathbb{E}[f(\mathbf{y})] && (4.11)\\
&= -\log(|W|) + \mathbb{E}[\sum_{i=1}^{m} f(y_i)] && (4.12)
\end{aligned}$$

where $\mathbf{y} = W\mathbf{x}$ and y_i is the i-th component of $\mathbf{y}$. At line (4.9) we used the change of variable $\mathbf{s} = W\mathbf{x}$.

4.1.2.3 *Optimizing the maximum likelihood: relative gradient, Hessian and approximations*

The matrix W needs to be invertible. In practice, if the constraint is not enforced, we could see numerical instabilities appear. A simple rule that preserves the invertibility of W is to use updates of the form:

$$W \leftarrow (I + \alpha D)W \qquad (4.13)$$

where α is a small step-size and D is a direction to be found. Following the same reasonning as in section 2.2.1.1, we assume a small step-size and write at first order:

$$\begin{aligned}
\underset{D, \|D\|=1}{\operatorname{argmin}} \mathcal{L}((I + \alpha D)W) &= \underset{D, \|D\|=1}{\operatorname{argmin}} \mathcal{L}(W) + \langle \frac{\partial \mathcal{L}(W)}{\partial W}, \alpha D W \rangle && (4.14)\\
&= \underset{D, \|D\|=1}{\operatorname{argmin}} \mathcal{L}(W) + \langle \frac{\partial \mathcal{L}(W)}{\partial W} W^\top, \alpha D \rangle && (4.15)\\
&= -\frac{\frac{\partial \mathcal{L}(W)}{\partial W} W^\top}{\|\frac{\partial \mathcal{L}(W)}{\partial W} W^\top\|} && (4.16)
\end{aligned}$$

Therefore the steepest direction for that update rule is $D = \frac{\partial \mathcal{L}(W)}{\partial W} W^\top$ and therefore updates are given by: $W \leftarrow (I - \alpha G)W$ where $G =$

$\frac{\partial \mathcal{L}(W)}{\partial W} W^\top$ is called the *relative gradient* [34]. Second order extensions of the method are obtained by following the steps in section 2.2.1.2. The corresponding Hessian is called the *relative Hessian*.

Let us follow [3] and use a quasi-Newton algorithm to minimize the likelihood. The relative gradient and Hessian of $\mathcal{L}$ are given by:

$$G = \mathbb{E}[f'(\mathbf{y})\mathbf{y}^\top] - I_p, \tag{4.17}$$

where $\mathbf{y} = W\mathbf{x}$ and $f'(\mathbf{y}) = \frac{\partial f(\mathbf{y})}{\partial \mathbf{y}}$ and

$$H_{abcd} = \delta_{ad}\delta_{bc} + \delta_{ac}\mathbb{E}[f''(y_a)y_b y_d] \tag{4.18}$$

Following [3], we can approximate the Hessian by

$$\tilde{H}_{abcd} = \delta_{ad}\delta_{bc} + \delta_{ac}\delta_{bd}\Gamma_{ab} \tag{4.19}$$

where $\Gamma_{ab} = \mathbb{E}[f''(y_a)y_b^2]$. The approximation is exact when the true unmixing matrix is found since $\mathbb{E}[s_b s_d] = s_b^2\delta_{bd}$.

The updates are then given by:

$$W \leftarrow (I - \alpha\tilde{H}^{-1}G)W \tag{4.20}$$

The approximated Hessian $\tilde{H}$ is block diagonal. Indeed we have for any matrix M:

$$\begin{aligned} \begin{bmatrix} (\tilde{H}M)_{ab} \\ (\tilde{H}M)_{ba} \end{bmatrix} &= \begin{bmatrix} \Gamma_{ab} & 1 \\ 1 & \Gamma_{ba} \end{bmatrix} \begin{bmatrix} M_{ab} \\ M_{ba} \end{bmatrix} \qquad \text{if } a \neq b \\ (\tilde{H}M)_{aa} &= (1 + \Gamma_{aa})M_{aa} \end{aligned} \tag{4.21}$$

So that each block corresponds to a pair (a, b) and is of size 2 if $a \neq b$ and of size 1 when $a = b$. Therefore $\tilde{H}$ can be easily regularized and inverted.

4.1.2.4 *Robustness to density mismatch*

In practice, we observe that sometimes, the quasi-Newton algorithm described in the previous section fails to recover the true mixing matrix. This happens when the sources used in the model are too far from the actual generating sources. In this section, we explain why this happens and derive stability conditions. This follows the work done in [33].

When the density of the sources corresponds to the one used in the model, one recovers the correct unmixing matrices via the maximum likelihood estimate in the limit of large samples. This comes from the consistency of maximum likelihood estimators. When this is not the case, a mismatch appears which can be quantified. Let us denote $\mathbf{x}^*$ the true data. As highlighted in (2.54) the expected negative log-likelihood coincides, up to a constant, with the KL divergence between the true distribution of the data and the distribution of the data as hypothesized in the model:

$$\mathcal{L}(W) \tag{4.22}$$
$$= D_{KL}(p(\boldsymbol{x}^*), p(W^{-1}\mathbf{s})) \tag{4.23}$$
$$= D_{KL}(p(W\boldsymbol{x}^*), p(\mathbf{s})) \tag{4.24}$$
$$= D_{KL}(p(W\boldsymbol{x}^*), \prod_i p_i(\boldsymbol{w}_i\boldsymbol{x}^*)) + D_{KL}(\prod_i p_i(\boldsymbol{w}_i\boldsymbol{x}^*), p(\mathbf{s})) \tag{4.25}$$
$$= D_{KL}(p(W\boldsymbol{x}^*), \prod_i p_i(\boldsymbol{w}_i\boldsymbol{x}^*)) + D_{KL}(\prod_i p_i(\boldsymbol{w}_i\boldsymbol{x}^*), \prod_i p_i(s_i)) \tag{4.26}$$
$$= D_{KL}(p(W\boldsymbol{x}^*), \prod_i p_i(\boldsymbol{w}_i\boldsymbol{x}^*)) + \sum_i D_{KL}(p_i(\boldsymbol{w}_i\boldsymbol{x}^*), p_i(s_i)) \tag{4.27}$$

where equation (4.24) comes from the invariance of the KL divergence and we denote $\prod_i p_i(x_i)$ the product of the marginal densities of $\mathbf{x}$.

The first term in equation (4.27) is the mutual information. It quantifies the independence of unmixed data $W\mathbf{x}$. The second term quantifies the mismatch between the assumed distribution of the components and the marginals of unmixed data. Therefore if the assumed components are too far from the true components, the recovered unmixing matrices may be far from the true ones.

Looking at the relative gradient G, we see that if components are truly independent, the true unmixing matrix will be a stationary point of the loss up to a scaling: $G(\Lambda A^{-1}) = 0$ where $\Lambda = \text{diag}(\lambda_1 \dots \lambda_p)$. However, in order for the quasi-Newton to reach this point, it must be a local minimum (the Hessian must be definite positive).

Let us therefore consider the Hessian H at point $W = \Lambda A^{-1}$ where Λ is chosen such that $G = 0$. The unmixed data $\mathbf{y} = W\mathbf{x} = \Lambda\mathbf{s}$ are independent and therefore, we have $\tilde{H} = H$. As already mentioned, $\tilde{H}$ is block diagonal with blocks described by equation (4.21). The condition for stability is that the Hessian is positive definite. Therefore all the blocks need to be positive definite so we get the conditions:

$$\forall a, \ \ 1 + \Gamma_{aa} > 0 \qquad \text{for blocks of size 1} \tag{4.28}$$
$$\forall a \neq b, \ \ \Gamma_{ab}\Gamma_{ba} > 1 \qquad \text{for blocks of size 2} \tag{4.29}$$

When the conditions are satisfied, the quasi-Newton algorithms will recover the true unmixing matrices if initialized close to a solution. However, we have no theoretical guarantees of convergence because of the non-convexity of the problem.

4.1.3 *Non-stationary ICA and joint diagonalization*

Up to now we have assumed that samples were independent an identically distributed. In non-stationary ICA, samples are no longer identically distributed as the distribution can vary over time. This section follows the work in [120]. The components are assumed to be Gaussian with a variance that varies between samples but is assumed to be piece-wise constant $\Sigma_t = \Sigma_k$ for $t \in T_k$ where $(T_k)_k$ is a partition of $[1, 2, \dots, n]$ and Σ_k is a positive diagonal matrix.

Denote $X[i]$ to be sample i of the observed data X. The negative empirical expected log-likelihood is given by:

$$\mathcal{L} = -\log(|W|) + \frac{1}{n}\sum_k \left(\frac{1}{2}\sum_{i\in T_k} \|\Sigma_k^{-\frac{1}{2}} W X[i]\|^2 + \frac{1}{2}\log(|\Sigma_k|)\right) \tag{4.30}$$

Denoting $C_k = \sum_{i\in T_k} X[i]X[i]^\top$ we have

$$\mathcal{L} = -\log(|W|) + \frac{1}{n}\sum_k \left(\frac{1}{2}\mathrm{trace}(\Sigma_k^{-1} W C_k W^\top) + \frac{1}{2}\log(|\Sigma_k|)\right) \tag{4.31}$$

Minimizing $\mathcal{L}$ with respect to Σ_k yields $\Sigma_k = \mathrm{diag}(WC_kW^\top)$ and therefore up to a constant:

$$\mathcal{L} = -\log(|W|) + \frac{1}{n}\sum_k \left(\frac{1}{2}\log(|\mathrm{diag}(WC_kW^\top)|)\right) \tag{4.32}$$

which up to a constant can be rewritten:

$$\mathcal{L} = \frac{1}{2n}\sum_k \left(\frac{\log(|\mathrm{diag}(WC_kW^\top)|)}{\log(|WC_kW^\top|)}\right) \tag{4.33}$$

$\mathcal{L}$ is a joint diagonalization criterion. The optimal W can be found via a quasi-Newton method very similar to the one we used for non-Gaussian ICA [2]. However, the updates only depend on the covariance matrices C_k and no longer on the number of samples making this approach very fast when the number of samples is large.

4.1.4 *Other approaches and extensions*

NON-WHITE ICA While non-stationary ICA relaxes the assumption that samples are identically distributed, non-white ICA relaxes the assumption that samples are independent. Following the work in [118], the entropy of a stationary Gaussian process $\mathbf{y}(t)$ is given by:

$$h(\mathbf{y}(t)) = \frac{1}{4\pi}\int_{-\pi}^{\pi} \log(\det(4\pi^2 e F(\lambda)))d\lambda \tag{4.34}$$

where F is the spectral density matrix of the process. Then the authors define the Gaussian mutual information between stationary processes $y_1(t), \dots, y_k(t)$ by:

$$I_g(y_1, \dots, y_k) = \frac{1}{4\pi}\int_{-\pi}^{\pi} \log(\det(\mathrm{diag}(F(\lambda)))) - \log(\det(F(\lambda)))d\lambda \tag{4.35}$$

which is a joint diagonalization criterion similar to (4.33).

EXTENDED INFOMAX: MORE FLEXIBLE DENSITIES In Infomax, the densities are fixed once and for all. However, it can be useful to make them more flexible. Note that the updates in equation (4.20) and the stability criterion in equation (4.28) only depend on the score function $f'(\mathbf{y}) = \frac{\partial f(\mathbf{y})}{\partial \mathbf{y}}$. Choosing $f'(x) = x + \tanh(x)$ would only separate super-Gaussian sources while choosing $f'(x) = x - \tanh(x)$ would only separate sub-Gaussian sources. In [86], the authors therefore propose an extended Infomax algorithm, where the score function can change based on observed statistics so that the stability criterion is fulfilled in any case.

ICA VIA MUTUAL INFORMATION MINIMIZATION In the limit of large samples, we have seen in equation (4.27) that maximum likelihood minimizes the sum of two terms: the mutual information and a second term that quantifies the mismatch between the assumed distribution of the components and the marginals of unmixed data. It might be desirable to only minimize the mutual information and forget the second term. Since the mutual information depends on the entropy, different works such as [72] [88] [89] or [90] studied how the entropy can be estimated accurately and efficiently.

4.2 ANALYSIS OF MULTIVIEW DATA

In this section, we present multiview unsupervised techniques suited to analyze the data of multiple subjects exposed to the same complex stimuli. Such techniques assume some similarity between the data of different subjects. This assumption can be justified by the findings of [68] showing that brains exposed to the same natural stimuli exhibit synchronous activity. The task of finding common patterns or responses that are shared between subjects is called *shared response modeling*.

In the general linear model presented in section 3.1.6, the shared response is assumed to be known. Therefore, multiple subjects can be studied separately assuming that the data of different subjects are independent given the shared response. In the unsupervised setting it may not be so straightforward to deal with multiple subjects and therefore many different methods for data-driven multivariate analysis of neuroimaging group studies have been proposed. We summarize the characteristics of some of the most commonly used ones.

4.2.1 *Multiset canonical correlation analysis*

Canonical correlation analysis is initially designed to find a linear combination that maximizes the correlation between two datasets. The extension to more than two datasets is ambiguous, and many different generalized CCA methods have been proposed. [79] introduces 6 objective functions that reduce to CCA when $m = 2$ and [107] considered 4 different possible constraints leading to 24 different formulations of Multiset CCA.

In [79], the different formulations of multiset CCA are not derived from a probabilistic model. However, later works (see [92] or [5]) have shown that some formulations of multiset CCA can be related to probabilistic models.

In this section, we present the formulation refered to in [107] as "SUMCORR with constraint 4" which is one of the fastest to fit.

Let us consider $X_1, \dots, X_m \in \mathbb{R}^{p,n}$, m datasets and consider the following (SUMCORR) objective:

$$\max_{\mathbf{a}_1 \in \mathbb{R}^v, \dots, \mathbf{a}_m \in \mathbb{R}^v} \sum_{i=1}^{m} \sum_{j=1}^{m} \langle \mathbf{a}_i, X_i X_j^\top \mathbf{a}_j \rangle \tag{4.36}$$

This objective can be arbitrarily large if not constrained. Constraint 4 is given by:

$$\sum_{i=1}^{m} \langle \mathbf{a}_i, X_i X_i^\top \mathbf{a}_i \rangle = 1 \tag{4.37}$$

The Lagrangian is given by:

$$\sum_{i=1}^{m} \sum_{j=1}^{m} \langle \mathbf{a}_i, X_i X_j^\top \mathbf{a}_j \rangle - \lambda \left(\sum_{i=1}^{m} \langle \mathbf{a}_i, X_i X_i^\top \mathbf{a}_i \rangle - 1 \right) \tag{4.38}$$

Taking the gradient with respect to $\mathbf{a}_i$ we obtain

$$\sum_{j=1}^{m} X_i X_j^\top \mathbf{a}_j = \lambda X_i X_i^\top \mathbf{a}_i \tag{4.39}$$

This is a generalized eigenvalue problem of the form $C\mathbf{a} = \lambda D\mathbf{a}$ where C is a block matrix where block i, j is given by $X_i X_j^\top$, D is the block diagonal matrix formed by the block i, i of C and $\mathbf{a} \in \mathbb{R}^{m \times v}$ yields the dataset specific projections vectors: $\mathbf{a} = \begin{bmatrix} \mathbf{a}_1 \\ \dots \\ \mathbf{a}_m \end{bmatrix}$.

The leading eigenvector correspond to the first canonical vectors. The second canonical vectors is given by the second eigenvalues and so on. They are orthogonal for the scalar product: $\langle \mathbf{a}, \mathbf{b} \rangle_D = \langle \mathbf{a}, D\mathbf{b} \rangle$.

4.2.2 *Group independent component analysis*

Given the success of ICA in analyzing the data of one subject. It is natural to look for extensions of ICA in a multiview setting. Several works assume that the subjects share a common mixing matrix, but with different components [117] [142]. Instead, we focus on models where the subjects share a common components matrix, but have different mixing matrices.

4.2.2.1 *CanICA and ConcatICA*

In the single subject setting, we reduce the data (for example using PCA) and apply ICA on reduced data. Therefore a natural framework to perform group ICA is to first aggregate the data of individual subjects into a single dataset, often resorting to dimension reduction technique and then apply off-the-shelf ICA on the aggregated dataset. When PCA is used to aggregate the data, the method is referred to as ConcatICA [28]. An alternative is to use multiset canonical correlation analysis (CCA) leading to a method called CanICA [150].

This framework has the advantage of being simple and straightforward to implement since it resorts to customary single-subject ICA method.

When datasets are high-dimensional, a three steps procedure is often used: first dimensionality reduction is performed on data of each subject separately; then the reduced data are merged into a common representation; finally, an ICA algorithm is applied for shared components extraction.

CanICA and ConcatICA are popular methods for fMRI [29] and EEG [48] group studies. These methods directly recover only group level, shared components; when individual components are needed, a back-reconstruction step is required (such as GICA1 [28] GICA2, GICA3 [50] or dual-regression [17]).

In our experiments, when we fit CanICA or ConcatICA, we use Infomax and the picard solver [3] with the tanh non-linearity. To obtain k components, we first apply a subject specific PCA with k components and then aggregate the data using either a Group PCA (in ConcatICA) or a multiset CCA (in CanICA) with k components. When individual sources are needed, we use dual-regression. As already noted, many different procedure to perform single subject ICA exists (such as Infomax, EBM [89], ERBM [90]), the number of components in the first PCA can be chosen in many different ways and so can the back-reconstruction method. This leads to a number of different CanICA / ConcatICA algorithms. Many of these different versions of CanICA / ConcatICA are implemented in the GIFT toolbox `https://trendscenter.org/software/gift/`. We highlight that GIFT uses a slightly different terminology as us: dual-regression is referred to as spatial-temporal regression and ConcatICA is simply referred to as GroupICA. In this work, GroupICA describes instead the problem of recovering common sources from multiple subjects.

In the rest of the thesis, we often make the point that CanICA and ConcatICA are not maximum likelihood estimators. This is because the group PCA / multiset CCA used to fuse the data of different subjects does not come from a maximum likelihood approach.

4.2.2.2 *Likelihood based method*

While CanICA and ConcatICA are simple to implement and very fast to fit, they do not rely on maximum likelihood estimators. Therefore they do not benefit of advantages of such estimators such as asymptotic efficiency.

The model of [64] considers the very general model $\mathbf{x}_i = A_i \mathbf{s} + \mathbf{n}_i$ where A_i represent mixing matrices, $\mathbf{s}$ the common sources and $\mathbf{n}_i$ are view specific noise. Therefore this model is an instance of noisy ICA as defined in [75]. The noise covariance is learned from the data and each source is assumed to be a mixture of Gaussians $p(s_j) = \sum_{z=1}^{q} \mathcal{N}(\mu_{zj}, \sigma_{zj})$. The parameters of the Gaussian mixtures are learned which makes the E-step impossible to compute in closed form. In order to solve this issue, an approximate E-step is introduced. Unfortunately, it leads to an update rule involving a sum over q^p terms making their algorithm intractable when the number of components p is larger than 20.

In this section, we show why a sum of an exponential number of terms appears. We start with the same model as in [64] but to make the computations more tractable, we assume that the Gaussian mixture is given by:

$$p(s_j) = \frac{1}{q} \sum_{\alpha_j \in \mathcal{A}} p(s_j|\alpha_j) \tag{4.40}$$

$$p(s_j|\alpha_j) = \mathcal{N}(s_j; 0, \alpha_j), \tag{4.41}$$

where α_j takes its value in a known discrete set $\mathcal{A}$ with equal probability $\frac{1}{q}$ where q is the cardinal of $\mathcal{A}$. We call $\boldsymbol{\alpha}$ the random vector with independent coordinates such that coordinate j is given by α_j. We further assume that the noise distribution is the same for all components and all subjects leading to the formulation:

$$\mathbf{x} = A\mathbf{s} + \mathbf{n} \tag{4.42}$$

where $A = \begin{bmatrix} A_1 \\ \vdots \\ A_m \end{bmatrix}$, $\mathbf{x} = \begin{bmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_m \end{bmatrix}$ and $\mathbf{n} = \begin{bmatrix} \mathbf{n}_1 \\ \vdots \\ \mathbf{n}_m \end{bmatrix}$ with $\mathbf{n} \sim \mathcal{N}(0, \sigma^2 I)$. This formulation is a special case of [64] and constitues a single subject noisy ICA problem almost identical to [104].

We now follow [104] and write:

$$p(\mathbf{x}, \mathbf{s}, \boldsymbol{\alpha}) \tag{4.43}$$
$$= p(\mathbf{x}|\mathbf{s})p(\mathbf{s}|\boldsymbol{\alpha})p(\boldsymbol{\alpha}) \tag{4.44}$$
$$= \mathcal{N}(\mathbf{x}; A\mathbf{s}, \sigma^2 I_p)\mathcal{N}(\mathbf{s}; 0, \mathrm{diag}(\boldsymbol{\alpha}))\frac{1}{2^p} \tag{4.45}$$
$$\propto \frac{\exp\left(-\frac{1}{2}(\frac{1}{\sigma^2}\|\mathbf{x} - A\mathbf{s}\|^2 + \langle \mathbf{s}, \mathrm{diag}(\boldsymbol{\alpha})^{-1}\mathbf{s}\rangle)\right)}{(|\sigma^2 I_p\|\,\mathrm{diag}(\boldsymbol{\alpha})|)^{\frac{1}{2}}} \tag{4.46}$$
$$= \frac{\exp\left(-\frac{1}{2}(\frac{1}{\sigma^2}(\|\mathbf{x}\|^2 - 2\langle \mathbf{x}, A\mathbf{s}\rangle + \|A\mathbf{s}\|^2) + \langle \mathbf{s}, \mathrm{diag}(\boldsymbol{\alpha})^{-1}\mathbf{s}\rangle)\right)}{(|\sigma^2 I_p\|\,\mathrm{diag}(\boldsymbol{\alpha})|)^{\frac{1}{2}}} \tag{4.47}$$
$$\propto \frac{\exp\left(-\frac{1}{2}(\langle \mathbf{s} - \mu_{\boldsymbol{\alpha}}, V_{\boldsymbol{\alpha}}^{-1}(\mathbf{s} - \mu_{\boldsymbol{\alpha}})\rangle - \langle \mu_{\boldsymbol{\alpha}}, V_{\boldsymbol{\alpha}}^{-1}\mu_{\boldsymbol{\alpha}}\rangle)\right)}{(|\sigma^2 I_p\|\,\mathrm{diag}(\boldsymbol{\alpha})|)^{\frac{1}{2}}}, \tag{4.48}$$

where $V_{\boldsymbol{\alpha}} = (\frac{1}{\sigma^2}A^\top A + \mathrm{diag}(\boldsymbol{\alpha})^{-1})^{-1}$, $\mu_{\boldsymbol{\alpha}} = \frac{1}{\sigma^2}V_{\boldsymbol{\alpha}}A^\top\mathbf{x}$ and the proportionality constant contains terms that do not depend on $\mathbf{s}$ or $\boldsymbol{\alpha}$.

From $p(\mathbf{x}, \mathbf{s}, \boldsymbol{\alpha})$, we get:

$$p(\mathbf{s}|\mathbf{x}, \boldsymbol{\alpha}) = \mathcal{N}(\mathbf{s}, \mu_{\boldsymbol{\alpha}}, \Sigma_{\boldsymbol{\alpha}}) \tag{4.49}$$

Then, we have that:

$$p(\boldsymbol{\alpha}|\mathbf{x}) = \int_{\mathbf{s}} p(\boldsymbol{\alpha}, \mathbf{s}|\mathbf{x})d\mathbf{s} \tag{4.50}$$
$$\propto \int_{\mathbf{s}} p(\mathbf{x}, \mathbf{s}, \boldsymbol{\alpha})d\mathbf{s} \tag{4.51}$$
$$\propto \frac{\exp\left(\frac{1}{2}(\langle \mu_{\boldsymbol{\alpha}}, V_{\boldsymbol{\alpha}}^{-1}\mu_{\boldsymbol{\alpha}}\rangle)\right)}{(|\,\mathrm{diag}(\boldsymbol{\alpha})\|V_{\boldsymbol{\alpha}}^{-1}|)^{\frac{1}{2}}} \tag{4.52}$$
$$\tag{4.53}$$

where we leave out terms that do not depend on $\boldsymbol{\alpha}$. The normalizing constant can be computed by summing over possible values of $\boldsymbol{\alpha}$:

$$p(\boldsymbol{\alpha}|\mathbf{x}) = \frac{\frac{\exp\left(\frac{1}{2}(\langle \mu_{\boldsymbol{\alpha}}, V_{\boldsymbol{\alpha}}^{-1}\mu_{\boldsymbol{\alpha}}\rangle)\right)}{(|\,\mathrm{diag}(\boldsymbol{\alpha})\|V_{\boldsymbol{\alpha}}^{-1}|)^{\frac{1}{2}}}}{\sum_{\boldsymbol{\alpha}, \alpha_j \in \mathcal{A}} \frac{\exp\left(\frac{1}{2}(\langle \mu_{\boldsymbol{\alpha}}, V_{\boldsymbol{\alpha}}^{-1}\mu_{\boldsymbol{\alpha}}\rangle)\right)}{(|\,\mathrm{diag}(\boldsymbol{\alpha})\|V_{\boldsymbol{\alpha}}^{-1}|)^{\frac{1}{2}}}} \tag{4.54}$$

Then, we can obtain a formula in closed form for $p(\mathbf{s}|\mathbf{x})$ using

$$p(\mathbf{s}|\mathbf{x}) = \sum_{\boldsymbol{\alpha}, \alpha_j \in \mathcal{A}} p(\mathbf{s}|\mathbf{x}, \boldsymbol{\alpha})p(\boldsymbol{\alpha}|\mathbf{x}) \tag{4.55}$$

The problem here is that the size of the set $\{\boldsymbol{\alpha}, \alpha_j \in \mathcal{A}\}$ is q^p where $q = |\mathcal{A}|$ is the cardinal of $\mathcal{A}$. This quantity quickly gets large when p increases making $p(\mathbf{s}|\mathbf{x})$ difficult to compute.

In this section, we have studied a simplified version of the model in [64] and shown that it is difficult to fit. This is often the case with maximum likelihood approaches. Despite, their advantages they are often intractable.

4.2.3 *Independent vector analysis*

Independent vector analysis [85] (IVA) models the data as a linear mixture of independent components $\mathbf{x}_i = A_i \mathbf{s}_i$, where each component s_{ij} of a given view i can depend on the corresponding component in other views: $\mathbf{s}_{[j]} = (s_{ij})_{i=1}^m$ are not independent.

Introducing $\mathbf{x} \in \mathbb{R}^{m \times p}$ such that $\mathbf{x} = [\mathbf{x}_1, \dots, \mathbf{x}_m]^\top$, $\mathbf{s} \in \mathbb{R}^{m \times p}$ such that $\mathbf{s} = [\mathbf{s}_1, \dots, \mathbf{s}_m]^\top$ and $\mathbf{y} \in \mathbb{R}^{m \times p}$ such that $\mathbf{y} = [\mathbf{y}_1, \dots, \mathbf{y}_m]^\top$ where $\mathbf{y}_i = W_i \mathbf{x}_i$ with $W_i = A_i^{-1}$, the expected negative log-likelihood is given by:

$$\begin{aligned}
\mathcal{L} &= -\mathbb{E}[\log(p(\mathbf{x}))] \\
&= \sum_{i=1}^m -\log(|W_i|) - \mathbb{E}[\log(p(\mathbf{y}))] \\
&= \sum_{i=1}^m -\log(|W_i|) + \sum_{j=1}^p -\mathbb{E}[\log(p_{\mathbf{s}_{[j]}}(\mathbf{y}_{[j]}))]
\end{aligned}$$

where we used the notation $\mathbf{y}_{[j]} = (y_{ij})_{i=1}^m$.

The optimization can be carried out using alternate minimization keeping the mixing matrices of all subjects fixed but one. We can rely on the relative gradient as in section 4.1.2.3 and use update of the form $W_i \leftarrow (I - \alpha_i G_i) W_i$ where α_i is given by backtracking line search and G_i is the relative gradient given by:

$$G_i = -I_p + \mathbb{E}[\phi_i(\mathbf{y}_i)\mathbf{y}_i^\top] \tag{4.56}$$

where component j of $\phi_i(\mathbf{y}_i)$ is given by

$$\phi_{ij}(\mathbf{y}_i) = \frac{\partial - \log(p_{\mathbf{s}_{[j]}}(\mathbf{y}_{[j]}))}{\partial y_{ij}} \tag{4.57}$$

Practical implementations of this general model assume a distribution for $p_{\mathbf{s}_{[j]}}$. In IVA-L [85],

$$p_{\mathbf{s}_{[j]}}(\mathbf{y}_{[j]}) \propto \exp(-\sqrt{\sum_i (y_{ij})^2}) \tag{4.58}$$

and therefore

$$\phi_{ij}(\mathbf{y}_i) = \frac{y_{ij}}{\sqrt{\sum_i (y_{ij})^2}} \tag{4.59}$$

In IVA-G [7] [152],

$$p_{\mathbf{s}_{[j]}}(\mathbf{y}_{[j]}) = \mathcal{N}(\mathbf{y}_{[j]}; 0, \Sigma_j) \tag{4.60}$$

and therefore

$$\phi_{ij}(\mathbf{y}_i) = \sum_l \Sigma_j^{-1}[il] y_{lj} \tag{4.61}$$

where $\Sigma_j^{-1}[il]$ is the coordinate i, l of Σ_j^{-1} and y_{lj} is the jth coordinate of $\mathbf{y}_i$.

In IVA-G, an estimate of Σ_j is needed at each iteration. This is computed using the sample covariance:

$$\Sigma_j = \frac{1}{n} Y_{[j]} Y_{[j]}^\top \tag{4.62}$$

Second order extensions and Hessian approximations can be used in IVA as well. This is described in [7]. Also note that although IVA-G and IVA-L are the two most popular implementations of the IVA framework, others exist (see for instance the work in [8]).

4.2.4 *Hyperalignment*

Hyperalignment is a model initially designed for fMRI data to reduce inter-subject variability [69].

Let us assume we have access to the data of two subjects: $\mathbf{x}_1, \mathbf{x}_2$. Assuming these subjects are exposed to a time-locked stimuli (such as a movie), a possible alignment is given by the Procrustes transform:

$$\min_{P \in \mathbb{R}^{p \times p}, PP^\top = I_p} \mathbb{E}[\|P\mathbf{x}_1 - \mathbf{x}_2\|_2] \tag{4.63}$$

This can be solved efficiently by

$$P = \mathcal{P}(\mathbb{E}[\mathbf{x}_2 \mathbf{x}_1^\top]) \tag{4.64}$$

where $\mathcal{P}$ is the projection on the orthogonal manifold: $\mathcal{P}(M) = M(M^\top M)^{-\frac{1}{2}}$. In practice $\mathcal{P}(M)$ is computed by performing an SVD of M, $M = U_M D_M V_M$ so that $\mathcal{P}(M) = U_M V_M$.

Hyperalignment is the combination of the Procrustes transform and an iterative procedure to produce a template from multiple alignments. In an initialization step, a random subject i is chosen and the alignment between all subjects $s \neq i$ and the target are computed. The initial template $\mathbf{t}$ is given by the averaged aligned data. Then, all subjects are aligned to the current template $\mathbf{t}$ and the template is recomputed using the averaged aligned data. This procedure is repeated for a given number of iterations until convergence.

The intuition behind the iterative procedure is that averaging the aligned data will tend to move the template away from the initial target. However, they are no theoretical guarantees associated with this procedure and even no associated loss. We describe formally the method in Algorithm 1.

Algorithm 1: Hyperalignment

Input: Data $\mathbf{X}_1, \dots, \mathbf{X}_m \in \mathbb{R}^{p \times n}$, number of iterations n_{iter}
▶ Select a random subject
$i \sim \mathcal{U}(1, m)$
▶ Initialize the alignment operators
for $s = 1 \dots m$ **do**
  if $s = i$ **then**
    $P_{st} = I_p$
  end
  else
    $P_{st} = \mathcal{P}(X_t X_s^\top)$
  end
end
$T = \frac{\sum_{s=1}^m P_{st} X_s}{m}$
▶ Main loop
for $it = 1 \dots n_{iter}$ **do**
  ▶ Align data and the current template
  for $s = 1 \dots m$ **do**
    $P_{st} = \mathcal{P}(T X_s^\top)$
  end
  ▶ Compute the template as the mean of aligned data
  $T = \frac{\sum_{s=1}^m P_{st} X_s}{m}$
end
return *Estimated template* T *and operators* P_{st}

4.2.5 *The shared response model (SRM)*

The shared response model [36] is a multi-view latent factor model. The data $\mathbf{x}_1 \dots \mathbf{x}_m$ are modeled as random vectors following the model:

$$\mathbf{x}_i = A_i \mathbf{s} + \mathbf{n}_i \tag{4.65}$$

$$A_i^\top A_i = I_p \tag{4.66}$$

where $\mathbf{x}_i \in \mathbb{R}^v$ is the data of view i, $A_i \in \mathbb{R}^{p \times v}$ is the mixing matrix of view i, $\mathbf{n}_i$ is the noise of view i and $\mathbf{s} \in \mathbb{R}^p$ are the shared components referred to as the *shared response* in fMRI applications. The mixing matrices A_i are assumed to be orthogonal so that $A_i^\top A_i = I_p$. However, in general the matrix $A_i A_i^\top$ is different from identity. The noise $\mathbf{n}_i$ is assumed to be Gaussian with covariance Σ_i and independent across views. We assume the number of features v to be much larger than the number of components p: $v \gg p$.

The conceptual figure 4.1 illustrates an application of the shared response model to fMRI data. The mixing matrices are spatial topogra-

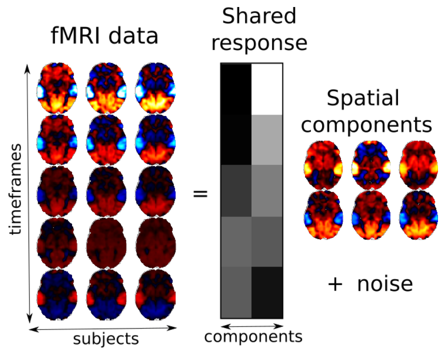


Figure 4.1: **Shared response model**: The raw fMRI data are modeled as a weighted combination of subject-specific spatial components with additive noise. The weights are shared between subjects and constitute the shared response to the stimuli.

phies specific to each subjects while the shared components give the common timecourses.

In [10, 36], two versions of the shared response model are introduced which we now present.

4.2.5.1 *Deterministic shared response model*

Let us consider n observations of $\mathbf{x}_i$ and $\mathbf{s}$ that we stack into matrices $X_i \in \mathbb{R}^{v,n}$ and $S \in \mathbb{R}^{p,n}$. The deterministic shared response model sees both the mixing matrices A_i and the n observations of the shared response S as parameters to be estimated. The noise variance is fixed to a multiple of identity: $\forall i, \Sigma_i = \sigma^2 I_v$ where σ is an hyper-parameter to choose. The model is optimized by maximizing the log-likelihood. The likelihood is given by: $p(\mathbf{x}) = \prod_i \mathcal{N}(\mathbf{x}_i; A_i \mathbf{s}, \sigma^2 I)$ and therefore the empirical expected negative log-likelihood is given up to a constant independent of A_i and S by:

$$\mathcal{L} = \frac{1}{n} \sum_i \|A_i S - X_i\|^2 = \frac{1}{n} (\|S\|^2 - 2\langle A_i S, X_i \rangle + \|X_i\|^2) \quad (4.67)$$

The negative log-likelihood $\mathcal{L}$ is optimized by performing alternate minimization on $(A_1 \dots A_m)$ and S. Note that the hyper-parameter σ does not have an influence on the loss and can therefore be safely ignored.

The gradient with respect to S is given by $\sum_i A_i^\top (A_i S - X_i) = \sum_i (S - A_i^\top X_i)$ yielding the closed form updates:

$$S \leftarrow \frac{1}{m} \sum_i (A_i^\top X_i) \quad (4.68)$$

From (4.67), minimizing $\mathcal{L}$ with respect to A_i is equivalent to maximizing $\langle A_i, X_i S^\top \rangle$ and therefore we have:

$$A_i \leftarrow \mathcal{P}(\frac{1}{n} X_i S^\top) \quad (4.69)$$

where $\mathcal{P}$ is the projection on the Stiefel manifold: $\mathcal{P}(M) = M(M^\top M)^{-\frac{1}{2}}$.

The complexity of Deterministic SRM is in $\tilde{O}(n_{\text{iter}} m p v n)$ where n is the number of samples and n_{iter} the number of iterations. We monitor the convergence by looking at the ℓ_∞ norm of the gradient. Note that we can monitor the gradient without any increase in complexity. Indeed, after the updates with respect to each mixing matrix have been carried out, only the gradient with respect to S remains: $\sum_i (S - A_i^\top \mathbf{x}_i)$. The algorithm is stopped when the gradient falls below a chosen tolerance.

4.2.5.2 *Probabilistic SRM*

In Probabilistic SRM , $\Sigma_i = \sigma_i^2 I_v$ and the shared components are assumed to be Gaussian $\mathbf{s} \sim \mathcal{N}(0, \Sigma_s)$. In [36], Σ_s is only assumed to be definite positive. However, as will be seen in the FastSRM chapter (chapter 5), enforcing a diagonal Σ_s ensures identifiability (provided the diagonal values are different). So we assume that Σ_s is diagonal.

The model is optimized via the expectation maximization algorithm. Denoting $\mathbb{V}[\mathbf{s}|\mathbf{x}] = (\sum_i \frac{1}{\sigma_i^2} I + \Sigma_s^{-1})^{-1}$ and $\mathbb{E}[\mathbf{s}|\mathbf{x}] = \mathbb{V}[\mathbf{s}|\mathbf{x}] \sum_i \frac{1}{\sigma_i^2} A_i^\top \mathbf{x}_i$, we have

$$p(\mathbf{x}, \mathbf{s}) = \prod_i \frac{\exp(-\frac{\|\mathbf{x}_i - A_i \mathbf{s}\|^2}{2\sigma_i^2})}{(2\pi\sigma_i^2 v)^{\frac{1}{2}}} \frac{\exp(-\frac{1}{2}\langle \mathbf{s}, \Sigma_s^{-1}\mathbf{s}\rangle)}{(2\pi|\Sigma_s|)^{\frac{1}{2}}} \tag{4.70}$$

$$= c_1 \exp(-\frac{1}{2}\left(\sum_i \frac{1}{\sigma_i^2}\|\mathbf{x}_i\|^2 - 2\langle \sum_i \frac{1}{\sigma_i^2} A_i^\top \mathbf{x}_i, \mathbf{s}\rangle\right. \tag{4.71}$$

$$\left. + \sum_i \frac{1}{\sigma_i^2}\|\mathbf{s}\|^2 + \langle \mathbf{s}, \Sigma_s^{-1}\mathbf{s}\rangle\right)) \tag{4.72}$$

$$= c_2(\mathbf{x}) \exp(-\frac{1}{2}\left(\langle \mathbf{s} - \mathbb{E}[\mathbf{s}, \mathbf{x}], \mathbb{V}[\mathbf{s}|\mathbf{x}]^{-1}(\mathbf{s} - \mathbb{E}[\mathbf{s}|\mathbf{x}])\rangle\right)) \tag{4.73}$$

where $c_1 = \frac{1}{(2\pi\sigma_i^2 v)^{\frac{1}{2}}} \frac{1}{(2\pi|\Sigma_s|)^{\frac{1}{2}}}$ and $c_2(\mathbf{x}) = c_1 \exp(-\frac{1}{2}(\sum_i \frac{1}{\sigma_i^2}\|\mathbf{x}_i\|^2 - \langle \mathbb{E}[\mathbf{s}, \mathbf{x}], \mathbb{V}[\mathbf{s}|\mathbf{x}]^{-1}\mathbb{E}[\mathbf{s}|\mathbf{x}]\rangle))$ are independent of $\mathbf{s}$. Therefore $\mathbf{s}|\mathbf{x} \sim \mathcal{N}(\mathbb{E}[\mathbf{s}|\mathbf{x}], \mathbb{V}[\mathbf{s}, \mathbf{x}])$

The negative expected completed log-likelihood is given by

$$\mathcal{L} = \sum_i \frac{1}{2} v \log(\sigma_i^2) + \frac{1}{2\sigma_i^2}\mathbb{E}[\|\mathbf{x}_i - A_i \mathbf{s}\|^2] \tag{4.74}$$

updates are therefore given by:

$$\sigma_i^2 \leftarrow \frac{1}{v}(\mathbb{E}[\|\mathbf{x}_i - A_i \mathbb{E}[\mathbf{s}|\mathbf{x}]\|^2] + \|\operatorname{diag}(\mathbb{V}[\mathbf{s}|\mathbf{x}])\|^2) \tag{4.75}$$

$$A_i \leftarrow \mathcal{P}(\mathbb{E}[\mathbf{x}_i \mathbb{E}[\mathbf{s}|\mathbf{x}]^\top]) \tag{4.76}$$

$$\Sigma_s \leftarrow \mathbb{V}[\mathbf{s}|\mathbf{x}] + \mathbb{E}[\mathbb{E}[\mathbf{s}|\mathbf{x}]\mathbb{E}[\mathbf{s}|\mathbf{x}]^\top] \tag{4.77}$$

It is useful to access the log-likelihood to check the implementation of the algorithm and monitor the convergence. From equation (4.73), the likelihood is given by:

$$p(\mathbf{x}) = c_2(\mathbf{x}) \int_{\mathbf{s}} \exp(-\frac{1}{2}\left(\langle \mathbf{s} - \mathbb{E}[\mathbf{s}, \mathbf{x}], \mathbb{V}[\mathbf{s}|\mathbf{x}]^{-1}(\mathbf{s} - \mathbb{E}[\mathbf{s}|\mathbf{x}])\rangle\right)) d\mathbf{s} \tag{4.78}$$

$$= c_2(\mathbf{x})(2\pi|\mathbb{V}[\mathbf{s}|\mathbf{x}]|)^{\frac{1}{2}} \tag{4.79}$$

replacing $c_2(\mathbf{x})$ by its expression and taking the log, the expected negative log-likelihood is (up to constants) given by:

$$\begin{aligned}\mathbb{E}[-\log(p(\mathbf{x}))] = & \sum_i \frac{v}{2}\log(\sigma_i^2) + \frac{1}{2}\log(|\Sigma_s|) - \frac{1}{2}\log(|\mathbb{V}[\mathbf{s}|\mathbf{x}]|) \\ & + \sum_i \frac{1}{2}\frac{1}{\sigma_i^2}\mathbb{E}[\|\mathbf{x}_i\|^2] - \frac{1}{2}\mathbb{E}[\langle \mathbb{E}[\mathbf{s}, \mathbf{x}], \mathbb{V}[\mathbf{s}|\mathbf{x}]^{-1}\mathbb{E}[\mathbf{s}|\mathbf{x}]\rangle]\end{aligned} \tag{4.80}$$

The complexity of Probabilistic SRM is $\tilde{O}(n_{\text{iter}}mpvn)$, the same as in Deterministic SRM. We can monitor the convergence by looking at the log-likelihood decrease at each iteration and stop the algorithm when the magnitude of the decrease is below some tolerance. The storage requirements of Deterministic or Probabilistic SRM are in $\tilde{O}(mvn)$ which simply means that the dataset needs to hold in memory.

4.3 CONCLUSION

In this chapter, we have reviewed several methods to perform unsupervised analysis of neuroimaging data. We have introduced methods most suited to the analysis of the data of a single subject such as ICA and have explored some of the extensions to multiple subjects such as CanICA, ConcatICA, IVA or SRM. In the next chapter, we present our first contribution: an efficient implementation of SRM that we call FastSRM.

Part II

FASTSRM: AN EFFICIENT IMPLEMENTATION OF THE SHARED RESPONSE MODEL

5 FASTSRM THEORY

In the previous chapter, we have decribed multiple unsupervised methods to analyze multiview data. As described in section 4.2.5, the shared response model [36] (SRM) is a multi-view latent factor model. It sees the data $(\mathbf{x}_i)_{i=1}^m$ as:

$$\mathbf{x}_i = A_i \mathbf{s} + \mathbf{n}_i \tag{5.1}$$

$$A_i^\top A_i = I_p \tag{5.2}$$

with $\mathbf{n}_i \sim \mathcal{N}(0, \Sigma_i)$ where $\Sigma_i = \sigma^2 I_v$ in deterministic SRM and $\Sigma_i = \sigma_i^2 I_v$ in probabilistic SRM.

In practice, we have access to n observations of $\mathbf{x}_i$ stacked in a matrix $X_i \in \mathbb{R}^{v \times n}$. The corresponding observations of $\mathbf{s}$ called $S \in \mathbb{R}^{k \times n}$ are unknown.

When working with high dimensional data, SRM is particularly interesting as it provides a principled way to perform dimension reduction. Note that this contrasts with ICA-like methods that do not incorporate dimension reduction in their model.

However SRM has initially been designed to work within regions of interest using only few subjects. When using full brain data, computational costs become high. In addition, memory requirements are difficult to meet since the full dataset needs to hold in memory.

Fortunately, these high costs can be reduced. Intuitively, since the shared response lives in a reduced space, a compressed representation of the input is good enough to find a suitable estimate of the shared response. FastSRM implements this idea. It turns out that there exists an optimal compression of the input for which we can obtain the same solution as with the full data.

5.1 THE FASTSRM ALGORITHM

SRM algorithms use different set of parameters θ to represent the data. In deterministic SRM $\theta = (A_i)_{i=1}^m, S$ where $(A_i)_{i=1}^m$ are the mixing matrices and S are the n observations of the shared response S while in probabilistic SRM $\theta = (A_i)_{i=1}^m, \Sigma_s, (\sigma_i)_{i=1}^m$ where $(A_i)_{i=1}^m$ are the mixing matrices, $(\sigma_i)_{i=1}^m$ the noise levels and Σ_s the components variance.

In fMRI, the classical approach used to reduce the data is to apply an atlas. A deterministic atlas such as [20] is a parcellation of the brain into r regions. Reducing an image using a deterministic atlas corresponds to averaging the signal within each region of the atlas. A probabilistic atlases such as [41] describes each region as a set of

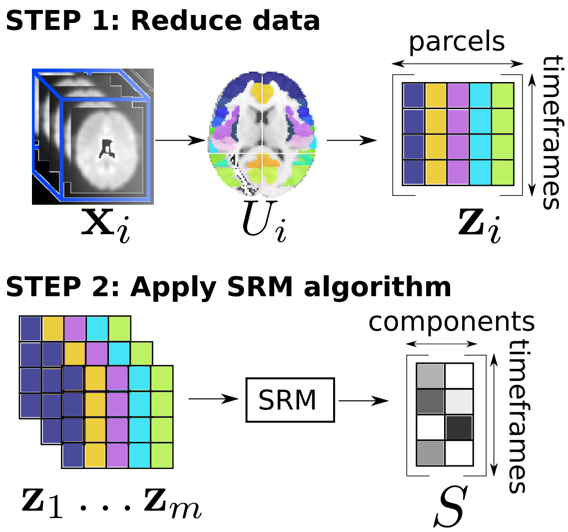


Figure 5.1: **FastSRM algorithm** In step 1, data x_i are projected onto an atlas U_i that may depend on the subject i (top). In step 2 a SRM algorithm is applied on reduced data to compute the shared response.

weights across the full brain. Therefore, the image reduction can be done with a matrix product.

In FastSRM we consider a set of view specific atlases $U_i \in \mathbb{R}^{v \times r}$ such that $U_i^\top U_i = I_r$ where r is the number of regions in the atlas. Data are reduced using $z_i = U_i^\top x_i$ and an SRM algorithm is applied on data z_i yielding parameters θ'. The figure 5.1 illustrates this process.

Note that the parameters obtained with FastSRM θ' are different from the parameters obtained with the corresponding SRM algorithm θ (the unmixing matrices in θ' do not even have the same shape as the unmixing matrices in θ). However, as we will see in the next section, there exists atlases such as the correspondence between θ and θ' can be made explicit.

From a computational stand point, the dimension reduction provides a large reduction in memory usage. Indeed as the original data are seen only once, it is no longer necessary to keep the full dataset in memory (we can load data X_i one after the other and similarly for the atlases U_i). Therefore the memory consumption is only in $\tilde{O}(vn)$ (where v is the number of voxels and n is the number of samples) which is lower than SRM by a factor of m, the number of subjects. The number of subjects is typically between 10 and 1000. This yields a practical benefit: on fMRI datasets with many subjects, one no longer needs a large cluster to run the shared response model but only a modern laptop. Additionally, low memory consumption reduces the risk of thrashing [45], a phenomenon that causes large increase in computation time when the memory used is close to the total available memory in the hardware.

After preprocessing, the reduced representation z_i is used instead of the original data x_i yielding a time complexity of $\tilde{O}(T_{\text{preprocessing}} +$

$n_{iter}mpnr$). Let us highlight that an experiment is often run multiple times such as when cross validated results are needed. In these cases, the pre-processing is performed only once and the apparent complexity becomes $\tilde{O}(n_{iter}mpnr)$ which is faster than SRM by a factor of $\frac{v}{r}$. The number of regions in large atlases is about $r = 1000$ and in full brain data, the number of voxels is about 300 000 so that $\frac{v}{r}$ is typically about 1000.

It remains to show how to draw a correspondence between FastSRM and SRM which is addressed in the following section.

5.2 OPTIMAL ATLASES

In principle, FastSRM can be used with any atlas. However, in general, working with reduced data induces a loss of information and therefore there is little hope to recover the parameters that would have been obtained from SRM from the parameters of FastSRM. However, in this section, we show that there exists an optimal atlas in the sense that SRM and FastSRM yield the same results.

Let us consider $\mathbf{x}_i = U_{\mathbf{x}_i}\mathbf{z}_i$ a PCA of $\mathbf{x}_i$ with the maximum number of components. As the number of samples n is lower than the number of features, $U_{\mathbf{x}_i} \in \mathbb{R}^{v \times n}$ and $\mathbf{z}_i \in \mathbb{R}^n$. We also have $U_{\mathbf{x}_i}^\top U_{\mathbf{x}_i} = I$. Therefore $U_{\mathbf{x}_i}$ constitutes a possible choice of subject specific atlas.

As the next property shows, $U_{\mathbf{x}_i}$ is an optimal atlas for deterministic FastSRM.

Proposition 9 (Optimal atlas for deterministic FastSRM). *Let $(A_i)_i, S$ be the solution obtained by deterministic SRM on data $(X_i)_i$ and $(A_i')_i, S'$ the solution obtained by deterministic FastSRM on data $(X_i)_i$ using atlases $(U_{X_i})_i$ where $X_i = U_{\mathbf{x}_i}Z_i$ is a PCA of X_i. Then $A_i = U_{X_i}A_i'$ and $S = S'$.*

Proof. Updates of the mixing matrices A_i in deterministic SRM equation (4.69) can be written:

$$A_i \leftarrow \mathcal{P}(\frac{1}{n}X_iS^\top) = U_{X_i}\mathcal{P}(\frac{1}{n}Z_iS^\top) \tag{5.3}$$

where $\mathcal{P}$ is the projection on the Stiefel manifold: $\mathcal{P}(M) = M(M^\top M)^{-\frac{1}{2}}$.

Therefore we can look for A_i as $A_i = U_{X_i}\tilde{A}_i$. $\tilde{A}_i$ is orthogonal. Indeed

$$A_i^\top A_i = I_p \tag{5.4}$$

$$\implies \tilde{A}_i^\top U_{X_i}^\top U_{X_i}\tilde{A}_i = I_p \tag{5.5}$$

$$\implies \tilde{A}_i^\top \tilde{A}_i = I_p \tag{5.6}$$

Then, we use the fact that

$$\|X_i - A_iS\|^2 = \|U_{X_i}Z_i - U_{X_i}\tilde{A}_iS\|^2 = \|Z_i - \tilde{A}_iS\|^2 \tag{5.7}$$

so that $A_i' = \tilde{A}_i$.

Therefore, the solution of deterministic SRM on data $(z_i)_{i=1}^m$ and $(x_i)_{i=1}^m$ are linked by the change of parameters $A_i = U_{x_i} A_i'$ and $S = S'$. This concludes the proof. □

In the case of probabilistic SRM we can obtain very similar results. However the algorithm applied on reduced data need to be slightly modified. We call probSRM(ψ) the probabilistic SRM algorithm modified such that updates

$$\sigma_i^2 \leftarrow \frac{1}{v}(\mathbb{E}[\|x_i - A_i \mathbb{E}[s|x]\|^2] + \|\operatorname{diag}(\mathbb{V}[s|x])\|^2) \tag{5.8}$$

are replaced by updates

$$\sigma_i^2 \leftarrow \frac{1}{\psi}(\mathbb{E}[\|x_i - A_i \mathbb{E}[s|x]\|^2] + \|\operatorname{diag}(\mathbb{V}[s|x])\|^2) \tag{5.9}$$

We have the following result:

Proposition 10 (Optimal atlas for probabilistic FastSRM). *Let* $(A_i)_i, \sigma_i, \Sigma_s$ *be the solution obtained by probabilistic SRM on data* x_i *and* $(A_i')_i, \sigma_i', \Sigma_s'$ *the solution obtained by ProbSRM(v) on data* $z_i = U_{x_i}^\top x_i$*.Then* $A_i = U_{x_i} A_i'$, $\sigma_i = \sigma_i'$ *and* $\Sigma_s = \Sigma_s'$.

Proof. Updates of the mixing matrices A_i in probabilistic SRM equation (4.76) can be written:

$$A_i \leftarrow U_{x_i} \mathcal{P}(\mathbb{E}[z_i \mathbb{E}[s|x_i]^\top]) \tag{5.10}$$

so we can look for A_i as $A_i = U_{x_i} \tilde{A}_i$ and, as in the deterministic case, $\tilde{A}_i$ is orthogonal. Therefore equality (5.7) holds.

Then we consider the expected negative log-likelihood of probabilistic srm:

$$\begin{aligned}
\mathcal{L} &= \sum_i \frac{1}{2} v \log(\sigma_i^2) + \frac{1}{2}\log(|\Sigma_s|) + \mathbb{E}[\int_s \sum_i \frac{1}{2\sigma_i^2}\|x_i - A_i s\|^2 \\
&\quad + \frac{1}{2}\langle s, \Sigma_s^{-1} s\rangle ds] && (5.11)\\
&= \sum_i \frac{1}{2} v \log(\sigma_i^2) + \frac{1}{2}\log(|\Sigma_s|) + \mathbb{E}[\int_s \sum_i \frac{1}{2\sigma_i^2}\|z_i - \tilde{A}_i s\|^2 \\
&\quad + \frac{1}{2}\langle s, \Sigma_s^{-1} s\rangle ds] && (5.12)
\end{aligned}$$

where we use equality (5.7). Optimizing the log-likelihood via expectation maximization yields the exact same updates as probabilistic srm on data z_i except that updates

$$\sigma_i^2 \leftarrow \frac{1}{t}(\mathbb{E}[\|z_i - \tilde{A}_i \mathbb{E}[s|z]\|^2] + \|\operatorname{diag}(\mathbb{V}[s|z])\|^2) \tag{5.13}$$

are replaced by updates

$$\sigma_i^2 \leftarrow \frac{1}{v}(\mathbb{E}[\|z_i - \tilde{A}_i \mathbb{E}[\mathbf{s}|z]\|^2] + \|\operatorname{diag}(\mathbb{V}[\mathbf{s}|z])\|^2) \tag{5.14}$$

so that $\tilde{A}_i = A_i'$.

Therefore, the updates in both algorithms are linked by $A_i = U_{x_i} A_i'$, $\sigma_i' = \sigma_i$ and $\Sigma_s' = \Sigma_s$.

This concludes the proof. □

The properties 9 and 10 show that no information is lost by replacing $x_i \in \mathbb{R}^v$ by its reduced representation $z_i \in \mathbb{R}^n$. A key property of the optimal atlas U_{x_i} is that it is valid whether or not the model for deterministic (respectively probabilistic) SRM holds.

A complexity analysis shows that finding the optimal atlas becomes the limiting step of the pipeline. Even with fast implementations, the subject specific PCA is costly. However FastSRM only works on z_i so we do not need to know the value of U_{x_i}. In practice, we observe data $X_i \in \mathbb{R}^{v \times n}$ and we want to get $Z_i \in \mathbb{R}^{n \times n}$ such that $X_i = U_{x_i} Z_i$. This can be done by performing an SVD of $X_i^\top X_i$ yielding $X_i^\top X_i = V_i D_i V_i^\top$ and setting $Z_i = D_i^{\frac{1}{2}} V_i^\top$. Although computing the product $X_i^\top X_i$ has time complexity $\tilde{O}(vt^2)$ there is exactly vt^2 multiplications and additions so it costs a lot less than the PCA on full data. When estimates of the mixing matrices are needed, they can be obtained by applying equation (5.3) in the deterministic SRM case and equation (5.10) in the probabilistic SRM case which only costs $\tilde{O}(mvp^2)$. In practice the cost of the matrix products $X_i^\top X_i$ is often still the limiting step of the pipeline (this depends on the number of iterations) but as we show in the next chapter, it is much more efficient than performing SRM on the full data. Note than if memory allows it, these matrix products can be computed in parallel.

Up to now, we have assumed that the covariance of components is diagonal in probabilistic SRM. In the next section we justify this assumption.

5.3 IDENTIFIABLITY OF THE SHARED RESPONSE MODEL

We first show why deterministic SRM and probabilistic SRM without any restrictions on the covariance of the components can only recover unmixing matrices up to an arbitrary rotation.

Let us consider data X_i generated from deterministic SRM (meaning deterministic SRM holds exactly for these data) with mixing matrices A_i and shared response S. Then deterministic SRM with parameters $A_i' = A_i R$ and $S' = R^\top$,where $R \in \mathbb{R}^{k \times k}$ is an orthogonal matrix, also generates X_i. This shows that deterministic SRM is not identifiable.

Similarly, in the probabilistic SRM case, if data x_i are generated from the probabilistic SRM model with parameters A_i, Σ_s, σ_i^2 (where Σ_s

can be any symmetric positive definite matrix) then the probabilistic SRM model with parameters $A_i R$, $R^\top \Sigma_s R$, σ_i^2 where $R \in \mathbb{R}^{k \times k}$ is an orthogonal matrix also generates $\mathbf{x}_i$. This shows that if no constraints are imposed on Σ_s, then probabilistic SRM is not identifiable.

In Proposition 11, we show that if Σ_s is assumed to be diagonal, then, under weak assumptions, probabilistic SRM is identifiable.

Proposition 11 (Identifiability of probabilistic SRM). *Let $\mathbf{x}_i$ be generated from a probabilistic SRM with parameters A_i, Σ_s, σ_i^2 (where Σ_s is diagonal positive with distinct values on the diagonal) and assume there exists another set of parameters A_i', Σ_s', ${\sigma_i'}^2$ (where Σ_s' is diagonal positive with distinct values on the diagonal) that also generate $\mathbf{x}_i$. Then if $m \geqslant 3$, $A_i' = A_i P^\top$, $\Sigma_s' = P \Sigma_s P^\top$ and ${\sigma_i^2}' = {\sigma_i}^2$ where P is a sign and permutation matrix (meaning P is the product of a diagonal matrix with values in $\{-1, 1\}$ and a permutation matrix).*

Proof. Let us consider $\mathbb{E}[\mathbf{x}_i \mathbf{x}_j^\top]$ for $i \neq j$. We have:

$$A_i \Sigma_s A_j^\top = A_i' \Sigma_s' {A_j'}^\top \tag{5.15}$$

up to re-ordering, equation (5.15) gives two singular value decompositions of the same matrix. Therefore, by unicity of the singular value decomposition we have: $\Sigma_s' = P_i \Sigma_s P_j^\top$ and $A_i' = A_i P_i^\top$ and $A_j = A_j P_j^\top$ where P_i and P_j are sign and permutation matrices. Since there are more than three subjects, there exists subject z such that $\Sigma_s' = P_i \Sigma_s P_z^\top$ and therefore $P_i \Sigma_s P_z^\top = P_i \Sigma_s P_j^\top$ so that $P_j = P_z$. So all sign and permutations are the same and we call P their common value. Then we consider $\mathbb{E}[\mathbf{x}_i \mathbf{x}_j^\top] = A_i \Sigma A_i^\top + \sigma^2 I_v = A_i' \Sigma_s' {A_i'}^\top + {\sigma^2}' I_v$ so we get ${\sigma^2}' = \sigma^2$ □

Proposition (11) justifies that Σ_s should be assumed to be diagonal in probabilistic SRM. In addition, working with a diagonal covariance matrix allows to speed up slightly the computations (although this does not change the time complexity of the algorithm).

5.4 CONCLUSION

In this chapter, we have presented FastSRM, an efficient implementation of SRM that uses optimal atlases to speed up computations and reduce memory requirements without loss of performance. We have also discussed the identifiability of SRM. In the next chapter, we measure the performance of FastSRM in practice and compare it to available implementations.

6 FASTSRM EXPERIMENTS

In the previous chapter we introduced the FastSRM algorithm, a framework to efficiently compute shared responses from the full brain data of multiple subjects. In this chapter, we investigate the practical benefits of FastSRM on synthetic and real fMRI data. FastSRM reduces the fitting time and memory requirements considerably, making it possible to compute shared responses using a laptop in a reasonable amount of time even on datasets that do not hold in RAM. The efficiency of FastSRM allows us to apply SRM on a large number of subjects. As an example, we apply FastSRM on 647 subjects from the CamCAN dataset and study how the mixing matrices of the SRM model are predictive of age.

6.1 COMPARING FITTING TIME AND PERFORMANCE OF FASTSRM AND SRM ON SYNTHETIC DATA

We generate synthetic data x_i according to the model of Probabilistic SRM. The parameters σ_i, A_i and Σ_s are generated randomly. We sample the value of the subject specific noise level from a normal distribution: $\sigma_i \sim \mathcal{N}(0, 0.1)$. The mixing matrices A_i are obtained by sampling their coefficient from a standardized normal distribution. Lastly, the covariance of the shared response Σ_s is diagonal and the diagonal values are obtained by sampling from a Dirichlet distribution with parameter $(1 \dots 1)$. We set the number of voxels to $v = 125\ 000$, the number of subjects to $m = 10$ and the number of components to $p = 50$. We generate $n = 1000$ samples.

We benchmark deterministic SRM, probabilistic SRM and their FastSRM counterparts in terms of fitting time and performance. Algorithms are designated by the atlas they use and therefore SRM algorithms described in section 4.2.5.2 and 4.2.5.1 are refered to as *None* because no atlas is used and FastSRM algorithms will have the label *Optimal*. Note that it would be possible to use FastSRM with sub-optimal atlases (there exists a wide variety of atlases available [20, 101, 136]) but without any guarantees that the performance are the same as SRM.

We use a number of iterations between 1 and 100 and report the performance, fitting time and a measure of convergence. In FastSRM, we do not compute the unmixing matrices but only the shared response. We measure the performance of an algorithm by computing the error between the true component $S \in \mathbb{R}^{p \times n}$ and the predicted component $\hat{S} \in \mathbb{R}^{p \times n}$ using the quantity: $\text{mse}(\hat{S}, S) = \min_{A \in \mathbb{R}^{p \times p}} \|A\hat{S} - S\|_F^2 =$

$\|S\hat{S}^\dagger\hat{S} - S\|_F^2$. This way of measuring errors is insensitive to the indeterminacies in DetSRM. We measure the fitting time in seconds. Lastly, we measure convergence by computing the gradient ℓ_∞ norm in case of DetSRM given by $\max(\text{abs}(G))$ where G is the gradient and use the distance between consecutive values of the loss for ProbSRM.

Results are plotted in figure 6.1. We empirically see that the optimal approach is equivalent to using no atlas in terms of performance. This is predicted by the theory in the previous chapter where we demonstrate that these two algorithms yield exactly the same output from the same input. In general probabilistic methods give much better results than their deterministic counterpart. This shows the superiority of likelihood based methods. In terms of fitting time, FastSRM is about a thousand time faster than SRM after 100 iterations. When no atlas is used, the number of iterations has a very strong impact on performance while it has a small impact when the optimal atlas is used. Lastly, looking at the convergence curves, we see that even after 100 iterations, algorithms did not fully converge. This means that in practice a much larger number of iterations is needed which would yield an even higher difference in fitting time between methods using no atlas and methods using the optimal atlas.

6.2 EXPERIMENTS ON FMRI DATA

We evaluate the performance of FastSRM on three fMRI datasets of subjects exposed to naturalistic stimuli: Sherlock, Forrest and Cam-CAN (more details about these datasets are available in section 3.1.5).

6.2.1 *Comparing fitting time, memory usage and performance on a timesegment matching experiment*

The timesegment matching experiment is first introduced in [36]. In a nutshell, the time-segment matching accuracy measures the similarity between two multivariate time-series by trying to localize a time-segment in one time-series by correlation with the other. In the context of movie watching, this measure has a lot of sense: if we split the movies in scenes and compute a representation per scene and per subject, it makes sense to assume that different subjects watching the movie would still have closer representation of the same scenes than of different scenes. This explains why timesegment matching is a standard evaluation of SRM-like methods also used in [63], [105] or [157].

We now describe more precisely the experimental design. We split the runs into a train and test set. After fitting the model on the training set, we apply the unmixing matrices $W_i = A_i^{-1}$ of each subject on the test set yielding individual components matrices. We estimate the shared responses by averaging the individual compo-

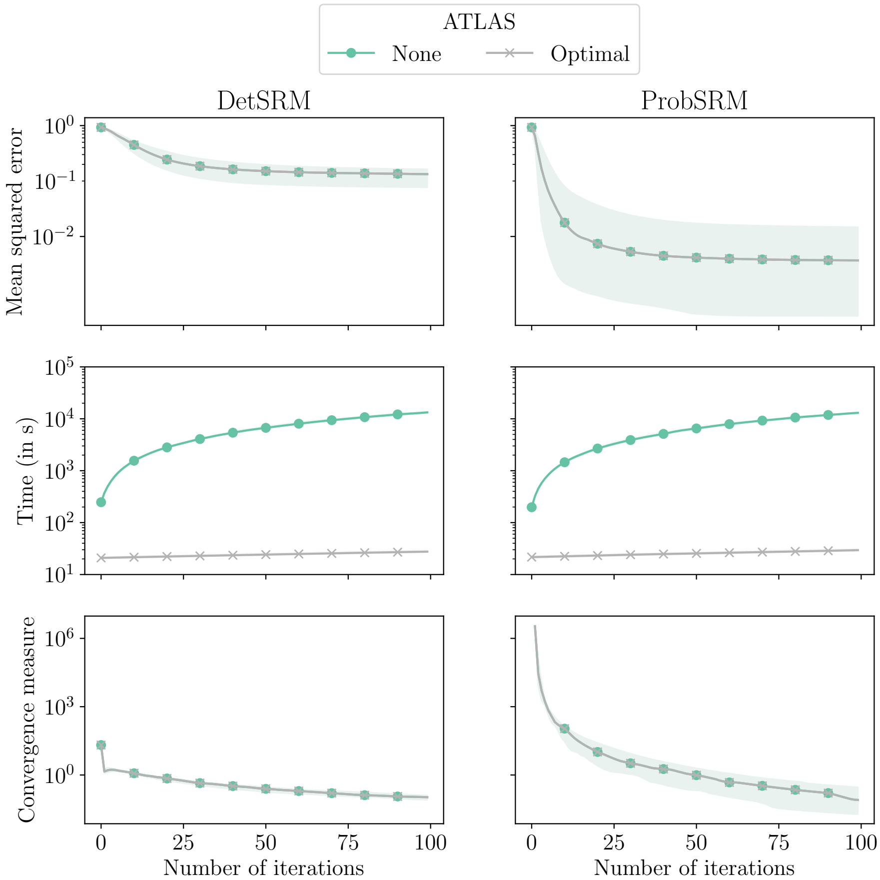


Figure 6.1: **Benchmark of SRM algorithms on synthetic data:** Performance, fitting time and convergence of SRM algorithms in the deterministic (left) or probabilistic (right) case. As expected, when optimal atlases are used, the performance is the same as if no atlas is used but the fitting time is much lower. This is even more pronounced when the number of iterations is high (and looking at convergence curves, we see that more iterations could be performed to be closer to a stationary point).

nents of each subjects but one. We select a target time-segment (9 consecutive timeframes) in the shared responses and try to localize the corresponding time segment in the components of the left-out subject using a maximum-correlation classifier. The time-segment is said to be correctly classified if the correlation between the sample and target time-segment is higher than with any other time-segment (partially overlapping time windows are excluded). We use 5-Fold cross-validation across runs: the training set contains 80% of the runs and the test set 20%, and repeat the experiment using all possible choices for left-out subjects. The mean accuracy is reported in Figure 6.2 (bottom). When the optimal atlas is used, the accuracy is the same as when no atlas is used but the fitting time is reduced by a factor 10 to 100 and so is the memory usage.

We would like to highlight here that these experiments are not exactly the same as in [36] as we use full brain data and they use regions of interest. However, the code used for this experiment is very similar to the tutorial in `https://brainiak.org/tutorials/11-SRM/`.

6.2.2 *Predict age from spatial components*

Because FastSRM is fast and memory efficient, it enables large-scale analysis of fMRI recordings of subjects exposed to the same naturalistic stimuli. We use all 647 subjects of the CamCAN dataset and demonstrate the usefulness of FastSRM by showing that the spatial components that it extracts from movie watching data are predictive of age. A key asset of FastSRM is that these spatial components can be visualized and therefore provide meaningful insights.

Note that while it is possible to do the same study using SRM, the table in figure 6.4 shows that the memory usage of SRM is 10x more important than FastSRM, almost reaching the memory limits of our cluster and is about 200x slower. FastSRM could be easily used with a much higher number of subjects while keeping the computation time and memory requirements reasonable with SRM would be impossible.

We now describe our age prediction pipeline. Functionally matched spatial components A_i are obtained using FastSRM. They are divided into two groups (train and test data) where the train set contains 80% of the data and the test set 20%. Within the train set we split again our data into two groups: the first group is used to train one Ridge model per spatial components to predict age, the second group is used to train a Random Forest to predict age from Ridge predictions. This way of stacking models is similar to the pipeline used in [130]. We use 5 fold cross validation to split the train set (so that the number of samples used to train the Random Forest is the number of elements in the train set). After the Random Forest is trained, we re-train one Ridge model per component using the full train set. On the test set each Ridge model makes a prediction and the predictions are aggregated

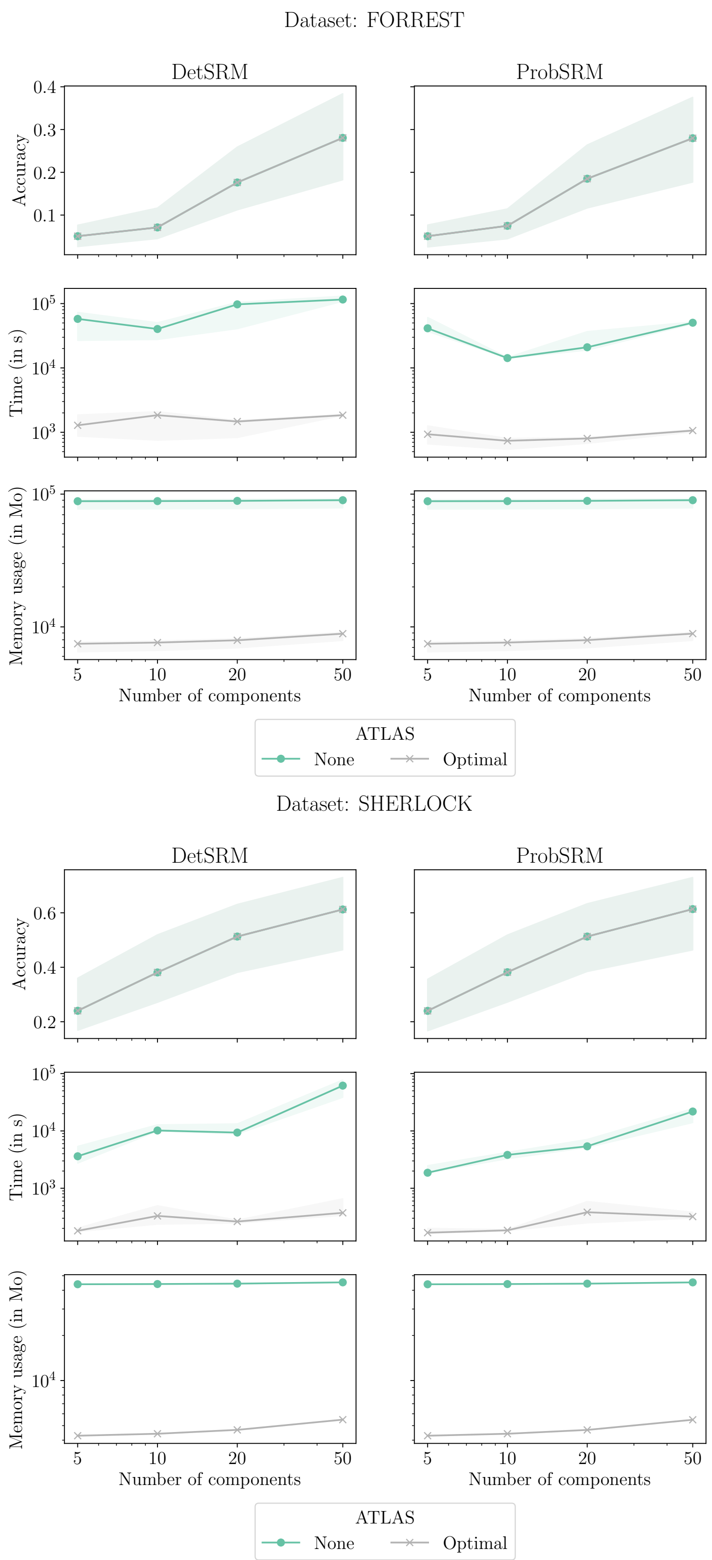


Figure 6.2: **Benchmark of SRM algorithms on fMRI data** (top) Timesegment matching accuracy (middle) Fitting time (bottom) Memory usage. When the optimal atlas is used, the accuracy is the same as when no atlas is used but the fitting time is reduced by a factor 10 to 100 and so is the memory usage.

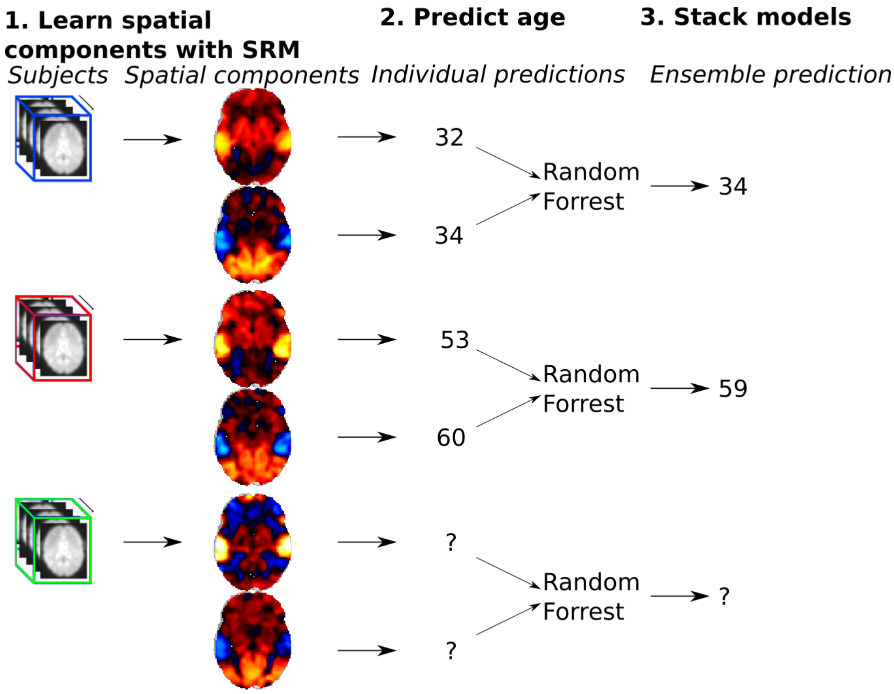


Figure 6.3: **Experiment — Predict age from spatial components extracted using FastSRM**: We first learn the spatial components from fMRI data using FastSRM. We learn one Ridge model per spatial components to predict age across subjects. Then, these models are aggregated using a Random Forest (like in [130]).

using the Random Forest model. An illustration of the process is available in Figure 6.3.

In each Ridge model, the coefficient that determines the level of l2 penalization is set by generalized cross validation, an efficient form of leave-one-out cross validation.

The train and test sets are chosen randomly. In Figure 6.4, we report the average mean absolute error (MAE) on the test set averaged over the 5 splits. FastSRM predicts age with an accuracy much better than chance resulting in a mean absolute error (MAE) of 7.5 years. Note that this is far from being the most accurate method. Using a combination of different modalities, it is possible to obtain a MAE of approximately 4.5 years [49]. However, as will be shown in the next paragraph, our method extracts a set of components specific to each individual making it more interpretable than other approaches.

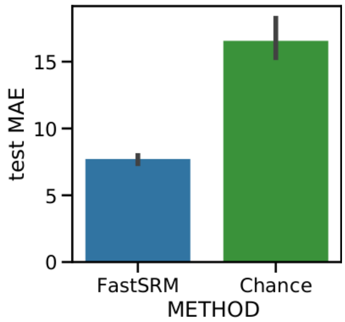
In order to assess which spatial components are most predictive of age, we assess the feature importance via the Gini index defined in [25] or [96] that measures the relative reduction in Gini impurity brought by each feature. Feature importance varies with different splits. We use the averaged feature importance over the 5 splits of our pipeline and plot the 3 most important spatial components according to this metric in Figure 6.4. These spatial components represent respectively 16%, 12% and 8% of the total feature importance and they highlight the the visual dorsal pathway, the precuneus and the visual ventral pathway respectively. The fact that averaged spatial components are interpretable and meaningful allows us to study the influence of age on brain networks involved in movie-watching. In Figure 6.5, we plot the most important spatial component averaged within groups of ages. We see that these spatial components evolve with age allowing us to visually identify which regions are meaningful. It turns out that aging is mostly reflected in brain activity as a fading of activity in the spatial correlates of movie watching, particularly in the dorsal visual cortex.

6.3 CONCLUSION

As studies using naturalistic stimuli will tend to become more common and large within and across subjects, we need scalable models in terms of computation time and memory usage. This is what FastSRM provides. We show that while FastSRM is provably equivalent to its SRM counterpart in terms of performance, it has much lower fitting time and memory requirements.

FastSRM allows large scale analysis of fMRI data of subjects exposed to naturalistic stimuli. As one example of such analysis, we show that it can be used to predict age from movie-watching data. Interestingly, although FastSRM is an unsupervised model, it extracts meaningful networks and as such constitutes a practical way of studying subjects exposed to naturalistic stimuli.

We also show that individual information can be extracted from the fMRI activity when subjects are exposed to naturalistic stimuli. Our predictive model is reminiscent of that of [23], that have shown that ICA components obtained from the decomposition of resting state data carry important information on individual characteristics. Our model inherits from all the weaknesses of SRM including the fact that mixing matrices are assumed to be orthogonal which is rather unrealistic. In later chapters, we see how this constraint can be released.



Algorithm	Memory usage (in Go)	Fitting time (in minutes)
FastSRM	18	9
ProbSRM	178	2232

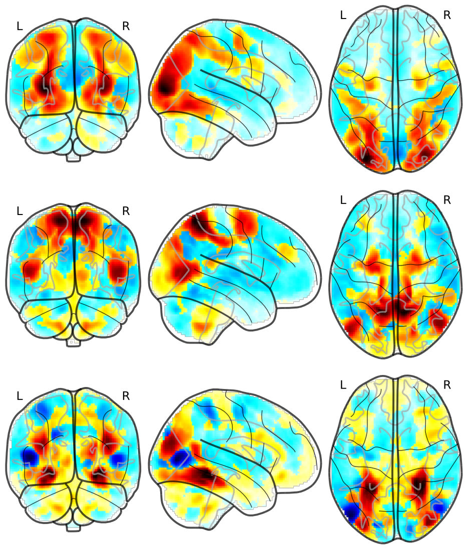


Figure 6.4: **Age prediction from spatial components**: (top) FastSRM predicts age with a better accuracy than chance resulting in a mean absolute error (MAE) of 7.5 years. (middle) FastSRM is more than 200x faster than ProbSRM and uses 10x less memory, hence it scales better than ProbSRM. (bottom) The three most important spatial components in terms of the reduction in Gini impurity they bring (see Gini importance or Feature importance in [25], [96]). From top to bottom, the most important spatial component (feature importance: 16%) highlights the visual dorsal pathway, the second most important spatial component (feature importance: 12%) highlights the precuneus and the third most important spatial component (feature importance: 8%) highlights the visual ventral pathway.

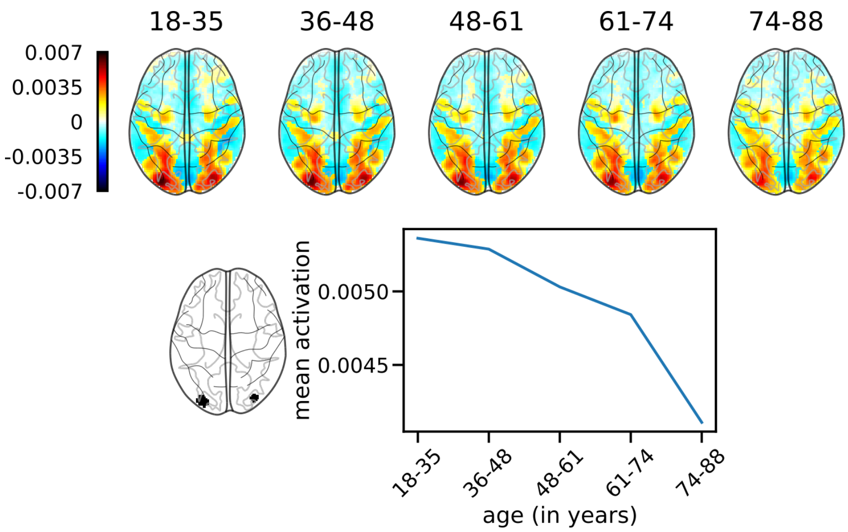


Figure 6.5: **Evolution of the most predictive spatial component with age**: (Top) Spatial component most predictive of age averaged within groups of different age (18-35, 36-48, 48-61, 61-74, 74-88). (Bottom) Mean activation in the region highlighted by the mask on the left. We see that the activity in the dorsal pathway decreases with age, which explains why this spatial component is a good predictor of age.

Part III

MULTIVIEW ICA

7

MULTIVIEW ICA THEORY

In chapter 5 and chapter 6, we have introduced a fast version of the shared response model (SRM). While SRM provides a useful dimension reduction framework, it assumes orthogonality of the mixing matrices, which is not biologically plausible.

In this chapter, we propose a novel group ICA method called *MultiView ICA*. In contrast to most Group ICA methods, MultiViewICA is grounded in a probabilistic model of the problem and comes with statistical guarantees such as asymptotic efficiency.

MultiViewICA models each subject's dataset as a linear combination of a common components matrix with additive Gaussian noise. Importantly, we consider that the noise is on the components and not on the sensors. This greatly simplifies the likelihood of the model which can even be written in closed-form.

Despite its simplicity, MultiView ICA allows for an expressive representation of inter-subject variability through subject-specific functional topographies (mixing matrices) and variability in the individual response (with noise in the component domain). To the best of our knowledge, this is the first time that such a tractable likelihood is proposed for multi-subject ICA. The likelihood formulation shares similarities with the usual ICA likelihood, which allows us to develop a fast and robust alternate quasi-Newton method for its maximization.

We first introduce the MultiView ICA model, and show that it is identifiable. We then write its likelihood in closed form, and maximize it using an alternate quasi-Newton method. We also provide a sensitivity analysis for MultiView ICA, and show that the choice of the noise parameter in the algorithm has little influence on the output.

7.1 MULTIVIEW ICA FOR SHARED RESPONSE MODELLING

7.1.1 *Model, likelihood and approximation*

Given m subjects, we model the data of subject i as a random vector $\mathbf{x}_i \in \mathbb{R}^p$ such that:

$$\mathbf{x}_i = A_i(\mathbf{s} + \mathbf{n}_i), \quad i = 1, \ldots, m \tag{7.1}$$

where $\mathbf{s} \in \mathbb{R}^p$ are the shared independent components, $\mathbf{n}_i \in \mathbb{R}^p$ is individual noise and $A_i \in \mathbb{R}^{p \times p}$ are the individual mixing matrices, assumed to be full-rank. In practice we have access to n observations of $\mathbf{x}_i$ assumed independent and identically distributed that we stack into a matrix $X_i \in \mathbb{R}^{p,n}$. For simplicity, we assume that the components share the same density δ, so that the independence assumption is $p(\mathbf{s}) = \prod_{j=1}^{p} \delta(s_j)$. Finally, we assume that the noise is Gaussian decorrelated with variance σ^2, $\mathbf{n}_i \sim \mathcal{N}(0, \sigma^2 I_p)$, and that the noise is independent across subjects and independent from the components. The assumption of additive white noise on the components models individual deviations from the shared components $\mathbf{s}$. It is equivalent to having noise on the sensors with covariance $\sigma^2 A_i (A_i)^\top$, i.e. a scaled version of the data covariance without noise.

Since the components are shared by the subjects, there are many more observed variables than components in the model: there are p components, while there are $p \times m$ observations. Therefore, model (7.1) can be seen as an instance of *undercomplete* ICA. The goal of multiview ICA is to recover the mixing matrices A_i from observations of the $\mathbf{x}_i$. The following proposition extends the standard idenfitiability theory of ICA [38] to multiview ICA, and shows that recovering the components/mixing matrices is a well-posed problem up to scale and permutation.

Proposition 12 (Identifiability of MultiView ICA). *Consider $\mathbf{x}_i$, $i = 1 \dots m$, generated from (7.1). Assume that $\mathbf{x}_i = A_i'(\mathbf{s}' + \mathbf{n}_i')$ for some invertible matrices $A_i' \in \mathbb{R}^{p \times p}$, independent non-Gaussian components $\mathbf{s}' \in \mathbb{R}^p$ and Gaussian noise $\mathbf{n}_i'$. Then, there exists a scale and permutation matrix $P \in \mathbb{R}^{p \times p}$ such that for all i, $A_i' = A_i P$.*

The proof is available in appendix A.1.1.

We propose a maximum-likelihood approach to estimate the mixing matrices. We denote by $W_i = (A_i)^{-1}$ the unmixing matrices, and view the likelihood as a function of W_i rather than A_i.

To derive the likelihood, we start by conditioning on $\mathbf{s}$. Then, we make a variable transformation from $\mathbf{x}_i$ to $\mathbf{n}_i = W_i \mathbf{x}_i - \mathbf{s}$, as opposed to the transformation to $\mathbf{s}$ as is usual in ICA. Using the probability transformation formula, we obtain

$$p_{\mathbf{x}_i|\mathbf{s}}(\mathbf{x}_i|\mathbf{s}) = |W_i| p_{\mathbf{n}_i}(W_i \mathbf{x}_i - \mathbf{s}) \tag{7.2}$$

where $p_{\mathbf{n}_i}$ is the density of $\mathbf{n}_i$. Note that the $\mathbf{x}_i$ are conditionally independent given $\mathbf{s}$, so we have:

$$p_{\mathbf{x}|\mathbf{s}}(\mathbf{x}|\mathbf{s}) = \prod_{i=1}^{m} |W_i| p_{\mathbf{n}_i}(W_i \mathbf{x}_i - \mathbf{s}) \tag{7.3}$$

and we next get the joint density as:

$$p_{\mathbf{x},\mathbf{s}}(\mathbf{x}, \mathbf{s}) = p_{\mathbf{s}}(\mathbf{s}) \prod_{i=1}^{m} |W_i| p_{\mathbf{n}^i}(W_i \mathbf{x}_i - \mathbf{s}) \tag{7.4}$$

Integrating out $\mathbf{s}$ and taking the log and expectation gives the expected negative likelihood:

$$\begin{aligned}\mathcal{L}(W_1, \dots, W_m) = &-\sum_{i=1}^{m} \log |W_i| \\ &- \mathbb{E}[\log \left(\int_{\mathbf{s}} \exp \left(-\frac{1}{2\sigma^2} \sum_{i=1}^{m} \|W_i \mathbf{x}_i - \mathbf{s}\|^2 \right) p(\mathbf{s}) d\mathbf{s} \right)]\end{aligned} \tag{7.5}$$

up to additive constants.

The integral in 7.5 after factorization, is given by

$$\int_{\mathbf{s}} \prod_{j=1}^{p} \exp \left(-\frac{1}{2\sigma^2} \sum_{i=1}^{m} (W_{ij}^\top \mathbf{x}_i - s_j)^2 \right) \delta(s_j) d\mathbf{s} \tag{7.6}$$

where W_{ij} is the j-th line of W_i. Denote $y_{ij} = W_{ij}^\top \mathbf{x}_i$ and $\tilde{s}_j = \frac{1}{m}\sum_{i=1}^{m} y_{ij}$. Fix j, and drop it to simplify notation. Then we need to solve the integral

$$\begin{aligned}&\int_s \exp \left(-\frac{1}{2\sigma^2} \sum_{i=1}^{m} (y_i - s)^2 \right) \delta(s) ds \\ &= \int_s \exp \left(-\frac{1}{2\sigma^2} [m(\tilde{s} - s)^2 + \sum_{i=1}^{m} (y_i - \tilde{s})^2] \right) \delta(s) ds \\ &= \exp \left(-\frac{1}{2\sigma^2} \sum_{i=1}^{m} (y_i - \tilde{s})^2 \right) \int_z \exp \left(-\frac{m}{2\sigma^2} z^2 \right) \delta(\tilde{s} - z) dz\end{aligned}$$

where we have made the change of variable $z = \tilde{s} - s$. The remaining integral simply means that δ is smoothed by a Gaussian kernel, which can be computed exactly if δ is a Gaussian mixture. We therefore define $f(s) = -\log \left(\int_z \exp \left(-\frac{m}{2\sigma^2} z^2 \right) \delta(s - z) dz \right)$.

The expected negative log-likelihood becomes

$$\mathcal{L}(W_1, \dots, W_m) = -\sum_{i=1}^{m} \log |W_i| + \frac{1}{2\sigma^2} \sum_{i=1}^{m} \mathbb{E}[\|W_i \mathbf{x}_i - \tilde{\mathbf{s}}\|^2] + \mathbb{E}[f(\tilde{\mathbf{s}})] \tag{7.7}$$

Multiview ICA is then performed by minimizing $\mathcal{L}$, and the estimated shared components are $\tilde{S} = \frac{\sum_i W_i X_i}{m}$. The negative log-likelihood $\mathcal{L}$ is quite simple, and importantly, can be computed easily given the parameters of the model and the data; it does not involve any intractable integral.

For one subject ($m = 1$), $\mathcal{L}(W_1)$ simplifies to the negative log-likelihood of ICA and we recover Infomax [19, 32], where the component log-pdf is replaced with the smoothed f.

7.1.2 *Alternate quasi-Newton method for MultiView ICA*

The parameters of the model are estimated by minimizing $\mathcal{L}$. We propose a combination of quasi-Newton method and alternate minimization for this task. First, $\mathcal{L}$ is non-convex: it is only defined when the W_i are invertible, which is a non-convex set. Therefore, we only look for local minima as usual in ICA. We propose an alternate minimization scheme, where $\mathcal{L}$ is alternatively diminished with respect to each W_i. When all matrices $W_1, \dots, W_m$ are fixed but one, W_i, $\mathcal{L}$ can be rewritten, up to an additive constant

$$\begin{aligned}\mathcal{L}_i(W_i) &= -\log|W_i| \\ &+ \frac{1-1/m}{2\sigma^2}\mathbb{E}[\|W_i\mathbf{x}_i - \frac{m}{m-1}\tilde{\mathbf{s}}_{-i}\|^2 + f(\frac{1}{m}\, W_i\mathbf{x}_i + \tilde{\mathbf{s}}_{-i})]\end{aligned} \quad (7.8)$$

with $\tilde{\mathbf{s}}_{-i} = \frac{1}{m}\sum_{j\neq i} W_j\mathbf{x}_j$. This function has the same structure as the usual maximum-likelihood ICA cost function: it is written $\mathcal{L}_i(W_i) = -\log|W_i| + \mathbb{E}[g(W_i\mathbf{x}_i)]$, where $g(\mathbf{y}) = \sum_{j=1}^{p} f(\frac{y_j}{m} + \tilde{\mathbf{s}}_{-i,j}) + \frac{1-1/m}{2\sigma^2}(y_j - \frac{m}{m-1}\tilde{\mathbf{s}}_{-i,j})^2$ where $\tilde{\mathbf{s}}_{-i,j}$ is the j-th component of $\tilde{\mathbf{s}}_{-i}$. Fast quasi-Newton algorithms [3, 160] have been proposed for minimizing such functions. We employ a similar technique as [160], which we now describe.

Quasi-Newton methods are based on approximations of the Hessian of $\mathcal{L}_i$. The relative gradient (resp. Hessian) [6, 34] of $\mathcal{L}_i$ is defined as the matrix $G_i \in \mathbb{R}^{p\times p}$ (resp. tensor $\mathcal{H}_i \in \mathbb{R}^{p\times p\times p\times p}$) such that as the matrix $E \in \mathbb{R}^{p\times p}$ goes to 0, we have $\mathcal{L}_i((I_p + E)W_i) \simeq \mathcal{L}_i(W_i) + \langle G_i, W_i\rangle + \frac{1}{2}\langle E, \mathcal{H}_i E\rangle$. Standard manipulations yield:

$$G_i = \mathbb{E}[\frac{1}{m}f'(\tilde{\mathbf{s}})(\mathbf{y}_i)^\top + \frac{1-1/m}{\sigma^2}(\mathbf{y}_i - \frac{m}{m-1}\tilde{\mathbf{s}}_{-i})(\mathbf{y}_i)^\top] - I_p \quad (7.9)$$

where $\mathbf{y}_i = W_i\mathbf{x}_i$.

$$(\mathcal{H}_i)_{abcd} = \delta_{ad}\delta_{bc} + \delta_{ac}\mathbb{E}[\left(\frac{1}{m^2}f''(\tilde{s}_a) + \frac{1-1/m}{\sigma^2}\right) y_{ib}y_{id}] \quad (7.10)$$

$$\text{for } a, b, c, d = 1 \dots p$$

Newton's direction is then $-(\mathcal{H}_i)^{-1} G_i$. However, this Hessian is costly to compute (it has $\simeq p^3$ non-zero coefficients) and invert (it can be seen as a big $p^2 \times p^2$ matrix). Furthermore, to enforce that Newton's direction is a descent direction, the Hessian matrix should be regularized in order to eliminate its negative eigenvalues [109], and $\mathcal{H}_i$ is not guaranteed to be positive definite. These obstacles render the computation of Newton's direction impractical. Luckily, if we assume that the signals in $\mathbf{y}_i$ are independent, several coefficients cancel, and the Hessian simplifies to the approximation

$$(H_i)_{abcd} = \delta_{ad}\delta_{bc} + \delta_{ac}\delta_{bd}\Gamma^i_{ab} \tag{7.11}$$
$$\text{with } (\Gamma_i)_{ab} = \mathbb{E}[\left(\frac{1}{m^2}f''(\tilde{s}_a) + \frac{1-1/m}{\sigma^2}\right)(y_{ib})^2]$$

This approximation is sparse: it only has $p(2p-1)$ non-zero coefficients. In order to better understand the structure of the approximation, we can compute the matrix $(H_i M)$ for $M \in \mathbb{R}^{p\times p}$. We find $(H_i M)_{ab} = (\Gamma_i)_{ab}M_{ab} + M_{ba}$: $H_i M_{ab}$ only depends on M_{ab} and M_{ba}, indicating a simple block diagonal structure of H_i. The operator H_i is therefore easily regularized and inverted: $\left((H_i)^{-1}M\right)_{ab} = \frac{\Gamma^i_{ba}M_{ab} - M_{ba}}{\Gamma^i_{ab}\Gamma^i_{ba} - 1}$. Finally, since this approximation is obtained by assuming that the $\mathbf{y}_i$ are independent, the direction $-H_i^{-1}G_i$ is close to Newton's direction when the $\mathbf{y}_i$ are close to independence, leading to fast convergence. Algorithm 2 alternates one step of the quasi-Newton method for each subject until convergence. A backtracking line-search is used to ensure that each iteration leads to a decrease of $\mathcal{L}_i$. The algorithm is stopped when the maximum norm of the gradients over one pass on each subject is below some tolerance level, indicating that the algorithm is close to a stationary point.

Algorithm 2: Alternate quasi-Newton method for MultiView ICA

Input: Dataset $(\mathbf{x}_i)_{i=1}^m$, initial unmixing matrices W_i, noise parameter σ, function f, tolerance ε
Set tol$= +\infty$, $\tilde{\mathbf{s}} = \frac{1}{m}\sum_{i=1}^p W_i\mathbf{x}_i$
while *tol*$> \varepsilon$ **do**
  tol = 0
  for $i = 1 \dots m$ **do**
    Compute $\mathbf{y}_i = W_i\mathbf{x}_i$, $\tilde{\mathbf{s}}_{-i} = \tilde{\mathbf{s}} - \frac{1}{m}\mathbf{y}_i$, gradient G_i (eq. (7.9)) and Hessian H_i (eq. (7.11))
    Compute the search direction $D = -H_i^{-1}G_i$
    Find a step size ρ such that $\mathcal{L}_i((I_p + \rho D)W_i) < \mathcal{L}_i(W_i)$ with line search
    Update $\tilde{\mathbf{s}} = \tilde{\mathbf{s}} + \frac{\rho}{m}DW_i\mathbf{x}_i$, $W_i = (I_p + \rho D)W_i$, tol$= \max(\text{tol}, \|G_i\|)$
  end
end
return *Estimated unmixing matrices* W_i, *estimated shared components* $\tilde{\mathbf{s}}$

7.1.3 *Robustness to model misspecification*

Algorithm 2 has two hyperparameters: σ and the function f. The latter is usual for an ICA algorithm, but the former is not. We study the impact of these parameters on the separation capacity of the algorithm, when these parameters do not correspond to those of the generative model (7.1).

Proposition 13. *We consider the cost function* $\mathcal{L}$ *in eq.* (7.7) *with noise parameters* σ *and function* f*. Assume sub-linear growth on* f'*:* $|f'(x)| \leqslant c|x|^{\alpha} + d$ *for some* $c, d > 0$ *and* $0 < \alpha < 1$*. Assume that* $\mathbf{x}_i$ *is generated following model* (7.1)*, with noise parameter* σ' *and density of the component* d' *which need not be related to* σ *and* f*. Then, there exists a diagonal matrix* Λ *such that* $(\Lambda(A^1)^{-1}, \dots, \Lambda(A^m)^{-1})$ *is a stationary point of* $\mathcal{L}$*, that is* $G^1, \dots, G^m = 0$ *at this point.*

The proof is available in appendix A.1.2.

The sub-linear growth of f' is a customary hypothesis in ICA which implies that d has heavier-tails than a Gaussian, and in appendix A.1.2 we provide other conditions for the result to hold. In this setting, the shared components estimated by the algorithm are $\tilde{S} = \Lambda(S + \frac{1}{m}\sum_{i=1}^{m} N_i)$, which is a scaled version of the best estimate of the shared components under the Gaussian noise hypothesis.

This proposition shows that, up to scale, the true unmixing matrices are a stationary point for Algorithm 2: if the algorithm starts at this point it will not move. The question of stability is also interesting: if the algorithm is initialized *close* to the true unmixing matrices, will it converge to the true unmixing matrix? In the appendix A.1.3, we provide an analysis similar to [33], and derive sufficient numerical conditions for the unmixing matrices to be local minima of $\mathcal{L}$.

7.2 RELATED WORK

In contrast to ConcatICA or CanICA (see section 4.2.2.1), MultiView ICA maximizes a likelihood, which brings statistical guarantees like consistency or asymptotic efficiency. Furthermore MultiViewICA finds individual and shared independent components in a single step. This differs from ConcatICA or GroupICA that require additional steps when individual components are needed such as back-projection [28] or dual-regression [17].

The approach of [64] (see section 4.2.2.2) optimizes the more general model $\mathbf{x}_i = A_i\mathbf{s} + \mathbf{n}_i$. The likelihood for this model involves an intractable high dimensional integral that is cumbersome to evaluate, and is then optimized with an EM algorithm using an inexact E-step. Having the simpler model $\mathbf{x}_i = A_i(\mathbf{s} + \mathbf{n}_i)$ leads to a closed-form likelihood, that can then be optimized by more efficient means. Note that in MultiView ICA, the noise can be interpreted as individual variability rather than sensor noise.

The SR-ICA approach of [157] performs dimension reduction, merging of individual data and independent component estimation. It is therefore similar to our method. However, they propose to modify the

FastICA algorithm [74] in a rather heuristic way, without specifying an optimization problem, let alone maximizing a likelihood. In the experiments on fMRI data in appendix A.1.4, we obtain better performance with MultiView ICA than the reported performance of SR-ICA.

One strength of our model is that we only assume that the mixing matrices are invertible and still enjoy identifiability whereas some other approaches impose additional constraints. For instance tensorial methods [18] assume that the mixing matrices are the same up to diagonal scaling. Other methods impose a common mixing matrix [27, 39, 62, 103]. Like PCA, the Shared Response Model [36] (SRM) assumes orthogonality of the mixing matrices. While the model defines a simple likelihood and provides an efficient way to reduce dimension, the orthogonal constraint may not be plausible.

Deep Learning methods, such as convolutional auto-encoders (CAE), can also be used to find the subject specific unmixing [37]. While these nonlinear extensions of the aforementioned methods are interesting, these models are hard to train and interpret. In the experiments on fMRI data in appendix A.1.4, we obtain better accuracy with MultiView ICA than that of CAE reported in [37].

A different path to multi-subject ICA is to extract independent components with individual ICA in each subject and align them. We propose a simple baseline approach to do so called *PermICA*. Inspired by the heuristic of the hyperalignment method [69] we choose a reference subject and first match the components of all other subjects to the components of the reference subject. The process is then repeated multiple times, using the average of previously aligned components as a reference. Finally, group components are given by the average of all aligned components. We use the Hungarian algorithm to align pairs of mixing matrices [147]. Alternative approaches involving clustering have also been developed [22, 53].

Lastly, IVA based methods (see section 4.2.3) estimate view-specific components but shared components are not modeled explicitly.

7.3 CONCLUSION

In this chapter, we have proposed a novel unsupervised algorithm that reveals latent components observed through different views. Using an independence assumption, we have demonstrated that the model is identifiable. In contrast to previous approaches, the proposed model leads to a closed-form likelihood, which we then optimize efficiently using a dedicated alternate quasi-Newton approach. Our approach enjoys the statistical guarantees of maximum-likelihood theory, while still being tractable.

In the next chapter, we evaluate the performance of MultiView ICA on synthetic data, on EEG and fMRI data and compare its performance to other GroupICA methods.

8 MULTIVIEW ICA IN PRACTICE

In chapter 7, we have introduced MultiViewICA and presented its theoretical properties. In this chapter, we empirically verify through extensive experiments on fMRI and MEG data that MultiView ICA improves component identification with respect to competing methods, suggesting that the expressiveness and robustness of this model make it a useful tool for multivariate neural signal analysis.

8.1 EXPERIMENTAL SETTING

In the following, the noise parameter in MultiviewICA is always fixed to $\sigma = 1$. We use the function $f(\cdot) = \log\cosh(\cdot)$, giving $f'(\cdot) = \tanh(\cdot)$ (f' is called the learning function or non-linearity [73]). We use the Infomax cost function [19] with the same non-linearity to run standard ICA, with the Picard algorithm [3] for fast and robust minimization of the cost function. Picard is applied with the default hyper-parameters. The code for MultiViewICA is available online at `https://github.com/hugorichard/multiviewica`.

We compare the following methods to obtain p components:

PermICA is described in section 7.2. *SRM* is the FastSRM algorithm described in part II. *ConcatICA* is described in section 4.2.2.1. *PCA+ConcatICA* corresponds to ConcatICA applied on subject data that have been first individually reduced by PCA with p components. *CanICA* is described in section 4.2.2.1. We define the chance level as the performance of an algorithm that computes unmixing matrices and projections to lower dimensional space by sampling random numbers from a standard normal distribution.

The subject specific dimension reduction in MultiView ICA, PermICA, ConcatICA and CanICA is performed with SRM in fMRI experiments and subject-specific PCA in MEG experiments.

The shape of the input data (v, t, m) depend on the dataset used and the experiment performed. A description of the datasets used is available in section 3.1.5 for fMRI data and in section 3.2.5 for MEG data. The values for p (the number of components) is given in the descriptions of the experiments. Note that in the special case of the experiments on the sherlock fMRI dataset, the experimental pipeline is available on github `https://github.com/hugorichard/multiviewica/tree/master/real_data_experiments` (including downloading and preprocessing of the data).

Since the cost function $\mathcal{L}$ is non-convex, having a good initialization can make a difference in the final result. We propose a two stage

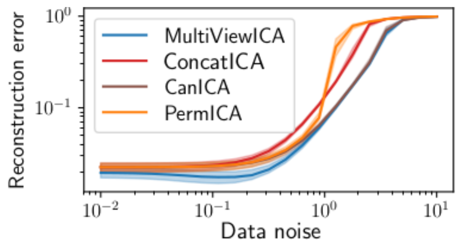


Figure 8.1: **Synthetic experiment**: reconstruction error of the algorithms on data following model $\mathbf{x}_i = A_i(\mathbf{s} + \mathbf{n}_i)$.

approach. We begin by applying PermICA on the datasets, which gives us a first set of unimixing matrices $W_1, \dots, W_m$. Note that we could also use ConcatICA for this task. Next, we perform a diagonal scaling of the mixing matrices, i.e. we find the diagonal matrices $\Lambda_1, \dots, \Lambda_m$ such that $\mathcal{L}(\Lambda_1 W_1, \dots, \Lambda_m W_m)$ is minimized. To do so, we employ Algorithm 2 but only take into account the diagonal of the descent direction at each step: the update rule becomes $W_i \leftarrow (I_p + \rho \text{Diag}(D))W_i$. The initial unmixing matrices for Algorithm 2 are then taken as $\Lambda_1 W_1, \dots, \Lambda_m W_m$. Empirically, we find that this two stage procedure allows the algorithm to start close to a satisfactory solution. A summary of our quantitative results on real data is available in appendix A.6.

8.2 SYNTHETIC EXPERIMENT

We validate our method on synthetic data generated according to the model in equation (7.1). The components are generated i.i.d. from a Laplace density $d(x) = \frac{1}{2}\exp(-|x|)$. The mixing matrices $A_1, \cdots, A_m$ are generated with i.i.d. entries following a normal law. Each compared algorithm returns a sequence of estimated unmixing matrices $W_1, \dots, W_m$. The performance of an algorithm is measured by the reconstruction error between the estimated components and the true components.

We use $m = 10$ datasets, $p = 15$ components and $n = 1000$ samples. Each experiment is repeated with 100 random seeds. We vary the noise level in the data generation from 10^{-2} to 10 and display the performance of algorithms in Figure 8.1.

Multiview ICA has uniformly better performance than the other algorithms, which illustrates the strength of maximum-likelihood based methods. In accordance with results of section 7.1, it is able to separate the components even with misspecified noise parameter and component density.

8.3 fMRI experiments

We use the *Forrest*, *Sherlock*, *Raiders* and *CLIPS* datasets described in section 3.1.5.

8.3.1 *Reconstructing the BOLD signal of missing subjects*

We want to show that once unmixing matrices have been learned, they can be used to predict evoked responses across subjects. This can be used to perform missing data imputation when the sessions of some subjects are missing. In [158], the authors consider multiple fMRI datasets that share a subpart of their subjects and use the fact that some subjects are shared to transfer information across datasets. Our reconstruction experiment measures a related quantity: the methods ability to predict data of left-out subjects from other subjects.

In this experiment we apply a 6 mm spatial smoothing to all datasets. We split the data into three groups. First, we randomly choose 80% of all runs from all subjects to form the training set. Then, we randomly choose 80% of subjects and take the remaining 20% runs as testing set. The left-out runs of the remaining subjects form the validation set. The compared algorithms are run on the training set and evaluated using the testing and validation sets. After an algorithm is run on training data, it defines for each subject a *forward operator* that maps individual data to the space spanned by components and a *backward operator* that maps the component space to individual data. For instance in ICA the forward operator is the product of the dimensionality reduction projection and unmixing matrix. We estimate the shared responses on the testing set by applying the forward operators on the testing data and averaging. Finally, we reconstruct the individual data from subjects in the validation set by applying the backward operators to the shared responses. We measure the difference between the true signal and the reconstructed one using voxel-wise R^2 score. The R^2 score between two series $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{y} \in \mathbb{R}^n$ is defined as $R^2(\mathbf{x}, \mathbf{y}) = 1 - \frac{1}{n \operatorname{Var}(\mathbf{y})} \sum_{t=1}^{n} (x_t - y_t)^2$, where $\operatorname{Var}(\mathbf{y}) = \frac{1}{n} \sum_{t=1}^{n} (y_t - \frac{1}{n} \sum_{t'=1}^{n} y_{t'})^2$ is the empirical variance of $\mathbf{y}$. The R^2 score is always smaller than 1, and equals 1 when $\mathbf{x} = \mathbf{y}$. The experiment is repeated 25 times with random splits to obtain error bars. Figure 8.2 provides a visual summary of our experimental procedure.

The R^2 score per voxel depends heavily on which voxels are considered. For example voxels in the visual cortex are better reconstructed in the *sherlock* dataset than in the *forrest* dataset. Performances are therefore given in terms of mean R^2 score inside a region of interest (ROI) in order to leave out regions where there is no useful information. In order to determine the ROIs, we focus on the R2 score per voxel between the BOLD signal reconstructed by ConcatICA and the actual bold signal. We run ConcatICA with $10, 20$ and 50 components

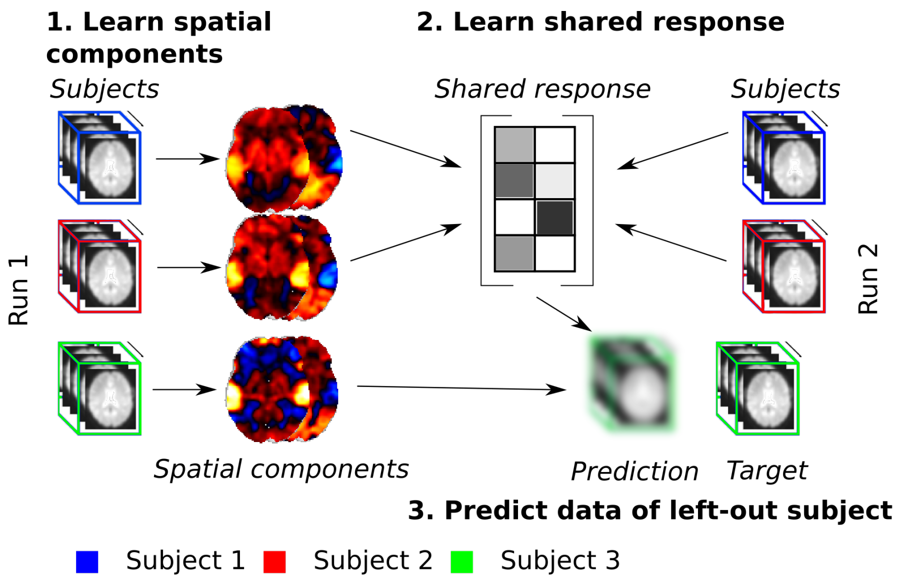


Figure 8.2: **Reconstructing the BOLD signal of missing subjects - Experimental procedure** 80% of the runs are used to compute forward operators (spatial components) for every subject (left). Then the forward operators and data from the left-out runs of 80% of the subjects one are used to compute the shared response during the left-out runs. Lastly, the shared response during the left-out runs and the forward operators of the test subjects are used to predict the data of the test subjects during the left-out runs. The performance of the model is measured by comparing the prediction and true data using the R^2 score.

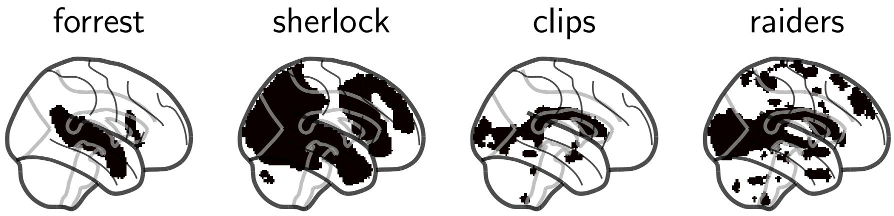


Figure 8.3: **Data-driven choice of ROI** Chosen ROIs for the experiment: Reconstructing the BOLD signal of missing subjects.

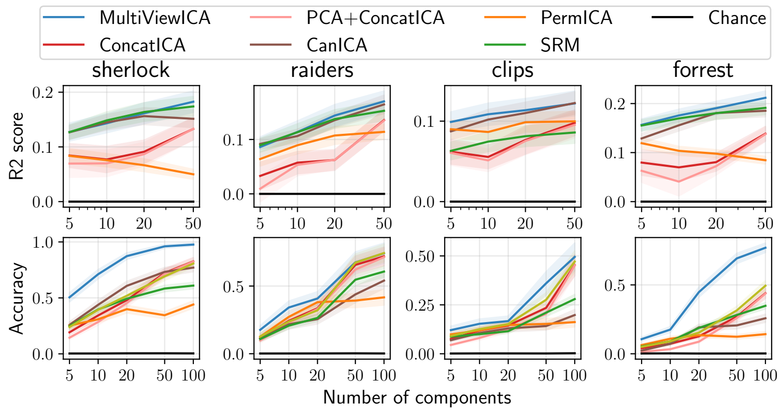


Figure 8.4: *Top:* **Reconstructing the BOLD signal of missing subjects**. Mean R^2 score between reconstructed data and true data (higher is better). *Bottom:* **Between subjects time-segment matching**. Mean classification accuracy. Error bars represent a 95 % confidence interval over cross validation splits.

and select the voxels that obtained a positive R2 score for all sets of components. We discard voxels with an R2 score above 80% as they visually correspond to artefacts and apply a binary opening using a unit cube as the structuring element. The chosen regions are plotted in figure 8.3.

In Figure 8.4 (top) we report the mean R^2 score within regions of interest. MultiView ICA has similar or better performance than the other methods on all datasets. This demonstrates its ability to capture inter-subject variability, making it a candidate of choice to handle missing data or perform transfer learning.

For completeness, we plot in Figure 8.5, for ConcatICA, SRM and MultiViewICA, the R2 score per voxel using 50 components for datasets *sherlock*, *forrest*, *raiders* and *clips*. As could be anticipated from the task definition, *forrest* obtains high reconstruction accuracy in the auditory cortices, while *clips* shows good reconstruction in the visual cortex (occipital lobe mostly); the richer *sherlock* and *raiders* datasets yield good reconstructions in both domains, but also in other systems (language, motor). We can also see that data reconstructed by MultiViewICA are a better approximation of the original data than other methods. This is particularly obvious for the *clips* datasets where it is

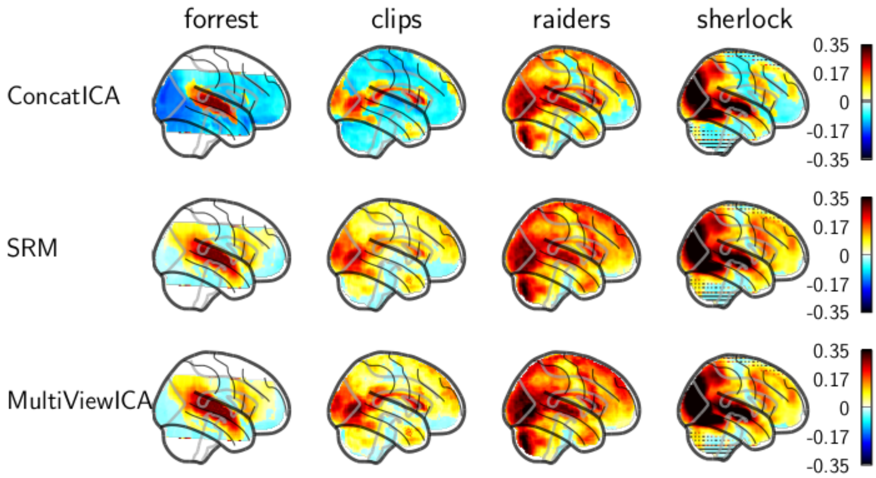


Figure 8.5: **Reconstructing the BOLD signal of missing subjects: Reconstruction R2 score per voxel** We plot for ConcatICA, SRM and MultiViewICA, the R2 score per voxel using 50 components for datasets *sherlock*, *forrest*, *raiders* and *clips*. We can see that data reconstructed by MultiViewICA are more faithful reproduction of the original data than other methods.

clear that voxels in the posterior part of the superior temporal sulcus are better recovered by MultiViewICA than by SRM or ConcatICA.

8.3.2 *Between subjects time-segment matching*

We reproduce the time-segment matching experiment described in section 6.2.1 and display the results in Figure 8.4 (bottom). MultiView ICA yields a consistent and substantial improvement in accuracy compared to other methods on the four datasets. We see a marked improvement on the sherlock and forrest datasets. A possible explanation lies in the preprocessing pipeline. Sherlock data undergo a 6 mm spatial smoothing and Forrest data are acquired at a higher resolution (7T vs 3T for other data). This affects the signal to noise ratio.

In order to investigate the practical impact of the choice of hyperparameter σ, we compute the accuracy of the MultiView ICA algorithm with different choice of σ. Results are reported in Figure 8.6. MultiviewICA performs consistently well for a wide range of noise parameter values, and only breaks at very high values. It supports the theoretical claim issued in Proposition 13 that the noise parameter is of little importance.

In appendix A.4, we plot the average forward operator across subjects of MultiView ICA and ConcatICA with 5 components on the forrest, sherlock, raiders and clips datasets.

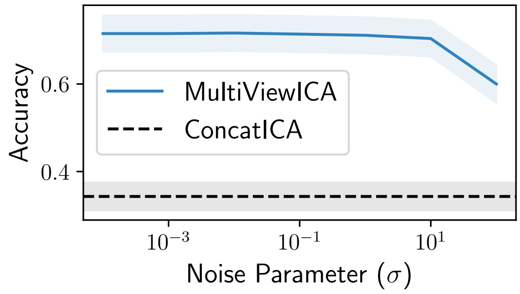


Figure 8.6: **Effect of the parameter** σ: We compute the accuracy of the MultiViewICA pipeline on the time-segment matching experiment for various values of the σ hyperparameter over a grid. The accuracy varies only marginally with σ.

8.3.3 *Between-runs time-segment matching*

We measure the ability of each algorithm to extract meaningful shared components that correlate more when they correspond to the same stimulus than when they correspond to distinct stimuli. We use the *raiders-full* dataset, which allows this kind of analysis because subjects watch some selected scenes from the movie twice, during the first two runs (1 and 2) and the last two (11 and 12). First, the forward operators are learned by fitting each algorithm with 20 components on the data of all 11 subjects using all 12 runs. We then select a subset of 8 subjects and the shared components are computed by applying the forward operators and averaging. We select a large target time-segment (50 timeframes) taken at random from run 1 and 2, and we try to localize the corresponding sample time-segment from the 10 last runs using a single component of the shared components. The time-segment is said to be correctly classified if the correlation between the target and corresponding sample time-segment is higher than with any other time-segment (partially overlapping windows are excluded). In contrast to the *between subject time-segment matching* experiment, we obtain one accuracy score per component. We repeat the experiment 10 times with different subsets of subjects randomly chosen and report the mean accuracy of the three best performing components in Figure 8.7. Error bars correspond to a 95 % confidence interval. MultiView ICA achieves the highest accuracy.

We then focus on the 3 best performing components of MultiView ICA. For each component, we plot in Figure 8.8 (left) the shared components during two sets of runs where subjects were exposed to the same scenes of the movie. We then study the localisation of these components. We average the forward operators across subjects and plot the columns corresponding to the components of interest in Figure 8.8 (right). As each column is seen as a set of weights over all voxels, it represents a spatial map.

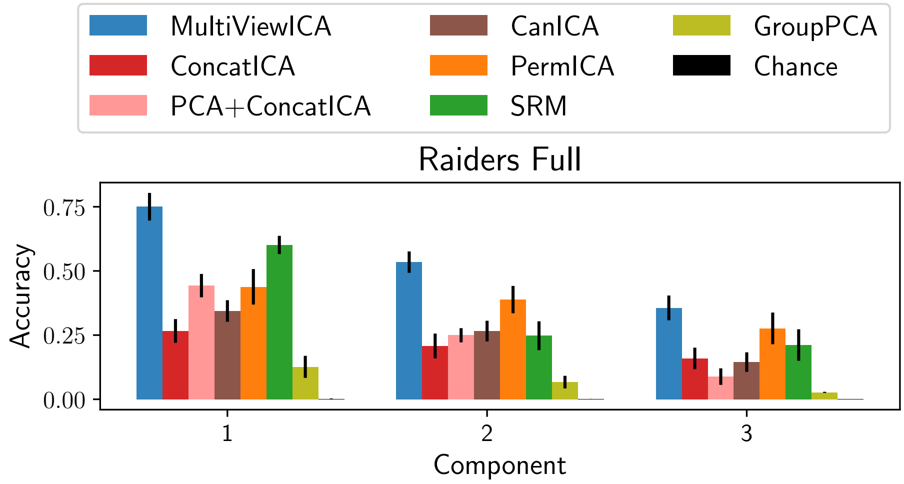


Figure 8.7: **Between runs time-segment matching**. Interesting components correlates more when they correspond to the same stimulus (same scenes of the movie) than when they correspond to distinct stimuli (different scenes). We extract 20 components and report the mean accuracy of the three best performing components

The component 1 of the shared responses follows almost the same pattern in the two set of runs corresponding to the same scenes of the movie. The spatial map corresponding to component 1 highlights the language network. In component 2, the temporal patterns during the viewing of identical scenes are also very similar. The corresponding spatial map highlights the visual network especially the visual dorsal pathway. In component 3, there exists a similarity however less striking than with the two previous components. The corresponding spatial map highlights a contrast between the spatial attention network and the auditory network.

8.4 PHANTOM MEG DATA

We demonstrate the usefulness of our approach on MEG data using the *Sinusoidal Phantom MEG* dataset where $m = 8$ dipoles at different locations produce a known sinusoidal oscillation (more details in section 3.2.5). 100 epochs are available. For each dipole, we chose $N_e = 2, \dots, 16$ epochs at random among our set of 100 epochs and concatenate them in the temporal dimension. We then apply algorithms on these data to extract $p = 20$ shared components. As we know the true component (the timecourse of the dipole), we can compute the reconstruction error of each component as the squared norm of the difference between the estimated component and the true component, after normalization to unit variance and fixing the sign. We only retain the component yielding minimal error. We also estimate for each forward operator the localization of the component by performing dipole fitting using its column corresponding to the component of minimal

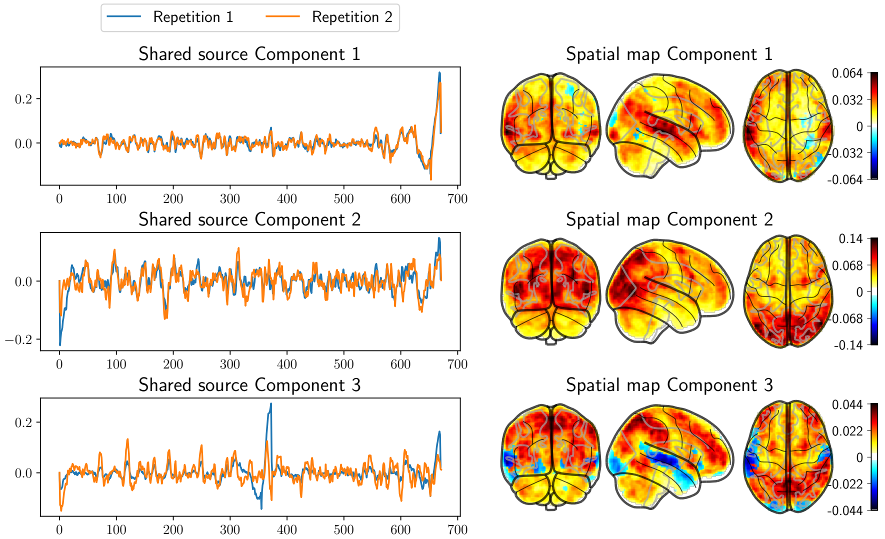


Figure 8.8: **Between-runs time segment matching: spatial maps and timecourses** *Left:* Timecourses of the 3 shared components yielding the highest accuracy. The two displayed set of runs correspond to the same scenes in the movie. *Right:* Localisation of the same shared components in the brain

error. We then compute the distance of the estimated dipole to the true dipole. These metrics are reported in figure 8.9 when the number of epochs considered N_e varies. MultiView ICA requires fewer epochs to correctly reconstruct and localize the true component.

8.5 Experiment on CamCAN dataset

Finally, we apply MultiView ICA on the CamCAN dataset [146]. A detailed description of the CamCAN dataset is available in section 3.2.5. We use the magnetometer data from the MEG of 200 subjects chosen randomly. Each subject is repeatedly presented an audio-visual stimulus. The MEG signal corresponding to these trials are then time-averaged to isolate the evoked response, yielding individual data. MultiView ICA is then applied to extract 20 shared components. 9 components were found to correspond to noise by visual inspection, and the 11 remaining are displayed in Figure 8.9. We observe that MultiView ICA recovers a very clean sequence of evoked potentials with sharp peaks for early components and slower responses for late components. In order to visualize their localization, we perform component localization for each subject by solving the inverse problem using sLORETA [113], providing a component estimate for each component. Then, we register each component estimate to a common reference brain. Finally, the component estimates are averaged, and thresholded maps are displayed in Figure 8.9. Individual maps corresponding to each component are displayed in Appendix A.3. The figure highlights both early auditory and visual cortices, also sug-

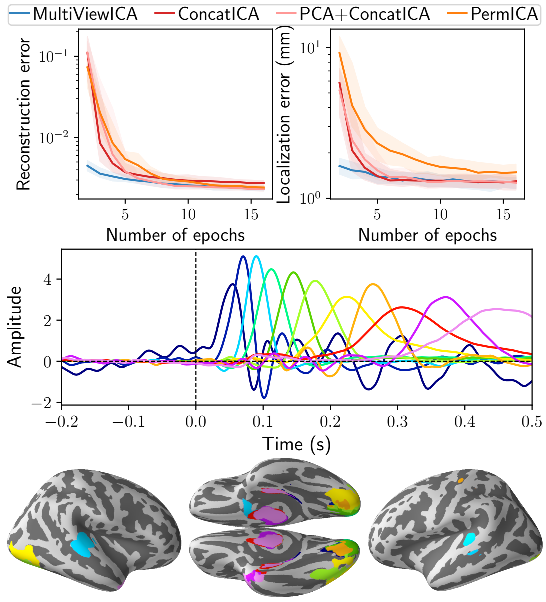


Figure 8.9: *Top:* **Experiment on MEG Phantom data**. Reconstruction error is the norm of the difference between the estimated and true component. Localization error is the distance between the estimated and true dipole. *Middle and Bottom:* **Experiment on 200 subjects from the CAM-can dataset** *Middle:* Time course of 11 shared components (one color per component). We recover clean evoked potentials. *Bottom:* Associated brain maps, obtained by averaging component estimates registered to a common reference.

gesting a propagation of the activity towards the ventral regions and higher level visual areas.

8.6 CONCLUSION

In this chapter, we have demonstrated the usefulness of MultiView ICA for neuroimaging group studies both on fMRI and MEG data, where it outperforms other methods. A limiting aspect of MultiView ICA is the assumption that the noise variance is the same across subjects. This is limiting because it does not properly model between-subjects variability. In the next chapter, we propose an extension of MultiView ICA with a more realistic noise model.

Part IV

SHARED ICA

9 SHARED ICA THEORY

In chapter 7 and chapter 8, we have introduced MultiView ICA, a well principled method to perform shared response modeling. While MultiView ICA yields good practical results, it does not model subject specific deviations from the shared response. Yet, the magnitude of the response may differ across subjects [115], as does any noise due to heart beats, respiratory artefacts or head movements [93].

This drawback is shared by most GroupICA methods that often rely on single subject ICA to recover the shared response. In addition, such methods are typically unable to separate Gaussian components.

In contrast, the framework of Independent vector analysis (IVA) [7, 85] allows subject specific variability in the unmixed data. Current implementations such as IVA-L [85], IVA-G [7], IVA-L-SOS [21], IVA-GGD [9] or IVA with Kotz distribution [8] estimate view-specific components but shared components are not modeled explicitly. Studying the components post IVA can give great insights about which components are shared and by which subjects (see [94]) and individual responses that are closed enough could be merged to produce a shared response. However, the flexibility in modeling different components from each subject comes at the cost of not having a well principled way of aggregating the components into a common shared response.

In this chapter, we introduce Shared ICA (ShICA), where each dataset is modeled as a linear transform of shared independent components contaminated by additive Gaussian noise. ShICA allows for principled extraction of the shared components (or responses) in addition to view-specific components. Since it incorporates a statistically sound noise model, it enables optimal inference of the shared responses (minimizing the mean squared error).

We first analyse the theoretical properties of the ShICA model, before providing powerful inference algorithms. First, we exhibit necessary and sufficient conditions for ShICA to be identifiable (previous work only shows local identifiability [9]), in the presence of Gaussian or non-Gaussian components. We then introduce an algorithm called ShICA-J that uses Multiset CCA to fit the model when all the components are assumed to be Gaussian. We exhibit necessary and sufficient conditions for Multiset CCA to be able to fit the model (previous work only gives sufficient conditions [92]) and provide examples on which ShICA-J can recover the mixing matrices, while Multiset CCA cannot. We next point out a practical problem, namely that even a small sampling noise can lead to large rotations of unmixing matrices when Multiset CCA is used. To address this issue and recover the correct unmixing matrices, we propose to apply joint diagonalization to the result of Multiset CCA. We further introduce ShICA-ML, a maximum likelihood estimator of ShICA that models non-Gaussian components using a Gaussian mixture model. While ShICA-ML yields more accurate components, ShICA-J is significantly faster and offers a great initialization to ShICA-ML.

9.1 SHARED ICA (SHICA): AN IDENTIFIABLE MULTI-VIEW MODEL

We assume a similar generative model as in MultiViewICA:

$$\mathbf{x}_i = A_i(\mathbf{s} + \mathbf{n}_i) \tag{9.1}$$

Like in MultiView ICA we assume that the shared components are statistically independent $p(\mathbf{s}) = \prod_{j=1}^{p} p(s_j)$, and that the individual noises are Gaussian and independent from the shared components. We assume $\mathbf{n}_i \sim \mathcal{N}(0, \Sigma_i)$, where the matrices Σ_i are assumed diagonal and positive. This contrasts with MultiView ICA: the individual noise variances are learned and are not assumed to be the same across components or subjects. We further assume that there are at least 3 views: $m \geqslant 3$.

In contrast to almost all existing works, we assume that some components (possibly all of them) may be Gaussian, and denote $\mathcal{G}$ the set of Gaussian components: $\mathbf{s}_j \sim \mathcal{N}(0, 1)$ for $j \in \mathcal{G}$. The other components are non-Gaussian: for $j \notin \mathcal{G}$, $\mathbf{s}_j$ is non-Gaussian.

IDENTIFIABILITY The parameters of the model are $\Theta = (A_1, \dots, A_m, \Sigma_1, \dots, \Sigma_m)$. We are interested in the identifiability of this model: given observations $\mathbf{x}_1, \dots, \mathbf{x}_m$ generated with parameters Θ, are there some other Θ' that can generate the same observations? Let us consider the following assumption that requires that the individual noises for Gaussian components are sufficiently diverse:

Assumption 14 (Noise diversity in Gaussian components)**.** *For all $j, j' \in \mathcal{G}, j \neq j'$, the sequences $(\Sigma_{ij})_{i=1\dots m}$ and $(\Sigma_{ij'})_{i=1\dots m}$ are different where Σ_{ij} is the j, j entry of Σ_i*

It is readily seen that there is one trivial set of indeterminacies in the problem: if $P \in \mathbb{R}^{p \times p}$ is a sign and permutation matrix the parameters $(A_1 P, \dots, A_m P, P^\top \Sigma_1 P, \dots, P^\top \Sigma_m P)$ also generate $\mathbf{x}_1, \dots, \mathbf{x}_m$. The following theorem shows that under the above assumption, these are the only indeterminacies of the problem.

Theorem 15 (Identifiability)**.** *We suppose Assumption 14. We let $\Theta' = (A'_1, \dots, A'_m, \Sigma'_1, \dots, \Sigma'_m)$ another set of parameters, and assume that they also generate $\mathbf{x}_1, \dots, \mathbf{x}_m$. Then, there exists a sign and permutation matrix P such that for all i, $A'_i = A_i P$, and $\Sigma'_i = P^\top \Sigma_i P$.*

Proof. By hypothesis, the covariances verify $C_{ij} = \mathbb{E}[\mathbf{x}_i\mathbf{x}_j^\top] = A_i(I_p + \delta_{ij}\Sigma_i)A_j^\top = A_i'(I_p + \delta_{ij}\Sigma_i')A_j'^\top$ for all i, j. We let $P_i = A_i^{-1}A_i'$. The previous relationship for $j \neq i$ gives $P_iP_j^\top = I_p$. Because there are more than 3 views, there is another integer $k \notin \{i, j\}$, and we have $P_iP_k^\top = P_jP_k^\top = I_p$. This shows that $P_i = P_j$: all these matrices are equal, and we call P their common value. The previous equation also gives $PP^\top = I_p$, so P is orthogonal. We have that $s + n_i$ and $s' + n_i'$ have independent components and $s + n^i = P(s' + n_i')$. Lemma 21 in appendix B.1 (a direct consequence of classical ICA results [38], Theorem 10) gives $P = \Pi^{-1}\Omega\Pi'$ where Π and Π' are sign and permutation matrices such that the first g components of $\Pi(s + n_i)$ and $\Pi'(s' + n_i')$ are Gaussian, and Ω is a block diagonal matrix given by

$$\Omega = \begin{bmatrix} \Omega_g & 0 \\ 0 & I_{p-g} \end{bmatrix}$$

where Ω_g is orthogonal. We call $A^{(g)}$ the first $g \times g$ block of a matrix A so that $\Omega^{(g)} = \Omega_g$.

Then, considering only the Gaussian components, we can write for $i = j$: $(\Pi\Sigma_i)^{(g)} = \Omega_g(\Pi'\Sigma_i')^{(g)}\Omega_g^\top$ for all i. This, combined with Assumption 14, implies that Ω_g is a sign and permutation matrix (see Lemma 22 in appendix B.1) and therefore P is a sign and permutation matrix. Then it follows that $I + \Sigma_i = P(I + \Sigma_i')P^\top$ and therefore $\Sigma_i = P\Sigma_i'P^\top$ so $\Sigma_i' = P^\top\Sigma_iP$. □

Identifiability in the Gaussian case is a consequence of the identifiability results in [153] and in the general case, local identifiability results can be derived from the work of [9]. However local identifiability only shows that for a given set of parameters there exists a neighborhood in which no other set of parameters can generate the same observations [135]. In contrast, the proof of Theorem 15 shows global identifiability.

Theorem 15 shows that the task of recovering the parameters from the observations is a well-posed problem, under the sufficient condition of Assumption 14. We also note that Assumption 14 is necessary for identifiability. For instance, if j and j' are two Gaussian components such that $\Sigma_{ij} = \Sigma_{ij'}$ for all i, then a global rotation of the components j, j' yields the same covariance matrices. The current work assumes $m \geqslant 3$. In appendix B.2 we give an identifiability result for $m = 2$.

9.2 ESTIMATION OF COMPONENTS WITH NOISE DIVERSITY VIA JOINT-DIAGONALIZATION

We now consider the computational problem of efficient parameter inference. This section considers components with noise diversity, while the next section deals with non-Gaussian components.

9.2.1 *Fitting ShICA via Multiset CCA*

If we assume that the components are all Gaussian, the covariance of the observations given by $C_{ij} = \mathbb{E}[x_i x_j^\top] = A_i(I_p + \delta_{ij}\Sigma_i)A_j^\top$ are sufficient statistics and methods using only second order information, like Multiset CCA, are candidates to estimate the parameters of the model. Consider the matrix $C \in \mathbb{R}^{pm \times pm}$ containing $m \times m$ blocks of size $p \times p$ such that the block i, j is given by C_{ij}. Consider the matrix D identical to C excepts that the non-diagonal blocks are filled with zeros. Multiset CCA (using the SUMCORR cost function under some constraint as described in section 4.2.1) consists in the following generalized eigenvalue problem:

$$Cu = \lambda Du, \quad \lambda > 0, \quad u \in \mathbb{R}^{pm} \quad . \tag{9.2}$$

Consider the matrix $U = [u^1, \dots, u^p] \in \mathbb{R}^{mp \times p}$ formed by concatenating the p leading eigenvectors of the previous problem ranked in decreasing eigenvalue order. Then, consider U to be formed of m blocks of size $p \times p$ stacked vertically and define $(W_i)^\top$ to be the i-th block. These m matrices are the output of Multiset CCA. We also denote $\lambda_1 \geqslant \dots \geqslant \lambda_p$ the p leading eigenvalues of the problem.

The next theorem shows that we only need $\lambda_1 \dots \lambda_p$ to be distinct for Multiset CCA to solve ShICA:

Assumption 16 (Unique eigenvectors). *$\lambda_1 \dots \lambda_p$ are distinct.*

Theorem 17. *We suppose Assumption 16 (only). Then, there exists a permutation matrix P and scale matrices Γ_i such that $W_i = P\Gamma_i A_i^{-1}$ for all i.*

Proof. Let us denote $W \in \mathbb{R}^{mp \times mp}$ the block diagonal matrix with block i given by $(A_i)^{-1}$. We have $Cu = \lambda Du \iff WCW^\top z = \lambda WDW^\top z$ where $u = W^\top z$. We call z a reduced eigenvector. Each block in $WCW^\top$ and in $WDW^\top$ is diagonal so any reduced eigenvector $z = \begin{bmatrix} z_1 \\ \vdots \\ z_m \end{bmatrix}$ is such that the matrix $Z = [z_1 \dots z_m]$ has exactly one non-zero line. Following Lemma 23 in appendix B.1, the first p leading reduced eigenvectors $z^1, \dots, z^p$ all have different first non-zero coordinates. Therefore the concatenation of the first p leading reduced eigenvectors is given by $[z^1, \dots z^p] = \begin{bmatrix} \Gamma_1 \\ \vdots \\ \Gamma_m \end{bmatrix} P^\top$ where $P^\top \in \mathbb{R}^{p \times p}$ is a permutation matrix and $\Gamma_i \in \mathbb{R}^{p \times p}$ is a diagonal matrix. Therefore, the first p eigenvectors are given by

$$[\mathbf{u}^1 \dots \mathbf{u}^p] = \begin{bmatrix} W_1^\top \\ \vdots \\ W_m^\top \end{bmatrix} = \begin{bmatrix} (A_1^{-1})^\top \Gamma_1 P^\top \\ \vdots \\ (A_m^{-1})^\top \Gamma_m P^\top \end{bmatrix} \text{ and so } W_i = P\Gamma_i A_i^{-1} \qquad \square$$

This theorem means that solving the generalized eigenvalue problem (9.2) allows to recover the mixing matrices up to a scaling and permutation: this form of generalized CCA recovers the parameters of the statistical model. Note that Assumption 16 is also a necessary condition. Indeed, if two eigenvalues are identical, the eigenvalue problem is not uniquely determined.

Note that the link between Multiset CCA and probabilistic models has been studied in other contexts [92] or [5] although the authors do not use the same formulation of MultisetCCA as ours. We highlight here, that the different versions of Multiset CCA are not equivalent and may very well produce different results.

We have two different Assumptions, 14 and 16, the first of which guarantees theoretical identifiability as per Theorem 15 and the second guarantees consistent estimation by Multiset CCA as per Theorem 17. Next we will discuss their connections, and show some limitations of the Multiset CCA approach. To begin with, we have the following result about the eigenvalues of the problem (9.2) and the Σ_{ij}.

Proposition 18. *For $j \leqslant p$, let λ_j the largest solution of $\sum_{i=1}^{m} \frac{1}{\lambda_j(1+\Sigma_{ij})-\Sigma_{ij}} = 1$. Then, $\lambda_1, \dots, \lambda_p$ are the p largest eigenvalues of problem (9.2).*

It is easy to see that we then have $\lambda_1, \dots, \lambda_p$ greater than 1, while the remaining eigenvalues are lower than 1. From this proposition, two things appear clearly. First, Assumption 16 implies Assumption 14. Indeed, if the λ_j's are distinct, then the sequences $(\Sigma_{ij})_i$ must also be different from the previous proposition. This is expected as from Theorem 17, Assumption 16 implies identifiability, which in turn implies Assumption 14.

Proposition 18 also allows us to derive cases where Assumption 14 holds but not Assumption 16. The following proposition shows that we can chose parameters of the model so that the model is identifiable but it cannot be solved using Multiset CCA:

Proposition 19. *Assume that for two integers j, j', the sequence $(\Sigma_{ij})_i$ is a permutation of $(\Sigma_{ij'})_i$, i.e. that there exists a permutation of $\{1, \dots, p\}$, π, such that for all i, $\Sigma_{ij} = \Sigma_{\pi(i)j'}$. Then, $\lambda_j = \lambda_{j'}$.*

In this setting, Assumption 14 holds so ShICA is identifiable, while Assumption 16 does not hold, so Multiset CCA cannot recover the unmixing matrices.

9.2.2 *Sampling noise and improved estimation by joint diagonalization*

The consistency theory for Multiset CCA developed above is conducted under the assumption that the covariances C_{ij} are the true covariances of the model, and not approximations obtained from observed samples. In practice, however, a serious limitation of Multiset CCA is that even a slight error of estimation on the covariances, due to "sampling noise", can yield a large error in the estimation of the unmixing matrices, as will be shown next.

We begin with an empirical illustration. We take $m = 3$, $p = 2$, and Σ_i such that $\lambda_1 = 2 + \varepsilon$ and $\lambda_2 = 2$ for $\varepsilon > 0$. In this way, we can control the *eigen-gap* of the problem, ε. We take W_i the outputs of Multiset CCA applied to the true covariances C_{ij}. Then, we generate a perturbation $\Delta = \delta \cdot S$, where S is a random positive symmetric $pm \times pm$ matrix of norm 1, and $\delta > 0$ controls the scale of the perturbation. We take Δ_{ij} the $p \times p$ block of Δ in position (i, j), and $\tilde{W}_i$ the output of Multiset CCA applied to the covariances $C_{ij} + \Delta_{ij}$. We finally compute the sum of the Amari distance between the W_i and $\tilde{W}_i$: the Amari distance measures how close the two matrices are, up to scale and permutation [6]. Fig 9.1 displays the median Amari distance over 100 random repetitions, as the perturbation scale δ increases. The different curves correspond to different values of the eigen-gap ε. We see clearly that the robustness of Multiset CCA critically depends on the eigen-gap, and when it is small, even a small perturbation of the input (due, for instance, to sampling noise) can lead to large estimation errors.

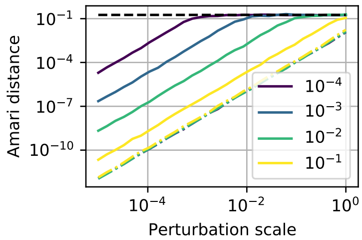


Figure 9.1: Amari distance between true mixing matrices and estimates of Multiset CCA when covariances are perturbed. Different curves correspond to different eigengaps. When the gap is small, a small perturbation can lead to complete mixing. Joint-diagonalization (colored dotted lines) fixes the problem.

This problem is very general and well studied [141]: the mapping from matrices to (generalized) eigenvectors is highly non-smooth. However, the gist of our method is that the *span* of the leading p eigenvectors is smooth, as long as there is a large enough gap between λ_p and λ_{p+1}. For our specific problem we have the following bounds, derived from Prop. 18.

Proposition 20. *We let* $\sigma_{\max} = \max_{ij} \Sigma_{ij}$ *and* $\sigma_{\min} = \min_{ij} \Sigma_{ij}$. *Then,* $\lambda_p \geqslant 1 + \frac{m-1}{1+\sigma_{\max}}$, *while* $\lambda_{p+1} \leqslant 1 - \frac{1}{1+\sigma_{\min}}$.

As a consequence, we have $\lambda_p - \lambda_{p+1} \geqslant \frac{m-1}{1+\sigma_{\max}} + \frac{1}{1+\sigma_{\min}}$: the gap between these eigenvalues increases with m, and decreases with the noise power.

Algorithm 3: ShICA-J

Input: Covariances $\tilde{C}_{ij} = \mathbb{E}[\mathbf{x}_i \mathbf{x}_j^\top]$
$(\tilde{W}_i)_i \leftarrow \text{MultisetCCA}((\tilde{C}_{ij})_{ij})$
$Q \leftarrow \text{JointDiag}((\tilde{W}_i \tilde{C}_{ii} \tilde{W}_i^\top)_i)$
$\Gamma_{ij} \leftarrow Q \tilde{W}_i \tilde{C}_{ij} \tilde{W}_j^\top Q^\top$
$(\Phi_i)_i \leftarrow \text{Scaling}((\Gamma_{ij})_{ij})$
return *Unmixing matrices* $(\Phi_i Q \tilde{W}_i)_i$

In this setting, when the magnitude of the perturbation Δ is smaller than $\lambda_p - \lambda_{p+1}$, [141] indicates that $\text{Span}([W_1, \dots, W_m]^\top) \simeq \text{Span}([\tilde{W}_1, \dots, \tilde{W}_m]^\top)$, where $[W_1, \dots, W_m]^\top \in \mathbb{R}^{pm \times p}$ is the vertical concatenation of the W_i's. In turn, this shows that there exists a matrix $Q \in \mathbb{R}^{p \times p}$ such that

$$W_i \simeq Q \tilde{W}_i \quad \text{for all} \quad i. \tag{9.3}$$

We propose to use joint-diagonalization to recover the matrix Q. Given the $\tilde{W}_i$'s, we consider the set of symmetric matrices $\tilde{K}_i = \tilde{W}_i \tilde{C}_{ii} \tilde{W}_i^\top$, where $\tilde{C}_{ii}$ is the contaminated covariance of $\mathbf{x}_i$. Following Eq. (9.3), we have $Q \tilde{K}_i Q^\top = W_i \tilde{C}_{ii} W_i^\top$, and using Theorem 17, we have $Q \tilde{K}_i Q^\top = P \Gamma_i A_i^{-1} \tilde{C}_{ii} A_i^{-\top} \Gamma_i P^\top$. Since $\tilde{C}_{ii}$ is close to $C_{ii} = A_i (I_p + \Sigma_i) A_i^\top$, the matrix $P \Gamma_i A_i^{-1} \tilde{C}_{ii} A_i^{-\top} \Gamma_i P^\top$ is almost diagonal. In other words, the matrix Q is an approximate diagonalizer of the $\tilde{K}_i$'s, and we approximate Q by joint-diagonalization of the $\tilde{K}_i$'s. In Fig 9.1, we see that this procedure mitigates the problems of multiset-CCA, and gets uniformly better performance regardless of the eigen-gap. In practice, we use a fast joint-diagonalization algorithm [2] to minimize a joint-diagonalization criterion for positive symmetric matrices [119]. The estimated unmixing matrices

$$U_i = Q \tilde{W}_i \tag{9.4}$$

correspond to the true unmixing matrices only up to some scaling: the information that the components are of unit variance is lost.

Scale estimation We form the matrices $\Gamma_{ij} = U_i \tilde{C}_{ij} U_j^\top$. In order to estimate the scalings, we solve

$$\mathcal{L}_\Phi = \min_{(\Phi_i)} \sum_{i \neq j} \| \Phi_i \operatorname{diag}(\Gamma_{ij}) \Phi_j - I_p \|_F^2 \tag{9.5}$$

where the Φ_i are diagonal matrices. The gradient is given by

$$\frac{\partial \mathcal{L}_\Phi}{\partial \Phi_i} = 2 \sum_{j \neq i} (\Phi_i \operatorname{diag}(Y_{ij}) \Phi_j - I_p) \Phi_j \tag{9.6}$$

Therefore we get

$$\frac{\partial \mathcal{L}_\Phi}{\partial \Phi_i} = 0 \tag{9.7}$$
$$\iff 2\sum_{j\neq i}(\Phi_i \operatorname{diag}(Y_{ij})\Phi_j - I_p)\Phi_j = 0 \tag{9.8}$$
$$\iff \Phi_i \sum_{j\neq i} \operatorname{diag}(Y_{ij})\Phi_j^2 - \sum_{j\neq i}\Phi_j = 0 \tag{9.9}$$
$$\iff \Phi_i = \frac{\sum_{j\neq i}\Phi_j}{\sum_{j\neq i}\operatorname{diag}(Y_{ij})\Phi_j^2} \tag{9.10}$$

We then iterate formula (9.10) over i until convergence. The final estimates of the unmixing matrices are given by

$$(\Phi_i U_i)_{i=1}^m \tag{9.11}$$

The full procedure, called ShICA-J, is summarized in Algorithm 3.

9.2.3 *Estimation of noise covariance and inference of shared components*

In practice, it is important to estimate noise co-variances Σ_i in order to take advantage of the fact that some views are noisier than others. As it is well known in classical factor analysis, modelling noise variances allows the model to virtually discard variables, or subjects, that are particularly noisy.

Using the ShICA model with Gaussian components, we derive noise covariances estimate directly from maximum likelihood. We use an expectation-maximization (EM) algorithm, which is especially fast because noise updates are in closed-form. Following derivations given in appendix B.3.1, the sufficient statistics in the E-step are given by

$$\mathbb{E}[\mathbf{s}|\mathbf{x}] = \left(\sum_{i=1}^m \Sigma_i^{-1} + I\right)^{-1} \sum_{i=1}^m (\Sigma_i^{-1}\mathbf{y}_i) \tag{9.12}$$
$$\mathbb{V}[\mathbf{s}|\mathbf{x}] = (\sum_{i=1}^m \Sigma_i^{-1} + I)^{-1} \tag{9.13}$$

Incorporating the M-step we get the following updates that only depend on the covariance matrices:

$$\Sigma_i \leftarrow \operatorname{diag}(\hat{C}_{ii} - 2\mathbb{V}[\mathbf{s}|\mathbf{x}]\sum_{j=1}^m \Sigma_j^{-1}\hat{C}_{ji} + \mathbb{V}[\mathbf{s}|\mathbf{x}]\sum_{j=1}^m\sum_{l=1}^m \left(\Sigma_j^{-1}\hat{C}_{jl}\Sigma_l^{-1}\right)\mathbb{V}[\mathbf{s}|\mathbf{x}] + \mathbb{V}[\mathbf{s}|\mathbf{x}]) \tag{9.14}$$

9.3 SHICA-ML: MAXIMUM LIKELIHOOD FOR NON-GAUSSIAN COMPONENTS

ShICA-J only uses second order statistics. However, the ShICA model (9.1) allows for non-Gaussian components. We now propose an algorithm for fitting the ShICA model that assumes non-Gaussian components so that it can separate Gaussian and non-Gaussian components. We estimate the parameters by maximum likelihood. Since most non-Gaussian components in real data are super-Gaussian, we assume that the non-Gaussian components $\mathbf{s}$ have the super-Gaussian density

$$p(s_j) = \frac{1}{2}\left(\mathcal{N}(s_j; 0, \frac{1}{2}) + \mathcal{N}(s_j; 0, \frac{3}{2})\right) \tag{9.15}$$

We propose to maximize the expected log-likelihood using a generalized EM [44, 106]. Derivations are available in Appendix B.4. Like in the previous section, the E-step is in closed-form yielding the following sufficient statistics:

$$\mathbb{E}[s_j|\mathbf{x}] = \frac{\sum_{\alpha \in \{\frac{1}{2}, \frac{3}{2}\}} \theta_\alpha \frac{\alpha \bar{y}_j}{\alpha + \bar{\Sigma}_j}}{\sum_{\alpha \in \{0.5, 1.5\}} \theta_\alpha} \tag{9.16}$$

$$\mathbb{V}[s_j|\mathbf{x}] = \frac{\sum_{\alpha \in \{\frac{1}{2}, \frac{3}{2}\}} \theta_\alpha \frac{\bar{\Sigma}_j \alpha}{\alpha + \bar{\Sigma}_j}}{\sum_{\alpha \in \{0.5, 1.5\}} \theta_\alpha} \tag{9.17}$$

where $\theta_\alpha = \mathcal{N}(\bar{y}_j; 0, \bar{\Sigma}_j + \alpha)$, $\bar{y}_j = \frac{\sum_i \Sigma_{ij}^{-1} y_{ij}}{\sum_i \Sigma_{ij}^{-1}}$ and $\bar{\Sigma}_j = (\sum_i \Sigma_{ij}^{-1})^{-1}$ with $\mathbf{y}_i = W_i \mathbf{x}_i$. Noise updates are in closed-form and given by:

$$\Sigma_i \leftarrow \mathrm{diag}(\mathbb{E}[(\mathbf{y}_i - \mathbb{E}[\mathbf{s}|\mathbf{x}])(\mathbf{y}_i - \mathbb{E}[\mathbf{s}|\mathbf{x}])^\top]) + \mathbb{V}[\mathbf{s}|\mathbf{x}]) \tag{9.18}$$

However, no closed-form is available for the updates of unmixing matrices. We therefore perform quasi-Newton updates given by

$$W_i \leftarrow (I - \rho(\widehat{\mathcal{H}^{W_i}})^{-1} \mathcal{G}^{W_i}) W_i \tag{9.19}$$

where $\rho \in \mathbb{R}$ is chosen by backtracking line-search

$$\widehat{\mathcal{H}^{W_i}_{a,b,c,d}} = \delta_{ad}\delta_{bc} + \delta_{ac}\delta_{bd}\frac{\mathbb{E}[(y_{ib})^2]}{\Sigma_{ia}} \tag{9.20}$$

is an approximation of the Hessian of the expected negative complete log-likelihood and

$$\mathcal{G}^{W_i} = -I + (\Sigma_i)^{-1}\mathbb{E}[(\mathbf{y}_i - \mathbb{E}[\mathbf{s}|\mathbf{x}])(\mathbf{y}_i)^\top] \tag{9.21}$$

is the gradient.

We alternate between computing the statistics $\mathbb{E}[\mathbf{s}|\mathbf{x}]$, $\mathbb{V}[\mathbf{s}|\mathbf{x}]$ (E-step) and updates of parameters Σ_i and W_i for $i = 1 \dots m$ (M-step). Let us highlight that our EM algorithm and in particular the E-step resembles the one used in [104]. However because they assume noise on the sensors and not on the components, their formula for $\mathbb{E}[\mathbf{s}|\mathbf{x}]$ involves a sum with 2^p terms whereas we have only 2 terms. The resulting method is called ShICA-ML.

MINIMUM MEAN SQUARED ERROR ESTIMATES IN SHICA In ShICA-J as well as in ShICA-ML, we have a closed-form for the expected components given the data $\mathbb{E}[\mathbf{s}|\mathbf{x}]$, shown in equation (9.12) and (9.16) respectively. This provides minimum mean squared error estimates of the shared components, and is an important benefit of explicitly modelling shared components in a probabilistic framework.

9.4 RELATED WORK

ShICA combines theory and methods coming from different branches of "component analysis". It can be viewed as a GroupICA method, as an extension of Multiset CCA, as an Independent Vector Analysis method or, crucially, as an extension of SRM. In the setting studied here, ShICA improves upon all existing methods.

ShICA inherits from all the advantages of MultiView ICA. Unlike CanICA or ConcatICA (see section 4.2.2.1), it optimizes a proper likelihood and unlike tensorial methods [18] or SRM, it does not assume any structure on the mixing matrices.

Compared to the likelihood based method of Guo presented in section 4.2.2.2 or to MultiView ICA, ShICA allows different subjects to have different noise variances. This last point is crucial as it allows to separate Gaussian components. In addition, the estimation in ShICA-ML relies on a very efficient closed form E-step. This differs from the likelihood based method of Guo that does not have a closed form E-step and has to rely on a first order approximation.

ShICA-J uses the Multiset CCA presented in section 4.2.1 which is one of the fastest as it reduces to solving a generalized eigenvalue problem. The fact that CCA solves a well defined probabilistic model has first been studied in [13] where it is shown that CCA is identical to multiple battery factor analysis [26] (restricted to 2 views). This latter formulation differs from our model in that the noise is added on the sensors and not on the components which makes the model unidentifiable. Identifiable variants and generalizations can be obtained by imposing sparsity on the mixing matrices such as in [11, 81, 156] or non-negativity [42]. Lastly, the work in [92] exhibits a set of sufficient (but not necessary) conditions under which a well defined model can be learnt by the formulation of Multiset CCA used in ShICA-J. The set of conditions we exhibit in this work are necessary and sufficient. We further emphasize that basic Multiset CCA provides a poor estimator, as explained in section 9.2.2.

Let us highlight that ShICA can be seen as a particular instance of IVA where subject specific components s_{ij} are such that $s_{ij} = s_j + n_{ij}$. However, current implementations such as IVA-L [85], IVA-G [7], IVA-L-SOS [21], IVA-GGD [9] or IVA with Kotz distribution [8] estimate view-specific components but shared components are not modeled explicitly. Studying the components post IVA can give great insights about which components are shared and by which subjects (see [94]) and individual responses that are closed enough could be merged to produce a shared response. However, the flexibility in modeling different components from each subject comes at the cost of not having a well principled way of aggregating the components into a common shared response. In contrast, ShICA specifically enables extraction of shared components from the subject specific components via its minimum mean squared error estimate.

The IVA theory provides global identifiability conditions in the Gaussian case (IVA-G) [153] and local identifiability conditions in the general case [9] from which local identifiability conditions of ShICA could be derived. However, in this work, we provide global identifiability conditions for ShICA. Lastly, IVA can be performed using joint diagonalization of cross covariances [40, 91] although multiple matrices have to be learnt and cross-covariances are not necessarily symmetric positive definite, which makes the algorithm slower and less principled.

Let us point out that extracting a shared response from multiple dataset is also the goal of SRM. Some deep variants [37] release the orthogonality constrain but they are much more computationally demanding. ShICA leverages ICA theory to provide a much more powerful model of shared responses.

LIMITATIONS The main limitation of this work is that the model does not reduce the dimension inside each view. In line with other methods, such view-specific dimension reduction has to be done by some external method, typically view-specific PCA. Using specialized methods for the estimation of covariances should also be of interest for ShICA-J, where it only relies on sample covariances. Finally, ShICA-ML uses a simple model of a super-Gaussian distribution, while modeling the non-gaussianities in more detail in ShICA-ML should improve the performance.

9.5 CONCLUSION

In this chapter, we introduced the ShICA model as a principled unifying solution to the problems of shared response modelling and GroupICA. ShICA is able to use both the diversity of Gaussian variances and non-Gaussianity for optimal estimation. We presented two algorithms to fit the model: ShICA-J, a fast algorithm that uses noise diversity, and ShICA-ML, a maximum likelihood approach that can separate Gaussian and non-Gaussian components. ShICA algorithms come with principled procedures for shared components estimation, as well as adaptation and estimation of noise levels in each view (subject) and component. In the next chapter, we evaluate ShICA using both real and synthetic data and compare with competitive approaches.

10 SHARED ICA IN PRACTICE

In the previous chapter, we have introduced the ShICA model, a principled unifying solution to the problems of shared response modeling and GroupICA. In this chapter we evaluate its practical utility on synthetic data and on brain imaging data.

10.1 SYNTHETIC EXPERIMENT

In the following synthetic experiments, data are generated according to model (9.1) with $p = 4$ components and $m = 5$ views and mixing matrices are generated by sampling coefficients from a standardized Gaussian.

10.1.1 *Separation performance: different use cases*

Gaussian components are generated from a standardized Gaussian and their noise has standard deviation $\Sigma_i^{\frac{1}{2}}$ where $\Sigma_i^{\frac{1}{2}}$ is a diagonal matrix. The diagonal coefficients of $\Sigma_i^{\frac{1}{2}}$ are obtained by sampling from a uniform density between 0 and 1. Non-Gaussian components are generated from a Laplace distribution and their noise standard deviations are equal. We study 3 cases where either all components are Gaussian, all components are non-Gaussian or half of the components are Gaussian and half are non-Gaussian. We vary the number of samples n between 10^2 and 10^5 and display in Fig 10.1 the mean Amari distance across subjects as a function of n. The experiment is repeated 100 times using different seeds. We report the median result and error bars represent the first and last deciles. When all components are Gaussian (Fig. 10.1 (a)), CanICA cannot separate the components at all. In contrast ShICA-J, ShICA-ML and Multiset CCA are able to separate them, but Multiset CCA needs many more samples

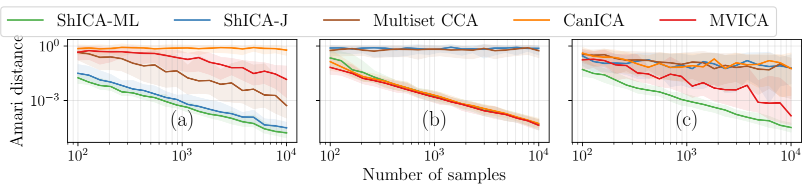


Figure 10.1: **Separation performance**: Algorithms are fit on data following model 9.1 **(a)** Gaussian components with noise diversity **(b)** Non-Gaussian components without noise diversity **(c)** Half of the components are Gaussian with noise diversity, the other half is non-Gaussian without noise diversity.

to reach the same Amari distance as ShICA-J or ShICA-ML, which shows that correcting for the rotation due to sampling noise improves the results. Looking at error bars, we also see that the performance of Multiset CCA varies quite a lot with the random seeds: this shows that depending on the sampling noise, the rotation can be very different from identity. MultiViewICA does achieve separation but obtains relatively poor separation performance compared to ShICA or Multiset CCA. When none of the components are Gaussian (Fig. 10.1 (b)), only CanICA, ShICA-ML and MultiView ICA are able to separate the components, as other methods do not make use of non-Gaussianity. Finally, in the hybrid case (Fig. 10.1 (c)), ShICA-ML is able to separate the components very well as it can make use of both non-Gaussianity and noise diversity. As we can see, MultiView ICA yields decent performance though uniformly worse than ShICA-ML. Also note that error bars are very large showing that for some seeds it gives poor results. Overall, MultiView ICA is a lot less reliable than ShICA-ML.

10.1.2 *Separation performance in function of non-Gaussianity*

We generate data according to model (9.1). Components $\mathbf{s}$ are generated using $s_j = d(x)$ with $d(x) = x|x|^{\alpha-1}$ and $x \sim \mathcal{N}(0, 1)$. We impose noise diversity: the noise of view i has standard deviation $\Sigma_i^{\frac{1}{2}}$ (obtained by sampling from a uniform density between 0 and 1). Mixing matrices A_i are generated by sampling their coefficients from a standardized Gaussian law. The number of samples is fixed to $n = 10^5$ and we vary α between 0.8 and 2. Each experiment is repeated 40 times using different seeds in the random number generator. We use $p = 4$ components and $m = 5$ views. We display in Fig 10.2 the mean Amari distance across subjects. The experiment is repeated 100 times using different seeds. We report the median result and error bars represent the first and last deciles. When α is close to 1 (components are almost Gaussian), ShICA-J, ShICA-ML and multiset CCA can separate components well (but multiset CCA reaches higher Amari distance than ShICA). In this regime, MultiViewICA yields much higher Amari distance than ShICA-J, ShICA-ML or Multiset CCA but is still better than CanICA which cannot separate components at all. As non-Gaussianity (α) increases, ICA based methods yield better results but ShICA-ML yields uniformly lower Amari distance.

10.1.3 *Computation time*

We generate components using a slightly super Gaussian density: $s_j = d(x)$ with $d(x) = x|x|^{0.2}$ and $x \sim \mathcal{N}(0, 1)$. We vary the number of samples n between 10^2 and 10^4. We compute the mean Amari distance across subjects and record the computation time. The experiment is repeated 40 times. We plot the Amari distance as a function of the

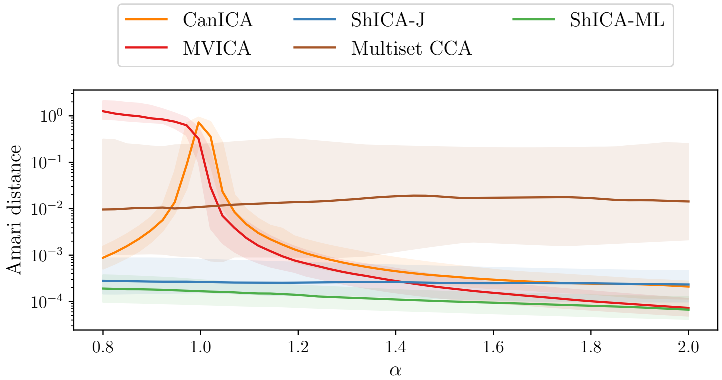


Figure 10.2: **Separation performance in function of non-Gaussianity** Separation performance of algorithms for sub-Gaussian $\alpha < 1$ and super-Gaussian $\alpha > 1$ components

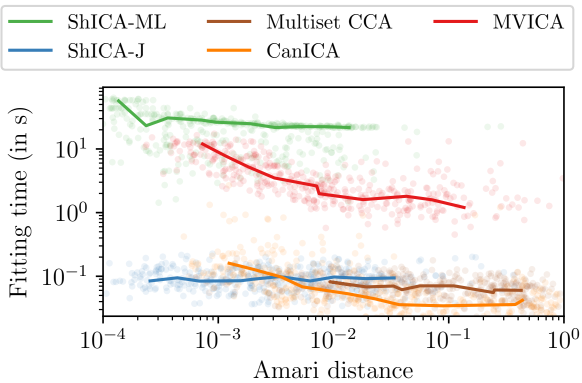


Figure 10.3: **Computation time:** Algorithms are fit on data generated from model (9.1) with a super-Gaussian density. For different values of the number of samples, we plot the Amari distance and the fitting time. Thick lines link median values across seeds.

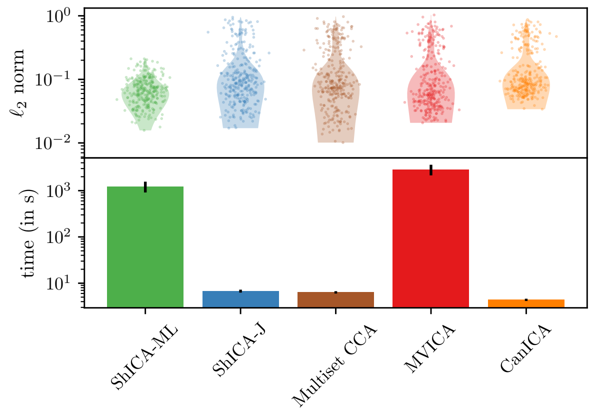


Figure 10.4: **Robustness w.r.t intra-subject variability in MEG**: (**top**) ℓ_2 distance between shared components corresponding to the same stimuli in different trials. (**bottom**) Fitting time.

computation time in Fig 10.3. Each point corresponds to the Amari distance/computation time for a given number of samples and a given seed. We then consider for a given number of samples, the median Amari distance and computation time across seeds and plot them in the form of a thick line. From Fig 10.3, we see that ShICA-J is the method of choice when speed is a concern while ShICA-ML yields the best performance in terms of Amari distance at the cost of an increased computation time. The thick lines for ShICA-J and Multiset CCA are quasi-flat, indicating that the number of samples does not have a strong impact on the fitting time as these methods only work with covariances. On the other hand CanICA or MultiviewICA computation time is more sensitive to the number of samples.

10.2 EXPERIMENTS ON BRAIN IMAGING DATA

The shape of the input data (v, t, m) depend on the dataset used and the experiment performed. A description of the datasets used is available in section 3.1.5 for fMRI data and in section 3.2.5 for MEG data. The values for p (the number of components) is given in the descriptions of the experiments. Note that in the special case of the experiments on the sherlock fMRI dataset, the experimental pipeline is available on github `https://github.com/hugorichard/multiviewica/tree/master/real_data_experiments` (including downloading and preprocessing of the data).

10.2.1 *Robustness w.r.t intra-subject variability in MEG*

In the following experiments we consider the Cam-CAN dataset [146]. We use the magnetometer data from the MEG of $m = 100$ subjects chosen randomly among 496. Each subject is repeatedly presented three audio-visual stimuli. For each stimulus, we divide the trials into two sets and within each set, the MEG signal is averaged across trials to isolate the evoked response. This procedure yields 6 chunks of individual data (2 per stimulus). We study the similarity between shared components corresponding to repetitions of the same stimulus. This gives a measure of robustness of each ICA algorithm with respect to intra-subject variability. Data are first reduced using a subject-specific PCA with $p = 10$ components. Algorithms are run 10 times with different seeds on the 6 chunks of data, and shared components are extracted. When two chunks of data correspond to repetitions of the same stimulus they should yield similar components. For each component and for each stimulus, we therefore measure the ℓ_2 distance between the two repetitions of the stimulus. This yields 300 distances per algorithm that are plotted on Fig 10.4.

The components recovered by ShICA-ML have a much lower variability than other approaches. The performance of ShICA-J is competitive with Multiview ICA while being much faster to fit. Multiset CCA yields satisfying results compared with ShICA-J. However we see that the number of components that do not match at all across trials is greater in Multiset CCA.

The mixing operators in ShICA define spatial maps. In appendix B.5, we plot the average spatial maps across subjects.

10.2.2 *MEG Phantom experiment*

10.2.2.1 *Elekta Phantom*

We use the *Elektra Phantom MEG* dataset described in 3.2.5 where dipoles at $m = 32$ different locations emit the same signal. We reduce the data by applying view specific PCA with $k = 20$ components and algorithms are applied on the reduced data. We select the component that is closer to the true one and compute the L2 norm between the predicted component and the true one after normalization. Then we attempt to recover the position of each dipole by performing dipole fitting on the mixing operator of each view (using only the column corresponding to the true component). The localization error is defined as the mean l2 distance between the true localization and the predicted localization where the mean is computed across dipoles. Each epoch corresponds to 301 samples and 20 epochs are available in total. We vary the number of epochs between 2 and 18 and display in Fig 10.5 the reconstruction error and the localization error as a function of the number of epochs used. ShICA-ML outperforms other methods. ShICA-J gives satisfying results while being much faster.

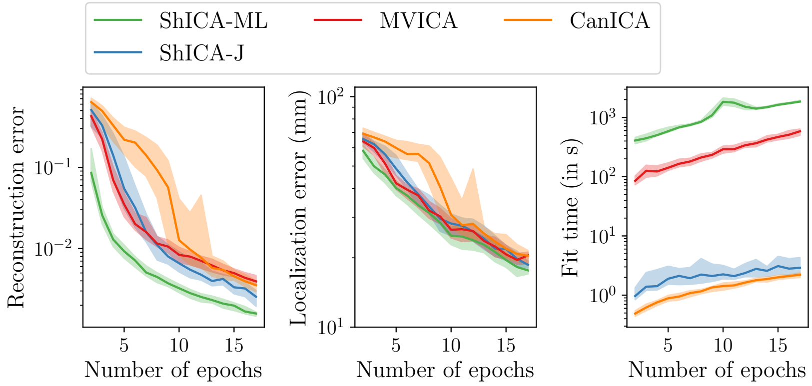


Figure 10.5: **MEG Phantom (Elekta)**: (left) L2 distance between the predicted and actual component (middle) Mean error (in mm) between predicted and actual dipoles localization (right) Fitting time (in seconds)

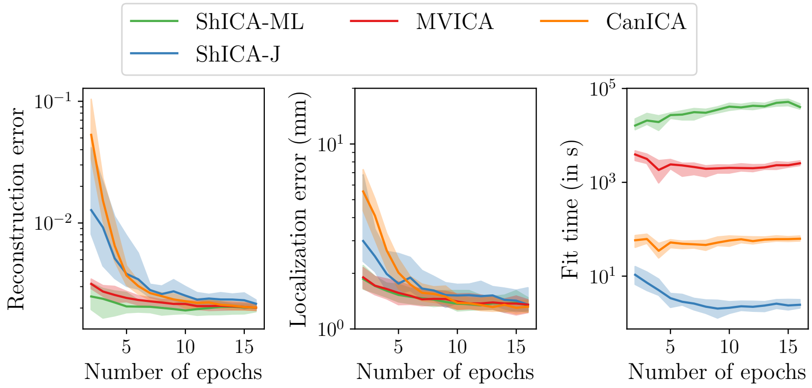


Figure 10.6: **MEG Phantom Sinusoidal components**: (left) L2 distance between the predicted and actual component (middle) Mean error (in mm) between predicted and actual dipoles localization (right) Fitting time (in seconds)

10.2.2.2 *MEG Phantom Sinusoidal components*

We reproduce the phantom experiment presented in section 8.4. ShICA-ML outperforms other methods. ShICA-J gives satisfying results while being much faster.

10.2.3 *Reconstructing the BOLD signal of missing subjects*

We reproduce the experiental pipeline described in section 8.3.1. ShICA-ML yields the best R^2 score in all datasets and for any number of components. ShICA-J yields competitive results with respect to Multiview ICA while being much faster to fit.

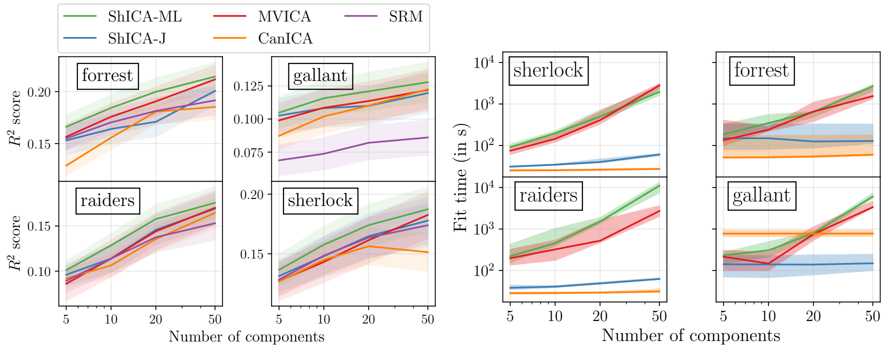


Figure 10.7: **Reconstructing the BOLD signal of missing subjects**. (**left**) Mean R^2 score between reconstructed data and true data. (**right**) Fitting time.

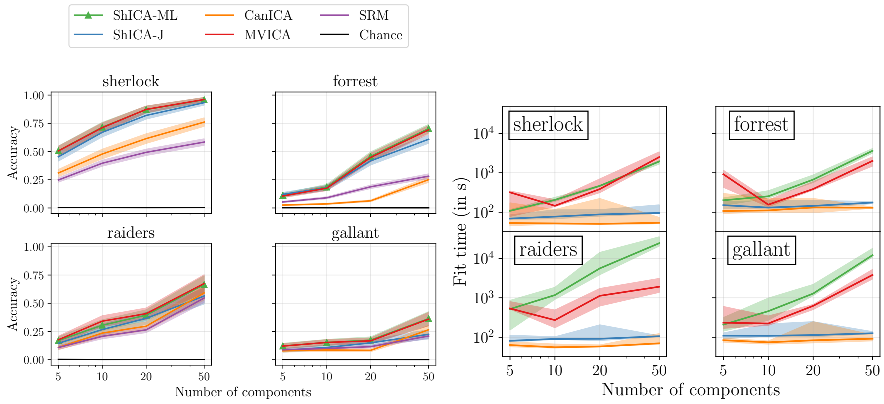


Figure 10.8: **Timesegment matching experiment**: (left) Accuracy (right) Fitting time (in seconds)

10.2.4 *fMRI timesegment matching experiment*

A popular benchmark especially in the SRM community is the timesegment matching experiment [36] which we describe in section 6.2.1. The left panel in Fig 10.8 shows that ShICA-ML, MultiViewICA and ShICA-J yield almost equal accuracy and outperform other methods by a large margin. The right panel in Fig 10.8 shows that ShICA-J is much faster to fit than MultiViewICA or ShICA-ML.

10.3 CONCLUSION

In this chapter, we have shown the practical benefits of ShICA. On simulated data, ShICA clearly outperforms all competing methods in terms of the trade-off between statistical accuracy and computation time. On brain imaging data, ShICA gives more stable decompositions for comparable computation times, and more accurately predicts the data of one subject from the data of other subjects, making it a good candidate to perform transfer learning. As ShiCA only involves linear transforms, decisions based on its output are easier to interpret, making it accessible to practitioners.

In the next section, we show that ICA can be used to perform data augmentation in fMRI.

Part V

CONDICA

11 CONDICA THEORY

In this chapter, we present an ICA-based method to achieve data augmentation for fMRI data.

Advances in computational cognitive neuroimaging research are related to the availability of large amounts of labeled brain imaging data, since classifiers used to decode brain maps have large sample-complexity. However, such data are scarce.

To tackle this problem, data generation is an attractive approach, as it could potentially compensate for the shortage of data. Conditional Generative Adversarial Networks (CGANs) are promising generative models [58] designed for computer vision. However, such improvements have not yet carried over to brain imaging. A likely reason is that CGANs are ill-suited to the noisy, high-dimensional and small-sample data available in functional neuroimaging. Furthermore the training of CGANs is notoriously unstable and there are many hyper-parameters to tune.

In this work, we introduce Conditional ICA: a novel data augmentation technique using ICA together with conditioning mechanisms to generate surrogate brain imaging data and improve image classification performance. Conditional ICA benefits from the abundant resting state data and can be trained with only few labeled samples.

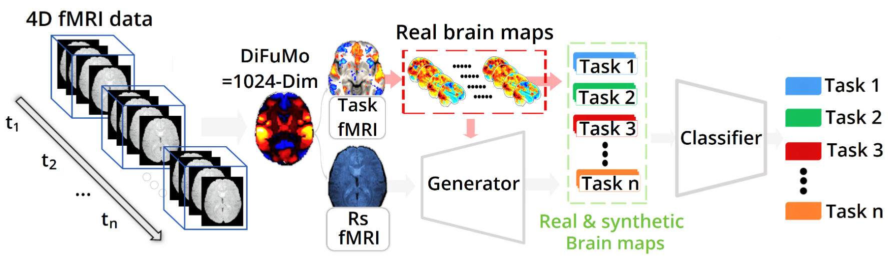


Figure 11.1: **Conditional ICA approach.** Our method aims to generate surrogate data from Task and Rest fMRI data by synthesizing statistical maps that qualitatively fit the distribution of the original maps. These can be used to improve the accuracy of machine learning models that identify contrasts from the corresponding brain activity patterns.

11.1 METHODS

11.1.1 *Spatial Dimension reduction*

The outline of the proposed approach is presented in Fig.11.1. While brain maps are high-dimensional, they span a smaller space than that of the voxel grid. For the sake of tractability, we reduce the dimension of the data by projecting the voxel values on the high-resolution version of the Dictionaries of Functional Modes *DiFuMo* atlas [41], i.e. with $p = 1024$ components. The choice of dimension reduction technique generally has an impact on the results. However, we consider this question to be out of the scope of the current study and leave this to future work.

11.1.2 *A generative model for task data*

Consider a task dataset X^{task} in $\mathbb{R}^{p,n}$ where n is the number of observations (samples) and $p = 1024$ the number of components in the atlas. X^{task} can be seen as n observations of a random vector $\mathbf{x}^{task} \in \mathbb{R}^p$. Let us consider how to learn the distribution of $\mathbf{x}^{task}$. Assuming a Gaussian distribution is standard in this setting, yet, as shown later, it misses key distributional features. Moreover, we consider a model that subsumes the distribution of any type of fMRI data (task or rest): a linear mixture of $k \leqslant p$ independent temporal signals. We therefore use temporal ICA to learn a dimension reduction and unmixing matrix $W^{task} \in \mathbb{R}^{k,p}$ such that the k components of $W^{task}\mathbf{x}^{task}$ are as independent as possible.

A straightforward method to generate new task data would be to independently sample them from the distribution of the components. This is easy because such distribution has supposedly independent marginals. We apply an invertible quantile transform q^{task} to the components of $W^{task}\mathbf{x}^{task}$ so that the distribution of $q^{task}(W^{task}\mathbf{x}^{task})$ has standardized Gaussian marginals. Since it also has independent marginals, it is given by $\mathcal{N}(0_k, I_k)$ from which we can easily sample.

However task datasets have a small number of samples ($10 \sim 10^2$). As a result, there are too few samples to learn a high quality unmixing matrix. In contrast, resting state datasets have a large number of samples ($10^4 \sim 10^5$). Therefore, we replace the unmixing matrix learned on task data W^{task} by the unmixing matrix learned on resting state data W^{rest}.

We form $\mathbf{z}^{task} = W^{rest}\mathbf{x}^{task}$ and learn its quantile transform q. The encoding model is thus given by:

$$\mathbf{z}^{task} = q(W^{rest}\mathbf{x}^{task}) \tag{11.1}$$

However, the independence assumption no longer holds and thus a latent structure among the marginals of $\mathbf{z}^{task}$ has to be taken into

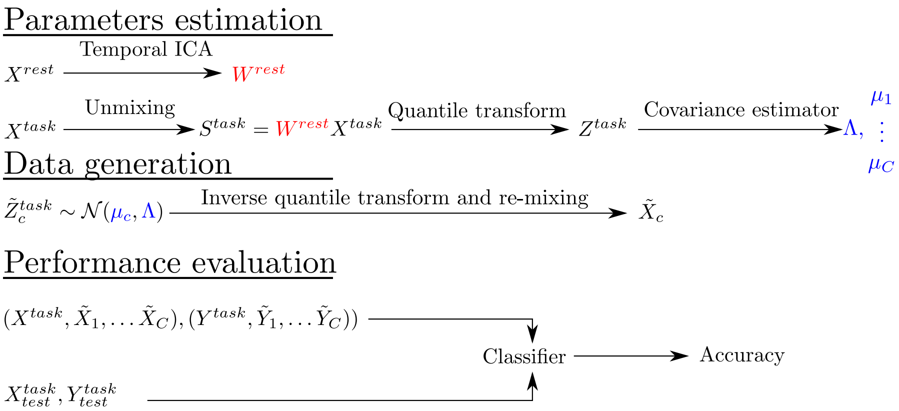


Figure 11.2: **Conditional ICA approach in depth.** The approach proceeds by learning a temporal ICA of rest data $X^{rest} \in \mathbb{R}^{p,n}$, resulting in independent components and unmixing matrix $W^{rest} \in \mathbb{R}^{k,p}$. Applying the unmixing matrix to the task data, we obtain samples in the component space that we map to a normal distribution, yielding $Z^{task} \in \mathbb{R}^{k,n}$. Then, we estimate the covariance $\Lambda \in \mathbb{R}^{k,k}$ (all classes are assumed to have the same covariance) and the class-specific means $\mu_1, \dots, \mu_C \in \mathbb{R}^k$ according to Ledoit-Wolf's method. For each class c, we can draw random samples $\tilde{Z}_c^{task} \in \mathbb{R}^{k,n_{\text{fakes}}}$ from the resulting multivariate Gaussian distribution $\mathcal{N}(\mu_c, \Lambda)$ and obtain fake data $\tilde{X}_c \in \mathbb{R}^{p,n_{\text{fakes}}}$ by applying the inverse quantile transform and re-mixing the data using the pseudo inverse of the unmixing matrix. We append these synthetic data to the actual data to create our new augmented dataset on which we train classifiers.

account. In addition the generative model needs to be conditioned to each class. We therefore assume that the samples in class c, $\mathbf{x}_c^{task}$ are such that:

$$q(W^{rest}\mathbf{x}_c^{task}) \sim \mathcal{N}(\mu_c, \Lambda) \tag{11.2}$$

In order to maximize the number of samples used to learn the parameters of the model, we assume that the quantile transform q and the latent covariance Λ do not depend on the class c. However, the mean μ_c, that can be learned efficiently using just a few tens of samples, depends on class c. Λ is learned using all task samples from a standard shrunk covariance estimator $\Lambda = \Sigma(1-\alpha) + \frac{\alpha}{k}\operatorname{tr}(\Sigma)I_k$ where α is given by the Ledoit-Wolf formula [84] and Σ is the sample covariance of $\mathbf{z}^{task}$.

The generative model of data for brain maps in a certain class c is given by the pseudo inverse of the encoding model:

$$\mathbf{x}_c = (W^{rest})^{\dagger} q^{-1}(\boldsymbol{\epsilon}) \tag{11.3}$$

with $\boldsymbol{\epsilon} \sim \mathcal{N}(\mu_c, \Lambda)$ and $(W^{rest})^{\dagger}$ is the Moore Penrose inverse of W^{rest}

An overview of our generative method is shown in Fig. 11.2.

11.2 RELATED WORK

In image processing, data augmentation is part of standard toolboxes and typically includes operations like cropping, rotation, translation. On fMRI data these methods do not make much sense as brain data are not invariant to such transformations. More advanced techniques [159] are based on generative models such as CGANs or variatonal auto-encoders [80]. Although CGAN-based method are powerful they are slow and difficult to train [12].

Our method is not an adversarial procedure, however it relates to other powerful generative models such as variatonal auto-encoders [80] with which it shares strong similarities. Indeed the analog of the encoding function in the variational auto-encoder is given by $e(\mathbf{x}) = \Lambda^{-\frac{1}{2}} q(W^{rest}\mathbf{x})$ in our model and the analog to the decoding function in the variational auto-encoder is given by $d(\mathbf{z}) = (W^{rest})^{\dagger} q^{-1}(\Lambda^{\frac{1}{2}}\mathbf{z})$ in our model. As in the variational auto-encoder, e approximately maps the empirical data distribution to a standardized Gaussian distribution, while the reconstruction error defined by the difference in l2 norm $\|d(e(\mathbf{x})) - \mathbf{x}\|_2^2$ must remain small. Lastly, another classical generative model related to ours is normalizing flows. We note that when W^{rest} is square (no dimension reduction in ICA), the decoding operator d is invertible (its inverse is e) making our model an instance of normalizing flows [131]. A great property is thus the simplicity and reduced cost of data generation.

11.3 CONCLUSION

In this chapter, we introduced Conditional ICA, a fast generative model for task data. Conditional ICA is essentially a linear generative model with pointwise non-linearity, which makes it cheap, easy to instantiate on new data, and to introspect. In the next chapter, we look at the performance of Conditional ICA on fMRI data.

12

CONDICA IN PRACTICE

In the previous chapter, we have introduced Conditional ICA, an efficient generative model for task fMRI data. In this chapter, we benchmark Conditional ICA as a generative model of task data against various augmentation methods by assessing their ability to improve classification accuracy on a large task fMRI dataset. We find that Conditional ICA yields highest accuracy improvements. In particular, Conditional ICA outperforms conditional GANs [102] while being much easier to optimize and interpret. Lastly, we show on 8 different datasets that the use of Conditional ICA results in systematic improvements in classification accuracy ranging from 1% to 5%.

12.1 DATASET, DATA AUGMENTATION BASELINES AND CLASSIFIERS USED

The unmixing matrices are learned on the rest HCP dataset [149] using 200 subjects. These data were used after standard preprocessing, including linear detrending, band-pass filtering ($[0.01, 0.1]$Hz) and standardization of the time courses. The other 8 datasets [111, 121–123, 127, 137, 149] are obtained from the Neurovault repository [59]. The classes used in each of these datasets correspond to the activation maps related to contrasts (such as "face vs tools") present in the set of tasks of each dataset. In table 12.1, we give references to the datasets used as well as the total number of samples (subjects), the size of train and test sets in each of the cross validation splits and the number of classes in each dataset.

We consider 4 alternative augmentation methods: *ICA*, *Covariance*, *ICA + Covariance* and *CGANs*. When no augmentation method is applied, we use the *Original* label.

The *ICA* method applies ICA to X^{task} to generate unmixing matrices W^{task} and components $S^{task} = W^{task}X^{task}$. To generate a sample $\tilde{\mathbf{x}}_c$ from class c, we sample independently from each component restricted to the samples of class c yielding $\tilde{\mathbf{s}}_c^{task}$ and mix the data: $\tilde{\mathbf{x}}_c = (W^{task})^\dagger \tilde{\mathbf{s}}_c^{task}$.

The *Covariance* method generates a new sample of synthetic data in class c by sampling from a Multivariate Gaussian with mean $\boldsymbol{\mu}_c$ and covariance Σ, where $\boldsymbol{\mu}_c$ is the class mean and Σ is the covariance of centered task data estimated using Ledoit-Wolf method. In brief, it assumes normality of the data per class.

The *ICA + Covariance* method combines the augmentation methods *ICA* and *Covariance*: samples are drawn following the ICA approach,

Dataset	Subjects, classes	Train/Test	Neurovault collection
hcp [149]	787, 23	100/687	4337
cam-can [137]	605, 5	100/505	4342
brainomics [111]	94, 19	50/44	4341
archi [123]	78, 30	40/38	4339
la5c [127]	191, 24	100/91	4343
pinel2012archi [123]	76, 10	40/36	1952
pinel2009twins [121]	65, 12	35/30	1952
pinel2007fast [122]	133, 10	70/63	1952

Table 12.1: **Datasets used in the experiments.** The table provides references to the datasets that were used for our experiments, with the number of subjects, the number of classes, the number of subjects in train and test set in each cross validation split and the collection number in Neurovault

but with some additive non-isotropic Gaussian noise. As in *ICA*, we estimate W^{task} and S^{task} from X^{task} via ICA. Then we consider $R_{task} = X_{task} - W_{task}S_{task}$ and estimate the covariance Σ_R of R_{task} via LedoitWolf's method. We then generate a data sample $\tilde{\mathbf{x}}_c$ from class c as with ICA and add Gaussian noise $\tilde{\mathbf{n}} \sim \mathcal{N}(0, \Sigma_R)$. Samples are thus generated as $\tilde{\mathbf{x}}_c + \tilde{\mathbf{n}}$.

CGANs can generate fake data from a given class. In the CGAN method, the generator and discriminator have a mirrored architecture with 2 fully connected hidden layer of size (256 and 512). The number of epochs, batch size, momentum and learning rate are set to 20k, 16, 0.9, 0.01 and we use the Leaky RELU activation function.

We evaluate the performance of augmentation methods through the use of classifiers: logistic regression (LogReg), linear discriminant analysis with Ledoit-Wold estimate of covariance (LDA), perceptron with two hidden layers (MLP) and random forrests (RF). The hyperparameters in each classifier are optimized through an internal 5-Fold cross validation. We set the number of iterations in each classifier so that convergence is reached. The exact specifications are given in table 12.2.

12.2 COMPARING CLASSIFICATION ACCURACY GAINS ON TASK HCP DATASET

In order to compare the different augmentation methods, we measure their relative benefit in the context of multi-class classification. We use 787 subjects from the HCP task dataset that contains 23 classes and randomly split the dataset into a train set that contains 100 subjects and

Methods	Optimizer	Hyper-parameters
LogReg	L-BFGS (20 000 iterations)	inverse L_2 regularization strength in $\{0.0001, 0.001, 0.01, 0.1, 1\}$
LDA	Least-squares solver	Estimation of covariance using Ledoit-Wolf's method
RF	-	Default parameters in sklearn
MLP	Adam (20 000 iterations, momentum: 0.9, batch size: 32, learning rate: 0.0001)	ReLU activation function, fully connected architecture with two hidden layers both of size 1024, L2 penalty coefficient: 10^{-5}

Table 12.2: **Optimizers and hyper-parameters of classifiers** For each classifier, we give the optimization method used as well as the value of hyper-parameters.

Models	LDA	LogReg	MLP	RF
Original	0.893	0.874	0.779	0.782
ICA	0.814	0.840	0.803	0.778
Covariance	0.895	0.876	0.819	0.780
ICA + Covariance	0.816	0.840	0.815	0.780
CGANs	0.874	0.874	0.726	0.779
Conditional ICA	**0.901**	**0.890**	**0.832**	**0.783**

Table 12.3: **Comparing augmentation methods based on classification accuracy on task HCP dataset** We compare augmentation methods based on the classification accuracy **(Acc)** obtained by 2 linear classifiers (LDA and LogReg) and two non-linear classifier (MLP and RF) trained on augmented datasets on HCP Task fMRI data. We report the mean accuracy across 5 splits.

a test set that contains 687 subjects. In each split, we run augmentation methods on the train set to generate fake samples corresponding to 200 subjects. These samples are then appended to the train set, resulting in an augmented train set on which the classifiers are trained. The results displayed in table 12.3 show that Conditional ICA always yields a higher accuracy than when no augmentation method is applied. The gains are over 1% on all classifiers tested excepts with the random forest classifier which yields much lower accuracy than other methods. By contrast, ICA+Covariance and ICA lead to a decrease in accuracy while the Covariance approach leads to non-significant gains.

In table 12.4, we give the running-time of the CGAN method and Conditional ICA. This shows that in contrast to deep learning based

Methods	Running-time (secs)
CGANs	11015.1 ($\approx$ 3,05 hr)
Conditional ICA	62 s

Table 12.4: **Running time.** We display the running time of conditional ICA and conditional GAN (CGANs) methods used to generate synthetic task fMRI data. Conditional ICA is several orders of magnitude faster than CGANs. In practice, the computational over-head induced by Conditional ICA is negligible.

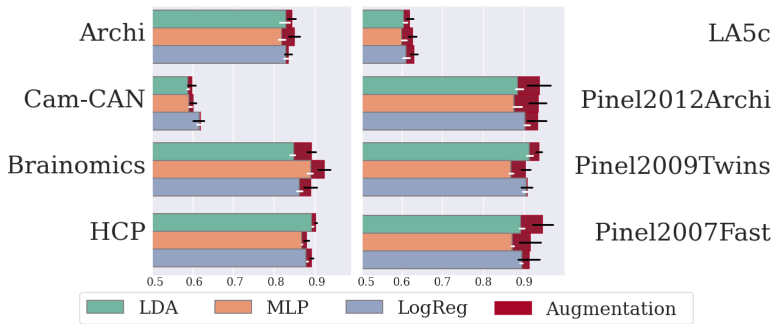


Figure 12.1: **Accuracy of models for eight multi-contrast datasets.** Cross validated accuracy of two linear (LDA and LogReg) and one non-linear classifier (MLP) with or without using data augmentation. The improvement yielded by data augmentation is displayed in red. Black error bars indicate standard deviation across splits while white error bars indicate standard deviation across splits with no augmentation.

methods, the computational over-head induced by CondICA is very low.

12.3 GAINS IN ACCURACY BROUGHT BY CONDITIONAL ICA ON EIGHT DATASETS.

In this experiment, we assess the gains brought by Conditional ICA data augmentation on the eight different task fMRI datasets refered to in section 12.1. The experimental pipeline is exactly the same as with the HCP task dataset. We report in Fig. 12.1 the cross-validated accuracy of classifiers with and without augmentation. We notice that the effect of data augmentation is consistent across datasets, classifiers and splits, with 1% to 5% net gains.

Lastly, we provide a sensitivity analysis on the number of components used in CondICA in figure 12.2. CondICA gives good performance for numbers of components between 800 and 1000 components. In all experiments we used $k = 900$ components.

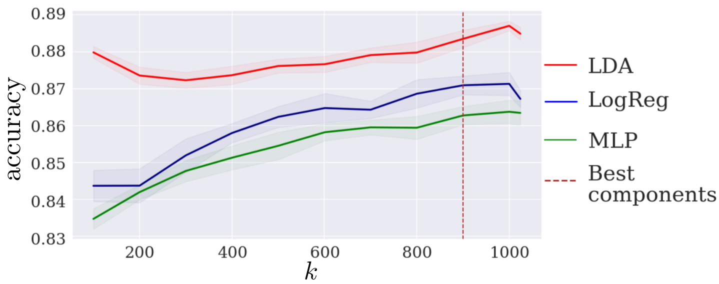


Figure 12.2: **Accuracy of augmented discriminative models when varying k.** We use 100 train subjects from the HCP task dataset to train Conditional ICA with k components and generate 200 fake subjects. Classifiers are trained on the train and fake subjects and tested on the left-out 687 subjects. We repeat the procedure for various values of k using 5 random splits per value and report the mean accuracy across splits as a function of k. The dotted line represents the number of components that has been used in our experiments ($k = 900$).

12.4 CONCLUSION

When Conditional ICA is used as a data augmentation method, it yields consistent improvement in classification accuracy: on 8 tasks fMRI datasets, we observe an increase in accuracy between 1% and 5% depending on the dataset and the classifier used. Importantly, this performance was obtained without any fine-tuning of the method, showing its reliability. One can also notice that our experiments cover datasets with different cardinalities, from tens to thousand, and different baseline prediction accuracy.

The systematic performance improvement CondICA yields makes it a promising candidate for data augmentation in a wide range of contexts. Future work may focus on its applicability to other decoding tasks such as the diagnosis of Autism Spectrum Disorder (ASD) [47, 51, 52] or Attention-Deficit/Hyperactivity Disorder detection (ADHD) [99]. Other extensions of the present work concern the adaptation to individual characteristics (e.g. age) prediction where fMRI has shown some potential.

Part VI

CONCLUSION

13 CONCLUSION

13.1 A note about resources used

All the code is written in Python. We use Matplotlib for plotting [70] , scikit-learn for machine-learning pipelines [114], MNE for MEG processing [60], Nilearn for fMRI processing and for its CanICA implementation [4], Brainiak [83] for its SRM implementation.

13.2 Contributions outside of the scope of the thesis

Some of the contributions we have made during the thesis extended beyond the scope of multivariate decompositions. We now present these contributions succinctly.

13.2.1 *A deep approach to model complex stimuli*

In this work, we learn a model to predict fMRI data of subjects watching a movie from the activities of a deep neural network exposed to the same movie. The neural network is previously trained to perform action recognition on a large corpus of movies. The association of activity in visual areas with the different layers of the deep architecture displays complexity-related contrasts across visual areas and reveals a striking foveal/peripheral dichotomy.

PUBLISHED WORK Hugo Richard et al. "Optimizing deep video representation to match brain activity." In: *Computational Cognitive Neuroscience* (2018)

13.2.2 *Predicting resting state from fMRI*

In this work, we predict task contrasts from rest fMRI data using a piecewise linear model. This model is shown to outperform linear models and a fully connected neural network.

PUBLISHED WORK Elvis Dohmatob et al. "Brain topography beyond parcellations: local gradients of functional maps." In: *NeuroImage* 229 (2021), p. 117706

13.2.3 *An optimal transport approach to hyperalignment*

In this work, we benchmark optimal transport, ridge regression and scaled Procrustes to align the data of two subjects. Optimal Transport and Ridge regression outperformed alternatives in that task.

PUBLISHED WORK Thomas Bazeille et al. "Local optimal transport for functional brain template estimation." In: *International Conference on Information Processing in Medical Imaging*. Springer. 2019, pp. 237–248

13.2.4 *Software*

The implementation of the methods developed in this thesis is freely available on Github `https://github.com/hugorichard`. Some of the code we wrote has made its way to bigger packages.

13.2.4.1 *Mvlearn*

Mvlearn [116] is a Python package for multiview learning tools. It offers reference implementations for algorithms and methods related to multiview learning. Its API is close to the scikit-learn [4] one, making it easy to learn. We have implemented the GroupICA, GroupPCA and MultiViewICA modules of mvlearn.

PUBLISHED WORK Ronan Perry et al. "mvlearn: Multiview Machine Learning in Python." In: *Journal of Machine Learning Research* 22.109 (2021), pp. 1–7. URL: `http://jmlr.org/papers/v22/20-1370.html`

13.2.4.2 *Brainiak*

Brainiak [82, 83], is a Python package that applies machine learning methods to neuroimaging data. Its API is the same as in scikit-learn and it includes modules such as Representational Similarity Analysis or Shared response modeling. We have implemented the FastSRM module of Brainiak.

PUBLISHED WORK Manoj Kumar et al. "BrainIAK: The brain imaging analysis kit." In: *Aperture* (2020). URL: `https://osf.io/db2ev/`

13.3 CONCLUSION

In this thesis, we have presented three methods to perform component analysis of multi-subject neuroimaging data and a data augmentation method for fMRI data.

First, in chapter 5 and chapter 6, we have developed an atlas based procedure that is shown to accelerate significantly the existing procedure for performing dimension reduction of fMRI data in a multi-subject context with provably no loss of performance. It is now possible to apply these algorithms in big datasets where the number of subjects is of the order of several hundreds, with several thousand samples and several hundred thousand features.

Then, we have proposed in chapter 7 and chapter 8 a novel unsupervised algorithm, MultiViewICA, that reveals latent sources observed through different views. Using an independence assumption, we have demonstrated that the model is identifiable, provided that the latent sources are not Gaussian. In contrast to previous approaches, the proposed model leads to a closed-form likelihood, which we then optimize efficiently using a dedicated alternate quasi-Newton approach. Therefore, MultiViewICA enjoys the statistical guarantees of maximum-likelihood theory, while still being tractable. MultiViewICA outperforms other unsupervised methods used to process fMRI and MEG data in the context of shared response modeling. However, it assumes the same level of noise in all subjects, which does not model properly between-subjects variability.

In chapter 9 and chapter 10, we have extended MultiViewICA in orter to deal with source noise heteroscedasticity. In practice ShICA outperforms MultiViewICA and other unsupervised methods used in the context of shared response modeling.

Lastly, we have introduced in chapter 11 and chapter 12 a data augmentation method based on ICA that outperforms deep learning algorithms in terms of decoding accuracy while being much faster.

13.4 FUTURE WORK AND PERSPECTIVES

Combined with the FastSRM algorithm, MultiViewICA and ShICA yield a novel way to make use of multi-view data. They yield a set of operators per view that map the data of each view to a shared response, reducing the variability between views. In principle, reducing the variability between views should facilitate the understanding of the data and therefore increase the performance of classification based tasks such as contrast maps labeling, automatic diagnosis or age prediction. Investigating to which extend these benefits are observed could be the topic of future research.

A second practical direction would be to apply our methods to different neuroimaging settings in which assumptions differ from ours. This thesis is geared towards naturalistic imaging, where the temporal response is assumed to be shared across subjects. We see at least two other neuroimaging settings in which our methods can be useful. The first one is the analysis of resting state data assuming that spatial topographies are shared across subjects. In this setting, the spatial topographies become the common components while the mixing operators correspond to a set of time-courses. In practice, transposing the data is enough to enforce such assumptions. A second one is the analysis of MEG / EEG data assuming both a spatial and temporal mixing. This can be done in practice by stacking the features of consecutive samples.

In terms of methods, we have treated separately dimension reduction (chapter 5) and source identification (chapter 7 and chapter 9). Future work might focus on understanding how these two steps can be performed jointly. Another possible extension in the case of naturalistic stimuli, is to assume that mixing matrices are close to each other. Indeed, as such matrices represent spatial topographies, such prior makes sense. Lastly, our data augmentation method does not make use of our understanding of multiview datasets. This constitutes an exciting direction of research.

BIBLIOGRAPHY

[1] Pierre Ablin. "Exploration of multivariate EEG/MEG signals using non-stationary models." PhD thesis. Université Paris-Saclay (ComUE), 2019.

[2] Pierre Ablin, Jean-François Cardoso, and Alexandre Gramfort. "Beyond Pham's algorithm for joint diagonalization." In: *arXiv preprint arXiv:1811.11433* (2018).

[3] Pierre Ablin, Jean-François Cardoso, and Alexandre Gramfort. "Faster independent component analysis by preconditioning with Hessian approximations." In: *IEEE Transactions on Signal Processing* 66.15 (2018), pp. 4040–4049.

[4] Alexandre Abraham, Fabian Pedregosa, Michael Eickenberg, Philippe Gervais, Andreas Mueller, Jean Kossaifi, Alexandre Gramfort, Bertrand Thirion, and Gaël Varoquaux. "Machine learning for neuroimaging with scikit-learn." In: *Frontiers in neuroinformatics* 8 (2014), p. 14.

[5] Tulay Adali, Matthew Anderson, and Geng-Shen Fu. "Diversity in independent component and vector analyses: Identifiability, algorithms, and applications in medical imaging." In: *IEEE Signal Processing Magazine* 31.3 (2014), pp. 18–33.

[6] Shun-ichi Amari, Andrzej Cichocki, and Howard H Yang. "A new learning algorithm for blind signal separation." In: *Advances in neural information processing systems*. 1996, pp. 757–763.

[7] Matthew Anderson, Tuelay Adali, and Xi-Lin Li. "Joint blind source separation with multivariate Gaussian model: Algorithms and performance analysis." In: *IEEE Transactions on Signal Processing* 60.4 (2011), pp. 1672–1683.

[8] Matthew Anderson, Geng-Shen Fu, Ronald Phlypo, and Tülay Adali. "Independent vector analysis, the Kotz distribution, and performance bounds." In: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE. 2013, pp. 3243–3247.

[9] Matthew Anderson, Geng-Shen Fu, Ronald Phlypo, and Tülay Adalı. "Independent vector analysis: Identification conditions and performance bounds." In: *IEEE Transactions on Signal Processing* 62.17 (2014), pp. 4399–4410.

[10] Michael J Anderson, Mihai Capota, Javier S Turek, Xia Zhu, Theodore L Willke, Yida Wang, Po-Hsuan Chen, Jeremy R Manning, Peter J Ramadge, and Kenneth A Norman. "Enabling factor analysis on thousand-subject neuroimaging datasets." In: *2016 IEEE International Conference on Big Data (Big Data)*. IEEE. 2016, pp. 1151–1160.

[11] Cédric Archambeau and Francis R Bach. "Sparse probabilistic projections." In: *NIPS*. 2008, pp. 73–80.

[12] Martin Arjovsky, Soumith Chintala, and Léon Bottou. "Wasserstein GAN." In: *arXiv:1701.07875 [cs, stat]* (2017).

[13] Francis R Bach and Michael I Jordan. *A probabilistic interpretation of canonical correlation analysis*. Tech. rep. University of California, 2005.

[14] Andreas Bartels and Semir Zeki. "Brain dynamics during natural viewing conditions—a new guide for mapping connectivity in vivo." In: *Neuroimage* 24.2 (2005), pp. 339–349.

[15] Thomas Bazeille, Hugo Richard, Hicham Janati, and Bertrand Thirion. "Local optimal transport for functional brain template estimation." In: *International Conference on Information Processing in Medical Imaging*. Springer. 2019, pp. 237–248.

[16] Christian F Beckmann, Marilena DeLuca, Joseph T Devlin, and Stephen M Smith. "Investigations into resting-state connectivity using independent component analysis." In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 360.1457 (2005), pp. 1001–1013.

[17] Christian F Beckmann, Clare E Mackay, Nicola Filippini, and Stephen M Smith. "Group comparison of resting-state fMRI data using multi-subject ICA and dual regression." In: *Neuroimage* 47.Suppl 1 (2009), S148.

[18] Christian F Beckmann and Stephen M Smith. "Tensorial extensions of independent component analysis for multisubject fMRI analysis." In: *Neuroimage* 25.1 (2005), pp. 294–311.

[19] Anthony J Bell and Terrence J Sejnowski. "An information-maximization approach to blind separation and blind deconvolution." In: *Neural computation* 7.6 (1995), pp. 1129–1159.

[20] Pierre Bellec, Pedro Rosa-Neto, Oliver C Lyttelton, Habib Benali, and Alan C Evans. "Multi-level bootstrap analysis of stable clusters in resting-state fMRI." In: *Neuroimage* 51.3 (2010), pp. 1126–1139.

[21] Suchita Bhinge, Rami Mowakeaa, Vince D Calhoun, and Tülay Adalı. "Extraction of time-varying spatiotemporal networks using parameter-tuned constrained IVA." In: *IEEE transactions on medical imaging* 38.7 (2019), pp. 1715–1725.

[22] Nima Bigdely-Shamlo, Tim Mullen, Kenneth Kreutz-Delgado, and Scott Makeig. "Measure projection analysis: a probabilistic approach to EEG source comparison and multi-subject inference." In: *Neuroimage* 72 (2013), pp. 287–303.

[23] Janine Diane Bijsterbosch, Mark W Woolrich, Matthew F Glasser, Emma C Robinson, Christian F Beckmann, David C Van Essen, Samuel J Harrison, and Stephen M Smith. "The relationship between spatial configuration and functional connectivity of brain regions." In: *Elife* 7 (2018), e32992.

[24] Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[25] Leo Breiman. "Random forests." In: *Machine learning* 45.1 (2001), pp. 5–32.

[26] Michael W Browne. "Factor analysis of multiple batteries by maximum likelihood." In: *British Journal of Mathematical and Statistical Psychology* 33.2 (1980), pp. 184–199.

[27] Vince D Calhoun, Tülay Adali, Vince B McGinty, James J Pekar, Todd D Watson, and Godfrey D Pearlson. "fMRI activation in a visual-perception task: network of areas detected using the general linear model and independent components analysis." In: *NeuroImage* 14.5 (2001), pp. 1080–1088.

[28] Vince D Calhoun, Tülay Adali, Godfrey D Pearlson, and James J Pekar. "A method for making group inferences from functional MRI data using independent component analysis." In: *Human brain mapping* 14.3 (2001), pp. 140–151.

[29] Vince D Calhoun, Jingyu Liu, and Tülay Adalı. "A review of group ICA for fMRI data and ICA for joint inference of imaging, genetic, and ERP data." In: *Neuroimage* 45.1 (2009), S163–S172.

[30] Vince D Calhoun, James J Pekar, Vince B McGinty, Tülay Adali, Todd D Watson, and Godfrey D Pearlson. "Different activation dynamics in multiple neural systems during simulated driving." In: *Human brain mapping* 16.3 (2002), pp. 158–167.

[31] Jean-Fran cois Cardoso. "Infomax and maximum likelihood for blind source separation." In: *IEEE Signal processing letters* (1997).

[32] Jean-François Cardoso. "Infomax and maximum likelihood for blind source separation." In: *IEEE Signal processing letters* 4.4 (1997), pp. 112–114.

[33] Jean-François Cardoso. "Blind signal separation: statistical principles." In: *Proceedings of the IEEE* 86.10 (1998), pp. 2009–2025.

[34] Jean-François Cardoso and Beate H Laheld. "Equivariant adaptive source separation." In: *IEEE Transactions on signal processing* 44.12 (1996), pp. 3017–3030.

[35] J. Chen, Y.C. Leong, K.A. Norman, and U. Hasson. "Shared experience, shared memory: a common structure for brain activity during naturalistic recall." In: *bioRxiv* (2016). DOI: 10.1101/035931. eprint: https://www.biorxiv.org/content/early/2016/01/05/035931.full.pdf. URL: https://www.biorxiv.org/content/early/2016/01/05/035931.

[36] Po-Hsuan Chen, Janice Chen, Yaara Yeshurun, Uri Hasson, James Haxby, and Peter J Ramadge. "A reduced-dimension fMRI shared response model." In: *Advances in Neural Information Processing Systems*. 2015, pp. 460–468.

[37] Po-Hsuan Chen, Xia Zhu, Hejia Zhang, Javier S Turek, Janice Chen, Theodore L Willke, Uri Hasson, and Peter J Ramadge. "A convolutional autoencoder for multi-subject fMRI data aggregation." In: *arXiv preprint arXiv:1608.04846* (2016).

[38] Pierre Comon. "Independent component analysis, a new concept?" In: *Signal processing* 36.3 (1994), pp. 287–314.

[39] Fengyu Cong, Zhaoshui He, Jarmo Hämäläinen, Paavo HT Leppänen, Heikki Lyytinen, Andrzej Cichocki, and Tapani Ristaniemi. "Validating rationale of group-level component analysis based on estimating number of sources in EEG through model order selection." In: *Journal of neuroscience methods* 212.1 (2013), pp. 165–172.

[40] Marco Congedo, Ronald Phlypo, and Jonas Chatel-Goldman. "Orthogonal and non-orthogonal joint blind source separation in the least-squares sense." In: *2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*. IEEE. 2012, pp. 1885–1889.

[41] Kamalaker Dadi, Gaël Varoquaux, Antonia Machlouzarides-Shalit, Krzysztof J Gorgolewski, Demian Wassermann, Bertrand Thirion, and Arthur Mensch. "Fine-grain atlases of functional modes for fMRI analysis." In: *NeuroImage* 221 (2020), p. 117126.

[42] Filip Deleus and Marc M. Van Hulle. "Functional connectivity analysis of fMRI data based on regularized multiset canonical correlation analysis." In: *Journal of Neuroscience Methods* 197.1 (2011), pp. 143–157. ISSN: 0165-0270. DOI: https://doi.org/10.1016/j.jneumeth.2010.11.029. URL: https://www.sciencedirect.com/science/article/pii/S0165027011000458.

[43] Arnaud Delorme, Jason Palmer, Julie Onton, Robert Oostenveld, and Scott Makeig. "Independent EEG sources are dipolar." In: *PloS one* 7.2 (2012).

[44] Arthur P Dempster, Nan M Laird, and Donald B Rubin. "Maximum likelihood from incomplete data via the EM algorithm." In: *Journal of the Royal Statistical Society: Series B (Methodological)* 39.1 (1977), pp. 1–22.

[45] Peter J Denning. "Thrashing: Its causes and prevention." In: *Proceedings of the December 9-11, 1968, fall joint computer conference, part I*. 1968, pp. 915–922.

[46] Elvis Dohmatob, Hugo Richard, Ana Luísa Pinho, and Bertrand Thirion. "Brain topography beyond parcellations: local gradients of functional maps." In: *NeuroImage* 229 (2021), p. 117706.

[47] Nicha C Dvornek, Pamela Ventola, Kevin A Pelphrey, and James S Duncan. "Identifying autism from resting-state fMRI using long short-term memory networks." In: *International Workshop on Machine Learning in Medical Imaging*. Springer. 2017, pp. 362–370.

[48] Tom Eichele, Srinivas Rachakonda, Brage Brakedal, Rune Eikeland, and Vince D Calhoun. "EEGIFT: group independent component analysis for event-related EEG data." In: *Computational intelligence and neuroscience* 2011 (2011).

[49] Denis Alexander Engemann, Oleh Kozynets, David Sabbagh, Guillaume Lemaitre, Gaël Varoquaux, Franziskus Liem, and Alexandre Gramfort. "Combining electrophysiology with MRI enhances learning of surrogate-biomarkers." In: *bioRxiv* (2019), p. 856336.

[50] Erik Barry Erhardt, Srinivas Rachakonda, Edward J Bedrick, Elena A Allen, Tülay Adali, and Vince D Calhoun. "Comparison of multi-subject ICA methods for analysis of fMRI data." In: *Human brain mapping* 32.12 (2011), pp. 2075–2095.

[51] Taban Eslami, Vahid Mirjalili, Alvis Fong, Angela R Laird, and Fahad Saeed. "ASD-DiagNet: a hybrid learning approach for detection of autism spectrum disorder using fMRI data." In: *Frontiers in neuroinformatics* 13 (2019), p. 70.

[52] Taban Eslami and Fahad Saeed. "Auto-ASD-network: a technique based on deep learning and support vector machines for diagnosing autism spectrum disorder using fMRI data." In: *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*. 2019, pp. 646–651.

[53] Fabrizio Esposito, Tommaso Scarabino, Aapo Hyvärinen, Johan Himberg, Elia Formisano, Silvia Comani, Gioacchino Tedeschi, Rainer Goebel, Erich Seifritz, and Francesco Di Salle. "Independent component analysis of fMRI group studies by self-organizing clustering." In: *Neuroimage* 25.1 (2005), pp. 193–205.

[54] Thomas S Ferguson. *A course in large sample theory*. Routledge, 2017.

[55] Karl J Friston, Andrew P Holmes, JB Poline, PJ Grasby, SCR Williams, Richard SJ Frackowiak, and Robert Turner. "Analysis of fMRI time-series revisited." In: *Neuroimage* 2.1 (1995), pp. 45–53.

[56] Philip E Gill and Walter Murray. "Quasi-Newton methods for unconstrained optimization." In: *IMA Journal of Applied Mathematics* 9.1 (1972), pp. 91–108.

[57] Gene H Golub. "Some modified matrix eigenvalue problems." In: *Siam Review* 15.2 (1973), pp. 318–334.

[58] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative adversarial nets." In: *Advances in neural information processing systems*. 2014, pp. 2672–2680.

[59] Krzysztof J Gorgolewski, Gael Varoquaux, Gabriel Rivera, Yannick Schwarz, Satrajit S Ghosh, Camille Maumet, Vanessa V Sochat, Thomas E Nichols, Russell A Poldrack, Jean-Baptiste Poline, et al. "NeuroVault. org: a web-based repository for collecting and sharing unthresholded statistical maps of the human brain." In: *Frontiers in neuroinformatics* 9 (2015), p. 8.

[60] Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A Engemann, Daniel Strohmeier, Christian Brodbeck, Roman Goj, Mainak Jas, Teon Brooks, Lauri Parkkonen, et al. "MEG and EEG data analysis with MNE-Python." In: *Frontiers in neuroscience* 7 (2013), p. 267.

[61] Luigi Gresele, Paul K. Rubenstein, Arash Mehrjou, Francesco Locatello, and Bernhard Schölkopf. "The Incomplete Rosetta Stone problem: Identifiability results for Multi-view Nonlinear ICA." In: *Proceedings of the Thirty-Fifth Conference on Uncertainty in Artificial Intelligence, UAI 2019, Tel Aviv, Israel, July 22-25, 2019*. Ed. by Amir Globerson and Ricardo Silva. AUAI Press, 2019, p. 53. URL: http://auai.org/uai2019/proceedings/papers/53.pdf.

[62] Vera A Grin-Yatsenko, Ineke Baas, Valery A Ponomarev, and Juri D Kropotov. "Independent component approach to the analysis of EEG recordings at early stages of depressive disorders." In: *Clinical Neurophysiology* 121.3 (2010), pp. 281–289.

[63] J Swaroop Guntupalli, Ma Feilong, and James V Haxby. "A computational model of shared fine-scale structure in the human connectome." In: *PLoS computational biology* 14.4 (2018), e1006120.

[64] Ying Guo and Giuseppe Pagnoni. "A unified framework for group independent component analysis for multi-subject fMRI data." In: *NeuroImage* 42.3 (2008), pp. 1078–1093.

[65] Michael Hanke, Florian J Baumgartner, Pierre Ibe, Falko R Kaule, Stefan Pollmann, Oliver Speck, Wolf Zinke, and Jörg Stadler. "A high-resolution 7-Tesla fMRI dataset from complex natural stimulation with an audio movie." In: *Scientific data* 1 (2014), p. 140003.

[66] Riitta Hari and Aina Puce. *MEG-EEG Primer*. Oxford University Press, 2017.

[67] Stefan Harmeling, Suvrit Sra, Michael Hirsch, and Bernhard Schölkopf. "Multiframe blind deconvolution, super-resolution, and saturation correction via incremental EM." In: *2010 IEEE International Conference on Image Processing*. IEEE. 2010, pp. 3313–3316.

[68] Uri Hasson, Yuval Nir, Ifat Levy, Galit Fuhrmann, and Rafael Malach. "Intersubject synchronization of cortical activity during natural vision." In: *science* 303.5664 (2004), pp. 1634–1640.

[69] James V Haxby, J Swaroop Guntupalli, Andrew C Connolly, Yaroslav O Halchenko, Bryan R Conroy, M Ida Gobbini, Michael Hanke, and Peter J Ramadge. "A common, high-dimensional model of the representational space in human ventral temporal cortex." In: *Neuron* 72.2 (2011), pp. 404–416.

[70] John D Hunter. "Matplotlib: A 2D graphics environment." In: *Computing in science & engineering* 9.3 (2007), pp. 90–95.

[71] Alexander G Huth, Shinji Nishimoto, An T Vu, and Jack L Gallant. "A continuous semantic space describes the representation of thousands of object and action categories across the human brain." In: *Neuron* 76.6 (2012), pp. 1210–1224.

[72] A Hyvärinen. "Analysis and projection pursuit." In: *Advances in neural information processing systems* 10 (1998), p. 273.

[73] Aapo Hyvärinen. "Independent component analysis in the presence of Gaussian noise by maximizing joint likelihood." In: *Neurocomputing* 22.1-3 (1998), pp. 49–67.

[74] Aapo Hyvarinen. "Fast and robust fixed-point algorithms for independent component analysis." In: *IEEE transactions on Neural Networks* 10.3 (1999), pp. 626–634.

[75] Aapo Hyvarinen. "Gaussian moments for noisy independent component analysis." In: *IEEE signal processing letters* 6.6 (1999), pp. 145–147.

[76] Aapo Hyvärinen and Erkki Oja. "Independent component analysis: algorithms and applications." In: *Neural networks* 13.4-5 (2000), pp. 411–430.

[77] Mainak Jas, Eric Larson, Denis A Engemann, Jaakko Leppäkangas, Samu Taulu, Matti Hämäläinen, and Alexandre Gramfort. "A reproducible MEG/EEG group study with the MNE software: recommendations, quality assessments, and good practices." In: *Frontiers in neuroscience* 12 (2018), p. 530.

[78] Tzyy-Ping Jung, Colin Humphries, Te-Won Lee, Scott Makeig, Martin J McKeown, Vicente Iragui, and Terrence J Sejnowski. "Extended ICA removes artifacts from electroencephalographic recordings." In: *Advances in neural information processing systems*. 1998, pp. 894–900.

[79] Jon R Kettenring. "Canonical analysis of several sets of variables." In: *Biometrika* 58.3 (1971), pp. 433–451.

[80] Diederik P Kingma and Max Welling. "Auto-encoding variational bayes." In: *arXiv preprint arXiv:1312.6114* (2013).

[81] Arto Klami, Seppo Virtanen, Eemeli Leppäaho, and Samuel Kaski. "Group factor analysis." In: *IEEE transactions on neural networks and learning systems* 26.9 (2014), pp. 2136–2147.

[82] Manoj Kumar, Michael J Anderson, James W Antony, Christopher Baldassano, Paula P Brooks, Ming Bo Cai, Po-Hsuan Cameron Chen, Cameron T Ellis, Gregory Henselman-Petrusek, David Huberdeau, et al. "BrainIAK: The brain imaging analysis kit." In: *Aperture* (2020). URL: https://osf.io/db2ev/.

[83] Manoj Kumar, Cameron T Ellis, Qihong Lu, Hejia Zhang, Mihai Capotă, Theodore L Willke, Peter J Ramadge, Nicholas B Turk-Browne, and Kenneth A Norman. "BrainIAK tutorials: User-friendly learning materials for advanced fMRI analysis." In: *PLoS computational biology* 16.1 (2020), e1007549.

[84] Olivier Ledoit and Michael Wolf. "A well-conditioned estimator for large-dimensional covariance matrices." In: *Journal of multivariate analysis* 88.2 (2004), pp. 365–411.

[85] Jong-Hwan Lee, Te-Won Lee, Ferenc A Jolesz, and Seung-Schik Yoo. "Independent vector analysis (IVA): multivariate approach for fMRI group study." In: *Neuroimage* 40.1 (2008), pp. 86–109.

[86] Te-Won Lee, Mark Girolami, and Terrence J Sejnowski. "Independent component analysis using an extended infomax algorithm for mixed subgaussian and supergaussian sources." In: *Neural computation* 11.2 (1999), pp. 417–441.

[87] Roger Levy. "Probabilistic models in the study of language." In: *Online Draft, Nov* (2012).

[88] Xi-Lin Li and Tulay Adali. "A novel entropy estimator and its application to ICA." In: *2009 IEEE International Workshop on Machine Learning for Signal Processing*. IEEE. 2009, pp. 1–6.

[89] Xi-Lin Li and Tülay Adali. “Blind spatiotemporal separation of second and/or higher-order correlated sources by entropy rate minimization.” In: *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE. 2010, pp. 1934–1937.

[90] Xi-Lin Li and Tülay Adali. “Independent component analysis by entropy bound minimization.” In: *IEEE Transactions on Signal Processing* 58.10 (2010), pp. 5151–5164.

[91] Xi-Lin Li, Tülay Adalı, and Matthew Anderson. “Joint blind source separation by generalized joint diagonalization of cumulant matrices.” In: *Signal Processing* 91.10 (2011), pp. 2314–2322.

[92] Yi-Ou Li, Tülay Adali, Wei Wang, and Vince D Calhoun. “Joint blind source separation by multiset canonical correlation analysis.” In: *IEEE Transactions on Signal Processing* 57.10 (2009), pp. 3918–3929.

[93] Thomas T Liu. “Noise contributions to the fMRI signal: An overview.” In: *NeuroImage* 143 (2016), pp. 141–151.

[94] Qunfang Long, Suchita Bhinge, Vince D Calhoun, and Tülay Adali. “Independent vector analysis for common subspace analysis: Application to multi-subject fMRI data yields meaningful subgroups of schizophrenia.” In: *NeuroImage* 216 (2020), p. 116872.

[95] João Loula, Gaël Varoquaux, and Bertrand Thirion. “Decoding fMRI activity in the time domain improves classification performance.” In: *NeuroImage* 180 (2018), pp. 203–210.

[96] Gilles Louppe, Louis Wehenkel, Antonio Sutera, and Pierre Geurts. “Understanding variable importances in forests of randomized trees.” In: *Advances in neural information processing systems*. 2013, pp. 431–439.

[97] Scott Makeig, Anthony J Bell, Tzyy-Ping Jung, and Terrence J Sejnowski. “Independent component analysis of electroencephalographic data.” In: *Advances in neural information processing systems*. 1996, pp. 145–151.

[98] Sanna Malinen, Yevhen Hlushchuk, and Riitta Hari. “Towards natural stimulation in fMRI—issues of data analysis.” In: *Neuroimage* 35.1 (2007), pp. 131–139.

[99] Zhenyu Mao, Yi Su, Guangquan Xu, Xueping Wang, Yu Huang, Weihua Yue, Li Sun, and Naixue Xiong. “Spatio-temporal deep learning method for adhd fmri classification.” In: *Information Sciences* 499 (2019), pp. 1–11.

[100] Martin J McKeown and Terrence J Sejnowski. “Independent component analysis of fMRI data: examining the assumptions.” In: *Human brain mapping* 6.5-6 (1998), pp. 368–372.

[101] Arthur Mensch, Julien Mairal, Bertrand Thirion, and Gaël Varoquaux. “Extracting Universal Representations of Cognition across Brain-Imaging Studies.” In: *arXiv preprint arXiv:1809.06035* (2018).

[102] Mehdi Mirza and Simon Osindero. “Conditional generative adversarial nets.” In: *arXiv preprint arXiv:1411.1784* (2014).

[103] R. P. Monti and A. Hyvärinen. “A unified probabilistic model for learning latent factors and their connectivities from high-dimensional data.” In: *Proc. 34th Conf. on Uncertainty in Artificial Intelligence (UAI2018)*. Monterey, California, 2018.

[104] Eric Moulines, Jean-Franccois Cardoso, and Elisabeth Gassiat. “Maximum likelihood for blind separation and deconvolution of noisy signals using mixture models.” In: *1997 ieee international conference on acoustics, speech, and signal processing*. Vol. 5. IEEE. 1997, pp. 3617–3620.

[105] Samuel A Nastase, Yun-Fei Liu, Hanna Hillman, Kenneth A. Norman, and Uri Hasson. “Leveraging shared connectivity to aggregate heterogeneous datasets into a common response space.” In: *bioRxiv* (2019). DOI: 10.1101/741975. eprint: https://www.biorxiv.org/content/early/2019/08/21/741975.full.pdf. URL: https://www.biorxiv.org/content/early/2019/08/21/741975.

[106] Radford M Neal and Geoffrey E Hinton. “A view of the EM algorithm that justifies incremental, sparse, and other variants.” In: *Learning in graphical models*. Springer, 1998, pp. 355–368.

[107] Allan Aasbjerg Nielsen. “Multiset canonical correlations analysis and multispectral, truly multitemporal remote sensing data.” In: *IEEE transactions on image processing* 11.3 (2002), pp. 293–305.

[108] S. Nishimoto, A.T. Vu, T. Naselaris, Y. Benjamini, B. Yu, and J. L Gallant. “Reconstructing visual experiences from brain activity evoked by natural movies.” In: *Current Biology* 21.19 (2011), pp. 1641–1646.

[109] Jorge Nocedal and Stephen Wright. *Numerical optimization*. Springer Science & Business Media, 2006.

[110] Seiji Ogawa, Tso-Ming Lee, Alan R Kay, and David W Tank. “Brain magnetic resonance imaging with contrast dependent on blood oxygenation.” In: *proceedings of the National Academy of Sciences* 87.24 (1990), pp. 9868–9872.

[111] Dimitri Papadopoulos Orfanos, Vincent Michel, Yannick Schwartz, Philippe Pinel, Antonio Moreno, Denis Le Bihan, and Vincent Frouin. “The brainomics/localizer database.” In: *Neuroimage* 144 (2017), pp. 309–314.

[112] Roberto D Pascual-Marqui. "Discrete, 3D distributed, linear imaging methods of electric neuronal activity. Part 1: exact, zero error localization." In: *arXiv preprint arXiv:0710.3341* (2007).

[113] Roberto Domingo Pascual-Marqui et al. "Standardized low-resolution brain electromagnetic tomography (sLORETA): technical details." In: *Methods Find Exp Clin Pharmacol* 24.Suppl D (2002), pp. 5–12.

[114] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. "Scikit-learn: Machine learning in Python." In: *the Journal of machine Learning research* 12 (2011), pp. 2825–2830.

[115] W Penny and A Holmes. "Random effects analysis." In: *Statistical parametric mapping: The analysis of functional brain images* 156 (2007), p. 165.

[116] Ronan Perry et al. "mvlearn: Multiview Machine Learning in Python." In: *Journal of Machine Learning Research* 22.109 (2021), pp. 1–7. URL: http://jmlr.org/papers/v22/20-1370.html.

[117] Niklas Pfister, Sebastian Weichwald, Peter Bühlmann, and Bernhard Schölkopf. "Robustifying independent component analysis by adjusting for group-wise stationary noise." In: *Journal of Machine Learning Research* 20.147 (2019), pp. 1–50.

[118] Dinh-Tuan Pham. "Blind separation of instantaneous mixture of sources via the Gaussian mutual information criterion." In: *Signal Processing* 81.4 (2001), pp. 855–870.

[119] Dinh Tuan Pham. "Joint approximate diagonalization of positive definite Hermitian matrices." In: *SIAM Journal on Matrix Analysis and Applications* 22.4 (2001), pp. 1136–1152.

[120] Dinh-Tuan Pham and J-F Cardoso. "Blind separation of instantaneous mixtures of nonstationary sources." In: *IEEE Transactions on signal processing* 49.9 (2001), pp. 1837–1848.

[121] Philippe Pinel and Stanislas Dehaene. "Genetic and environmental contributions to brain activation during calculation." In: *Neuroimage* 81 (2013), pp. 306–316.

[122] Philippe Pinel, Bertrand Thirion, Sébastien Meriaux, Antoinette Jobert, Julien Serres, Denis Le Bihan, Jean-Baptiste Poline, and Stanislas Dehaene. "Fast reproducible identification and large-scale databasing of individual functional cognitive networks." In: *BMC neuroscience* 8.1 (2007), p. 91.

[123] Philippe Pinel, Baudouin Forgeot d'Arc, Stanislas Dehaene, Thomas Bourgeron, Bertrand Thirion, Denis Le Bihan, and Cyril Poupon. "The functional database of the ARCHI project: Potential and perspectives." In: *NeuroImage* 197 (2019), pp. 527–543.

[124] Ana Luísa Pinho, Alexis Amadon, Baptiste Gauthier, Nicolas Clairis, André Knops, Sarah Genon, Elvis Dohmatob, Juan Jesús Torre, Chantal Ginisty, Séverine Becuwe-Desmidt, et al. "Individual Brain Charting dataset extension, second release of high-resolution fMRI data for cognitive mapping." In: *Scientific Data* 7.1 (2020), pp. 1–16.

[125] Ana Luísa Pinho, Alexis Amadon, Torsten Ruest, Murielle Fabre, Elvis Dohmatob, Isabelle Denghien, Chantal Ginisty, Séverine Becuwe-Desmidt, Séverine Roger, Laurence Laurier, et al. "Individual Brain Charting, a high-resolution fMRI dataset for cognitive mapping." In: *Scientific data* 5 (2018).

[126] Russell A Poldrack, Deanna M Barch, Jason Mitchell, Tor Wager, Anthony D Wagner, Joseph T Devlin, Chad Cumba, Oluwasanmi Koyejo, and Michael Milham. "Toward open sharing of task-based fMRI data: the OpenfMRI project." In: *Frontiers in neuroinformatics* 7 (2013), p. 12.

[127] Russell A Poldrack, Eliza Congdon, William Triplett, KJ Gorgolewski, KH Karlsgodt, JA Mumford, FW Sabb, NB Freimer, ED London, TD Cannon, et al. "A phenome-wide examination of neural and cognitive function." In: *Scientific data* 3.1 (2016), pp. 1–12.

[128] Russell A Poldrack, Jeanette A Mumford, and Thomas E Nichols. *Handbook of functional MRI data analysis*. Cambridge University Press, 2011.

[129] Jean-Baptiste Poline and Matthew Brett. "The general linear model and fMRI: does love last forever?" In: *Neuroimage* 62.2 (2012), pp. 871–880.

[130] Mehdi Rahim, Bertrand Thirion, Danilo Bzdok, Irène Buvat, and Gaël Varoquaux. "Joint prediction of multiple scores captures better individual traits from brain images." In: *NeuroImage* 158 (2017), pp. 145–154.

[131] Danilo Jimenez Rezende and Shakir Mohamed. "Variational inference with normalizing flows." In: *arXiv preprint arXiv:1505.05770* (2015).

[132] H. Richard, P. Ablin, B. Thirion, A. Gramfort, and A. Hyvarinen. "Shared Independent Component Analysis for Multi-Subject Neuroimaging." In: *Advances in Neural Information Processing Systems 33*. Dec. 2021.

[133] H. Richard, L. Gresele, A. Hyvarinen, B. Thirion, A. Gramfort, and P. Ablin. "Modeling Shared responses in Neuroimaging Studies through MultiView ICA." In: *Advances in Neural Information Processing Systems 33*. Dec. 2020.

[134] Hugo Richard, Ana Pinho, Bertrand Thirion, and Guillaume Charpiat. "Optimizing deep video representation to match brain activity." In: *Computational Cognitive Neuroscience* (2018).

[135] Thomas J Rothenberg. "Identification in parametric models." In: *Econometrica: Journal of the Econometric Society* (1971), pp. 577–591.

[136] Alexander Schaefer, Ru Kong, Evan M Gordon, Timothy O Laumann, Xi-Nian Zuo, Avram J Holmes, Simon B Eickhoff, and BT Thomas Yeo. "Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI." In: *Cerebral Cortex* 28.9 (2017), pp. 3095–3114.

[137] Meredith A Shafto, Lorraine K Tyler, Marie Dixon, Jason R Taylor, James B Rowe, Rhodri Cusack, Andrew J Calder, William D Marslen-Wilson, John Duncan, Tim Dalgleish, et al. "The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) study protocol: a cross-sectional, lifespan, multidisciplinary examination of healthy cognitive ageing." In: *BMC neurology* 14.1 (2014), p. 204.

[138] W. R. Shirer, S. Ryali, E. Rykhlevskaia, V. Menon, and M. D. Greicius. "Decoding Subject-Driven Cognitive States with Whole-Brain Connectivity Patterns." In: *Cerebral Cortex (New York, NY)* 22.1 (2012), pp. 158–165.

[139] Stephen M Smith, Aapo Hyvärinen, Gaël Varoquaux, Karla L Miller, and Christian F Beckmann. "Group-PCA for very large fMRI datasets." In: *Neuroimage* 101 (2014), pp. 738–749.

[140] Charles Stein et al. "Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution." In: *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. The Regents of the University of California. 1956.

[141] Gilbert W Stewart. "Error and perturbation bounds for subspaces associated with certain eigenvalue problems." In: *SIAM review* 15.4 (1973), pp. 727–764.

[142] Markus Svensén, Frithjof Kruggel, and Habib Benali. "ICA of fMRI group study data." In: *NeuroImage* 16.3 (2002), pp. 551–563.

[143] François Tadel, Sylvain Baillet, John C Mosher, Dimitrios Pantazis, and Richard M Leahy. "Brainstorm: a user-friendly application for MEG/EEG analysis." In: *Computational intelligence and neuroscience* 2011 (2011).

[144] Badr Tajini, Hugo Richard, and Bertrand Thirion. "Functional Magnetic Resonance Imaging data augmentation through conditional ICA." In: *MICCAI* (2021).

[145] Samu Taulu and Juha Simola. "Spatiotemporal signal space separation method for rejecting nearby interference in MEG measurements." In: *Physics in Medicine & Biology* 51.7 (2006), p. 1759.

[146] Jason R Taylor, Nitin Williams, Rhodri Cusack, Tibor Auer, Meredith A Shafto, Marie Dixon, Lorraine K Tyler, Richard N Henson, et al. "The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) data repository: structural and functional MRI, MEG, and cognitive data from a cross-sectional adult lifespan sample." In: *Neuroimage* 144 (2017), pp. 262–269.

[147] Petr Tichavsky and Zbynek Koldovsky. "Optimal pairing of signal components separated by blind techniques." In: *IEEE Signal Processing Letters* 11.2 (2004), pp. 119–122.

[148] Gautam Tripathi. "A matrix extension of the Cauchy-Schwarz inequality." In: *Economics Letters* 63.1 (1999), pp. 1–3.

[149] David C Van Essen, Stephen M Smith, Deanna M Barch, Timothy EJ Behrens, Essa Yacoub, Kamil Ugurbil, Wu-Minn HCP Consortium, et al. "The WU-Minn human connectome project: an overview." In: *Neuroimage* 80 (2013), pp. 62–79.

[150] Gaël Varoquaux, Sepideh Sadaghiani, Jean-Baptiste Poline, and Bertrand Thirion. "CanICA: Model-based extraction of reproducible group-level ICA patterns from fMRI time series." In: *arXiv preprint arXiv:0911.4650* (2009).

[151] Gael Varoquaux and Bertrand Thirion. "How machine learning is shaping cognitive neuroimaging." In: *GigaScience* 3 (2014), p. 28.

[152] Javier Vía, Matthew Anderson, Xi-Lin Li, and Tülay Adalı. "A maximum likelihood approach for independent vector analysis of Gaussian data sets." In: *2011 IEEE International Workshop on Machine Learning for Signal Processing*. IEEE. 2011, pp. 1–6.

[153] Javier Vía, Matthew Anderson, Xi-Lin Li, and Tülay Adalı. "Joint blind source separation from second-order statistics: Necessary and sufficient identifiability conditions." In: *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2011, pp. 2520–2523.

[154] Ricardo Vigário, Veikko Jousmäki, Matti Hämäläinen, Riitta Hari, and Erkki Oja. "Independent component analysis for identification of artifacts in magnetoencephalographic recordings." In: *Advances in neural information processing systems*. 1998, pp. 229–235.

[155] Ricardo Vigário, Jaakko Sarela, Veikko Jousmiki, Matti Hamalainen, and Erkki Oja. "Independent component approach to the analysis of EEG and MEG recordings." In: *IEEE transactions on biomedical engineering* 47.5 (2000), pp. 589–593.

[156] Daniela M Witten and Robert J Tibshirani. "Extensions of sparse canonical correlation analysis with applications to genomic data." In: *Statistical applications in genetics and molecular biology* 8.1 (2009).

[157] Hejia Zhang, Po-Hsuan Chen, Janice Chen, Xia Zhu, Javier S Turek, Theodore L Willke, Uri Hasson, and Peter J Ramadge. "A searchlight factor model approach for locating shared information in multi-subject fMRI analysis." In: *arXiv preprint arXiv:1609.09432* (2016).

[158] Hejia Zhang, Po-Hsuan Chen, and Peter Ramadge. "Transfer learning on fMRI datasets." In: *International Conference on Artificial Intelligence and Statistics*. 2018, pp. 595–603.

[159] Peiye Zhuang, Alexander G Schwing, and Oluwasanmi Koyejo. "Fmri data augmentation via synthesis." In: *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. IEEE. 2019, pp. 1783–1787.

[160] Michael Zibulevsky. "Blind source separation with relative newton method." In: *Proc. ICA*. Vol. 2003. 2003, pp. 897–902.

Part VII

APPENDICES

A MULTIVIEWICA

A.1 PROOFS OF SECTION 7.1

A.1.1 *Proof of Prop. 12*

We fix a subject i. Since $\mathbf{s}$ has independent components, so does $\mathbf{s}+\mathbf{n}_i$. Following [38], Theorem 11, there exists a scale-permutation matrix P^i such that $A_i' = A_i P^i$. As a consequence, we have $\mathbf{s}+\mathbf{n}_i = P^i(\mathbf{s}'+\mathbf{n}'^i)$ for all i.

Then, we focus on subject 1 and subject $i \neq 1$:

$$\mathbf{s}+\mathbf{n}^1 - (\mathbf{s}+\mathbf{n}_i) = P^1(\mathbf{s}'+\mathbf{n}'^1) - P^i(\mathbf{s}'+\mathbf{n}'^i) \tag{A.1}$$

$$\mathbf{n}^1 - \mathbf{n}_i = P^1(\mathbf{s}'+\mathbf{n}'^1) - P^i(\mathbf{s}'+\mathbf{n}'^i) \tag{A.2}$$

$$\iff P^1\mathbf{s}' - P^i\mathbf{s}' = P^i\mathbf{n}'^i - \mathbf{n}_i + \mathbf{n}^1 - P^1\mathbf{n}'^1 \tag{A.3}$$

Since the right hand side of equation (A.3) is a linear combination of Gaussian random variables, this would imply that $P^1\mathbf{s}' - P^i\mathbf{s}'$ is also Gaussian. However, given that $\mathbf{s}'$ is assumed to be non-Gaussian, the equality can only hold if $P^1 = P^i$ and both the right and the left hand side vanish. Therefore, the matrices P^i are all equal, and there exists a scale and permutation matrix P such that $A_i' = A_i P$.

A.1.2 *Proof of Prop. 13*

We consider $W_i = \Lambda(A_i)^{-1}$, where Λ is a diagonal matrix. We recall $\mathbf{x}_i = A_i(\mathbf{s}+\mathbf{n}_i)$, so that $\mathbf{y}_i = W_i\mathbf{x}_i = \Lambda(\mathbf{s}+\mathbf{n}_i)$. The gradient of $\mathcal{L}$ is given by equation (7.9):

$$\begin{aligned}\mathcal{G}^i &= \frac{1}{m}\mathbb{E}[f'(\tilde{\mathbf{s}})(\mathbf{s}+\mathbf{n}_i)^\top]\Lambda \\ &\quad + \frac{1-1/m}{\sigma^2}\Lambda\mathbb{E}[\left(\mathbf{n}_i - \frac{1}{m-1}\sum_{j\neq i}\mathbf{n}^j\right)(\mathbf{s}+\mathbf{n}_i)^\top]\Lambda - I_p \\ &= \frac{1}{m}\mathbb{E}[f'(\Lambda(\mathbf{s}+\frac{1}{m}\sum_j \mathbf{n}^j))(\mathbf{s}+\mathbf{n}_i)^\top]\Lambda + \frac{\sigma'^2(1-1/m)}{\sigma^2}\Lambda^2 - I_p\end{aligned} \tag{A.4}$$

where we write $f'(\mathbf{s}) = \begin{bmatrix} f'(s_1) \\ \vdots \\ f'(s_p)\end{bmatrix}$. Therefore, $\mathcal{G}^i$ is diagonal and constant across subjects (because $\mathbb{E}[f'(\Lambda(\mathbf{s}+\frac{1}{m}\sum_j \mathbf{n}^j))(\mathbf{n}_i)^\top] =$

$\mathbb{E}[f'(\Lambda(\mathbf{s}+\frac{1}{m}\sum_j \mathbf{n}^j))(\mathbf{n}^{i'})^\top])$. Let us therefore consider only its coefficient (a,a), and let $\lambda = \Lambda_{aa}$:

$$\mathcal{G}^i_{aa} = G(\lambda) = \phi(\lambda)\lambda + \frac{\sigma'^2(1-1/m)}{\sigma^2}\lambda^2 - 1,$$

where $\phi(\lambda) = \frac{1}{m}\mathbb{E}[f'(\lambda(s_a + \frac{1}{m}\sum_j n^j_a))(s_a + n^i_a)]$. One the one hand, we have $G(0) = -1$. On the other hand, if we assume for instance that f' has sub linear growth (i.e. $|f'(x)| \leqslant c|x|^\alpha + d$ for some $\alpha < 1$) or that ϕ is positive, we find that $G(+\infty) = +\infty$. Therefore, G cancels, which concludes the proof.

A.1.3 *Stability conditions*

We consider $W_i = \Lambda(A_i)^{-1}$ where Λ is such that the gradients $\mathcal{G}^i$ all cancel. We consider a small relative perturbation of W_i of the form $W_i \leftarrow (I_p + E^i)W_i$, and consider the effect on the gradient. We define $\Delta^i = \mathcal{G}^i\left((I_p + E^1)W_1, \dots, (I_p + E^m)W_m\right)$. Denoting $C = \frac{1-1/m}{\sigma^2}$ and $\tilde{\mathbf{n}} = \frac{1}{m}\sum_{i=1}^m \mathbf{n}_i$, we find:

$$\Delta^i = \Delta^i_1 + C\Delta^i_2 - I_p$$

where

$$\Delta^i_1 = \mathbb{E}[\frac{1}{m}f'\left(\Lambda(\mathbf{s}+\tilde{\mathbf{n}}) + \frac{1}{m}\sum_{j=1}^m E^j\Lambda(\mathbf{s}+\mathbf{n}^j)\right)(\mathbf{s}+\mathbf{n}_i)^\top\Lambda(I_p+E^i)^\top] \tag{A.5}$$

and

$$\Delta^i_2 = \mathbb{E}[\left(\Lambda\mathbf{n}_i - \frac{1}{m-1}\sum_{j\neq i}\Lambda\mathbf{n}^j + E^i\Lambda(\mathbf{s}+\mathbf{n}_i)\right.$$
$$\left.-\frac{1}{m-1}\sum_{j\neq i}E^j\Lambda(\mathbf{s}+\mathbf{n}^j)\right)(\mathbf{s}+\mathbf{n}_i)^\top\Lambda(I_p+E^i)^\top] \tag{A.6}$$

The first term is expanded at the first order, denoting $S = \sum_{j=1}^m E^j$:

$$\begin{aligned}\Delta^i_1 = \mathbb{E}[\frac{1}{m}&\left(f''(\Lambda(\mathbf{s}+\tilde{\mathbf{n}})) \odot \left(\frac{1}{m}\sum_{j=1}^m E^j\Lambda(\mathbf{s}+\mathbf{n}^j)\right)\right.\\ &\left.+ f'(\Lambda(\mathbf{s}+\tilde{\mathbf{n}}))\right)(\mathbf{s}+\mathbf{n}_i)^\top\Lambda(I_p+E^i)^\top]\\ &= \mathbb{E}[\frac{1}{m}f'(\Lambda(\mathbf{s}+\tilde{\mathbf{n}}))(\mathbf{s}+\mathbf{n}_i)^\top\Lambda(I_p+E^i)^\top\\ &+ \frac{1}{m^2}S\odot\left(f''(\Lambda(\mathbf{s}+\tilde{\mathbf{n}}))(\mathbf{s}^2)^\top\Lambda^2\right)\\ &+ \frac{1}{m^2}E^i\odot\left(f''(\Lambda(\mathbf{s}+\tilde{\mathbf{n}}))((\mathbf{n}_i)^2)^\top\Lambda^2\right)]\end{aligned} \tag{A.7}$$

The symbol $\odot$ denotes the element-wise multiplication, $f'(\mathbf{s}) = \begin{bmatrix} f'(s_1) \\ \vdots \\ f'(s_p) \end{bmatrix}$ and $f''(\mathbf{s}) = \begin{bmatrix} f''(s_1) \\ \vdots \\ f''(s_p) \end{bmatrix}$. Similarly, the second term gives at the first order:

$$\Delta_2^i = \mathbb{E}[\sigma'^2\Lambda^2(I_p + E^i)^\top + (1+\sigma'^2)E^i\Lambda^2 - \frac{1}{m-1}(S - E^i)\Lambda^2] \tag{A.8}$$

Combining this, we find:

$$\Delta^i = (E^i)^\top + E^i \odot \Gamma^E + S \odot \Gamma^S \tag{A.9}$$

where

$$\Gamma^E = \left(\frac{1}{m^2}\mathbb{E}[f''(\Lambda(\mathbf{s}+\tilde{\mathbf{n}}))((\mathbf{n}_i)^2)^\top] + (1 - \frac{1}{m})\frac{\sigma'^2}{\sigma^2} + \frac{1}{\sigma^2}\right)\Lambda^2$$

$$\Gamma^S = \left(\frac{1}{m^2}\mathbb{E}[f''(\Lambda(\mathbf{s}+\tilde{\mathbf{n}}))(\mathbf{s}^2)^\top] - \frac{1}{m\sigma^2}\right)\Lambda^2$$

are $p \times p$ matrices, independent of the subject. This linear operator is the Hessian block corresponding to the i-th subject: Denoting $\mathcal{H}$ the Hessian, it is the mapping $\mathcal{H}(E^1, \dots, E^m) = (\Delta^1, \dots, \Delta^m)$.

The coefficient Δ^i_{ab} only depends on $(E^i_{ab}, E^i_{ba}, E^1_{ab}, \dots, E^m_{ab})$. Therefore, the Hessian is block diagonal with respect to the blocks of coordinates $(E^1_{ab}, E^1_{ba}, \dots, E^m_{ab}, E^m_{ba})$. Denote $\varepsilon = \Gamma^E_{ab}$, $\varepsilon' = \Gamma^E_{ba}$, $\beta = \Gamma^S_{ab}$ and $\beta' = \Gamma^S_{ba}$. The linear operator for the block is:

$$K(\varepsilon, \varepsilon', \beta, \beta') = \left(\begin{array}{cc|cc|c|cc}
\varepsilon+\beta & 1 & \beta & 0 & \dots & \beta & 0 \\
1 & \varepsilon'+\beta' & 0 & \beta' & \dots & 0 & \beta' \\
\hline
\beta & 0 & \varepsilon+\beta & 1 & & \beta & 0 \\
0 & \beta' & 1 & \varepsilon'+\beta' & \ddots & 0 & \beta' \\
\hline
\vdots & \vdots & & \ddots & \ddots & \vdots & \vdots \\
\hline
\beta & 0 & \beta & 0 & \dots & \varepsilon+\beta & 1 \\
0 & \beta' & 0 & \beta' & \dots & 1 & \varepsilon'+\beta'
\end{array}\right)$$

The positivity of $\mathcal{H}$ is equivalent to the positivity of this operator for all pairs a, b. We now assume $\beta\beta' > 0$.

First, we should note that $K(\varepsilon, \varepsilon', \beta, \beta')$ is congruent to $K(\varepsilon\sqrt{\frac{\beta'}{\beta}}, \varepsilon'\sqrt{\frac{\beta}{\beta'}}, \sqrt{\beta\beta'}, \sqrt{\beta\beta'})$ via the basis

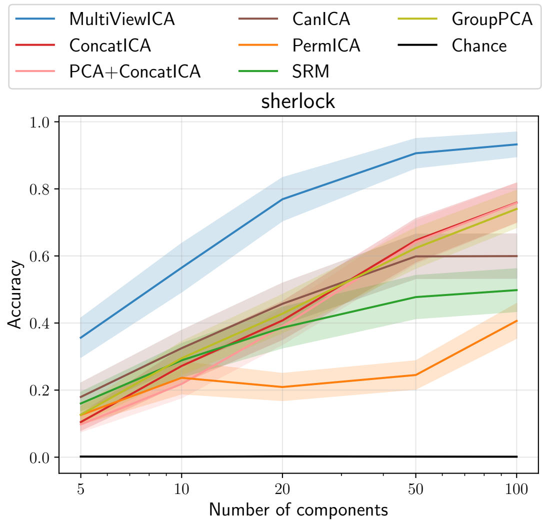


Figure A.1: **Reproducing the time-segment matching experiment of** [37] [157] Mean classification accuracy - error bars represent 95% confidence interval

$\mathrm{diag}((\frac{\beta'}{\beta})^{1/4}, (\frac{\beta}{\beta'})^{1/4}, \cdots, (\frac{\beta'}{\beta})^{1/4}, (\frac{\beta}{\beta'})^{1/4})$. We denote to simplify notation $\alpha = \epsilon\sqrt{\frac{\beta'}{\beta}}$, $\alpha' = \epsilon'\sqrt{\frac{\beta}{\beta'}}$ and $\gamma = \sqrt{\beta\beta'}$. We only have to study the positivity of $K(\alpha, \ \alpha', \gamma, \gamma)$. We have:

$$K(\alpha, \alpha', \gamma, \gamma) = I_m \otimes M_\alpha + \gamma \mathbb{1} \otimes I_2, \ \ M_\alpha = \begin{pmatrix} \alpha & 1 \\ 1 & \alpha' \end{pmatrix}$$

Since $I_m \otimes M_\alpha$ and $\gamma\mathbb{1} \otimes I_2$ commute, the minimum value of $\mathrm{Sp}(K)$ is $\min(I_m \otimes M_\alpha) + \min(\gamma \mathrm{Sp}(\mathbb{1})) = \frac{1}{2}(\alpha + \alpha' - \sqrt{(\alpha - \alpha')^2 + 4}) + m \min(0, \gamma)$. Since we assumed $\beta\beta' > 0$ we have $\gamma > 0$. This is similar to the usual ICA case, we find that the condition is $\alpha\alpha' > 1$.

If the following conditions hold for all pair of components a, b, the components are a local minimum of the cost function:

- $\Gamma^S_{ab}\Gamma^S_{ba} \geqslant 0$
- $\Gamma^E_{ab}\Gamma^E_{ba} > 1$

A.1.4 *Reproducing time-segment matching experiment*

We reproduce the time-segment matching experiments described in [37] and [157] and use two fold classification over runs instead of 5-fold as we have done in chapter 8. We used the sherlock data available at http://arks.princeton.edu/ark:/88435/dsp01nz8062179 and the full brain mask provided in the Python package associated with the paper. We applied high-pass filtering (140 s cutoff) and the time series of each voxel were normalized to zero mean and unit variance.

The results are available in Figure A.1.

A.2 RELATED WORK

The following table describes some usual method for extracting shared components from multiple subjects datasets. The column "Modality/-Components" describes the type of data for which each algorithm was *initially* proposed, even though each algorithm could be applied on any type of data. The components type can be either temporal if extracted components are time courses or spatial if they are spatial patterns.

Method	**Modality / Components**	**Dimension reduction**	**Description**
SRM [36]	fMRI / Temporal	SRM	The model is $\mathbf{x}_i = A_i\mathbf{s} + \mathbf{n}_i$, with *Gaussian* components and *orthogonal* mixing matrices A_i
GroupPCA [139]	fMRI / Spatial	GroupPCA	A memory efficient implementation of PCA applied on temporally concatenated data.
GIFT [28]	fMRI / Spatial	Individual PCA + Group PCA (on component-wise concatenated data)	Single-subject ICA is applied on the aggregated data
EEGIFT [48]	EEG / Temporal	Individual PCA + Group PCA (on component-wise concatenated data)	Single-subject ICA is applied on the aggregated data

PermICA	Any	Any	Single-subject ICA is applied on each subject's data, and the components are matched using the Hungarian algorithm
Clustering approach [53]	fMRI / Spatial	Individual PCA	Single-subject ICA is applied on each subject's data, and the components are matched using a hierarchical clustering algorithm.
Measure projection analysis [22]	EEG / Temporal	Individual PCA	Single-subject ICA is applied on each subject's data, and the components are matched using a hierarchical clustering algorithm.
TensorICA [18]	fMRI / Spatial	Group PCA (on spatially concatenated data)	TensorICA incorporates ICA assumptions into the PARAFAC model. The mixing matrices $A_1 \cdots A_n$ are such that $A_i = AD_i$ where A is common to all subjects and D_i are subject specific diagonal matrices.

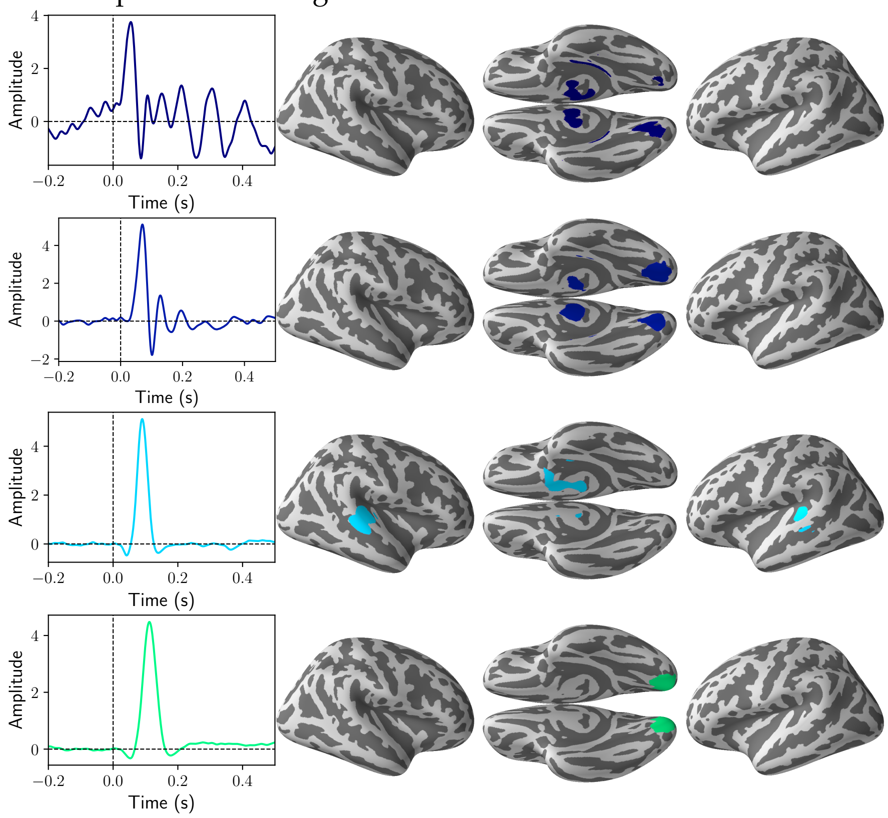
Unifying Approach of [64]	fMRI / Spatial	Group PCA (on spatially concatenated data) + Group-PCA (on component-wise concatenated data).	The model is $x_i = A_i s + n_i$ with a Gaussian mixture model on independent components and a matrix normal prior on the noise.
SR-ICA [157]	fMRI / Temporal	SR-ICA	SR-ICA incorporates ICA assumptions into the shared response model.
CAE-SRM [37]	fMRI / Temporal	CAE-SRM	A convolutional auto-encoder is used to perform the unmixing.
CanICA [150]	fMRI / Spatial	Individual PCA + multi set CCA (on component-wise concatenated data)	CanICA applies single-subject ICA on data reduced with PCA and CCA.
Spatial Concat-ICA [142]	fMRI / Spatial	Group PCA (on spatially concatenated data)	ICA is applied on spatially concatenated data. The mixing is constrained to be the same across all subjects.
Temporal Concat-ICA [39]	EEG / Temporal	Group PCA (on temporally concatenated data)	ICA is applied on temporaly concatenated data. The mixing is constrained to be the same across all subjects.

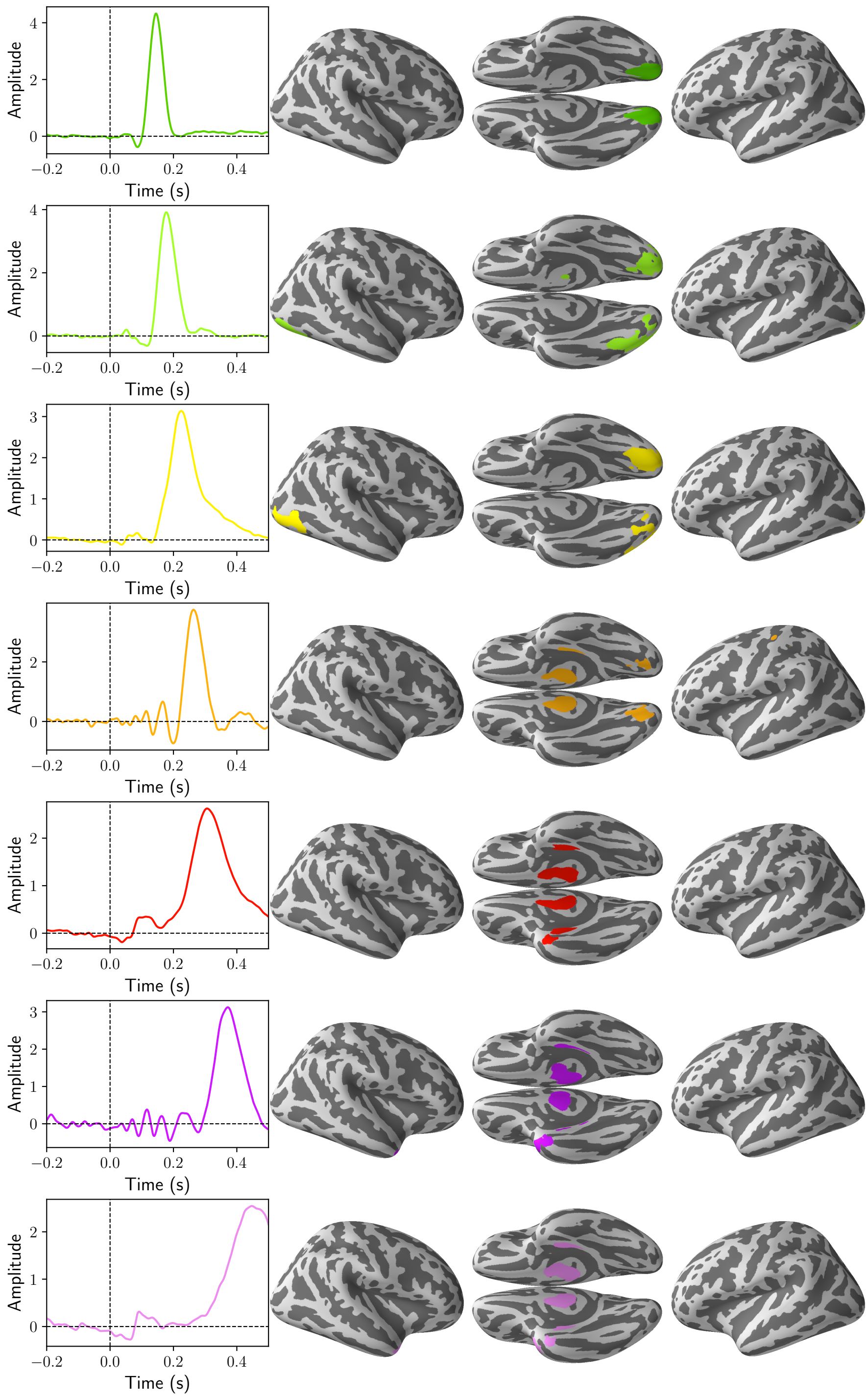
coroICA [117]	Any	Any	The model is $x_i = As_i + n_i$. The mixing is constrained to be the same across all subjects.

An additional related model is described in [61]. Similarly to our work, the ICA model has noise on the components side. However, the model involves nonlinear mixings, which are computationally unfeasible to optimize via maximum likelihood; a contrastive learning scheme is therefore adopted, and the likelihood is not derived in closed form. No evaluation on neuroimaging datasets is presented.

A.3 DETAILED CAM-CAN COMPONENTS

We display each of the 11 shared components found by Multiview ICA on the Cam-CAN. The time-courses are on the left, the corresponding brain maps are on the right.





A.4 AVERAGE FORWARD OPERATORS ON FMRI DATASETS

We display the average forward operator across subjects on the Raiders, Forrest, Clips and Sherlock datasets obtained with MultiViewICA and ConcatICA with 5 components. A 5 mm spatial smoothing was applied on all datasets, and the confound signals corresponding to the 5 components with the highest variance were removed before applying MultiViewICA or ConcatICA.

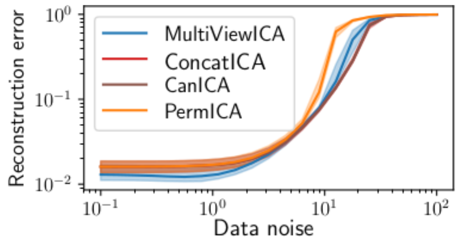
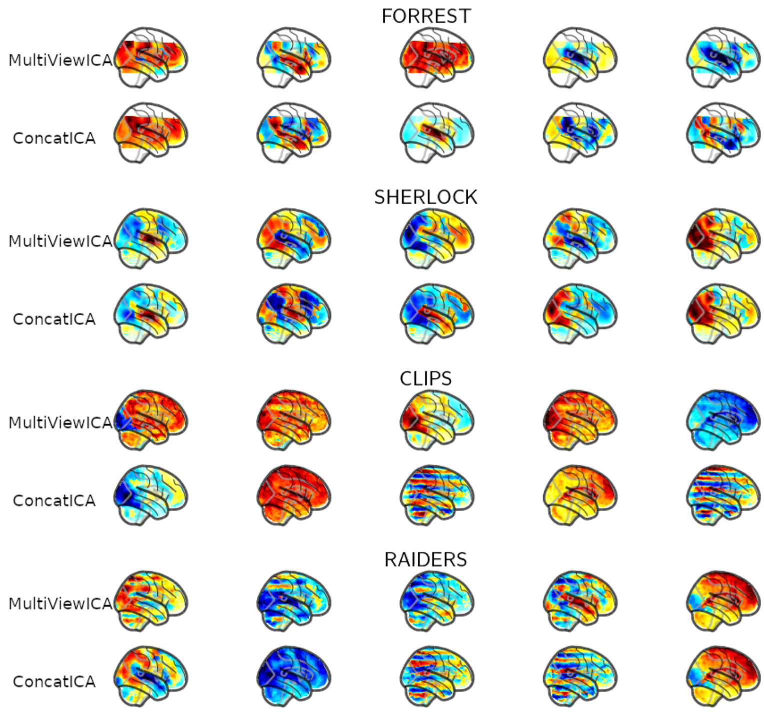


Figure A.2: Synthetic experiment with model $x_i = A_i s + n_i$



A.5 SYNTHETIC BENCHMARK USING ADDITIVE NOISE ON THE SENSORS

We generate data according to the model $\mathbf{x}_i = A_i \mathbf{s} + \mathbf{n}_i$, where $\mathbf{x}_i \in \mathbb{R}^{50}$, $\mathbf{s} \in \mathbb{R}^{20}$, and $\mathbf{n}_i \sim \mathcal{N}(0, \sigma^2 I_{50})$. After applying individual PCA to obtain signals of dimension 20, we apply the different ICA algorithms and report the reconstruction error in fig. A.2.

A.6 SUMMARY OF OUR QUANTITATIVE RESULTS

Our quantitative results for the fMRI experiments of time-segment matching and BOLD signal reconstruction and on for the MEG phantom data experiment are summarized, respectively, in Table A.2, Table A.3 and Table A.4. All methods are compared upon extraction of components with the same dimensionality (20 components).

Dataset	Method	Accuracy	Confidence interval
clips	Chance	0.002	[0.001, 0.003]
	CanICA	0.130	[0.112, 0.147]
	PCA + ConcatICA	0.124	[0.109, 0.139]
	ConcatICA	0.152	[0.133, 0.171]
	PermICA	0.147	[0.126, 0.169]
	SRM	0.115	[0.104, 0.126]
	MultiViewICA	**0.167**	[0.142, 0.192]
forrest	Chance	0.002	[0.001, 0.002]
	CanICA	0.192	[0.170, 0.214]
	PCA + ConcatICA	0.088	[0.077, 0.098]
	ConcatICA	0.154	[0.137, 0.170]
	PermICA	0.135	[0.118, 0.152]
	SRM	0.188	[0.173, 0.203]
	MultiViewICA	**0.448**	[0.411, 0.484]
raiders	Chance	0.002	[0.001, 0.003]
	CanICA	0.256	[0.220, 0.291]
	PCA + ConcatICA	0.331	[0.289, 0.372]
	ConcatICA	0.321	[0.281, 0.361]
	PermICA	0.381	[0.341, 0.421]
	SRM	0.265	[0.240, 0.289]
	MultiViewICA	**0.408**	[0.358, 0.458]
sherlock	Chance	0.005	[0.003, 0.006]
	CanICA	0.607	[0.567, 0.648]
	PCA + ConcatICA	0.454	[0.416, 0.492]
	ConcatICA	0.519	[0.481, 0.556]
	PermICA	0.399	[0.365, 0.434]
	SRM	0.493	[0.465, 0.520]
	MultiViewICA	**0.873**	[0.844, 0.903]

Table A.2: Timesegment matching: Summary of our quantitative results. We report the mean accuracy across cross-validation splits.

Dataset	Method	R2 score	Confidence interval
clips	Chance	0.000	[0.000 ,0.000]
	CanICA	0.110	[0.097 , 0.123]
	PCA + ConcatICA	0.075	[0.058 , 0.092]
	ConcatICA	0.077	[0.059 , 0.094]
	PermICA	0.099	[0.087 , 0.111]
	SRM	0.081	[0.069 , 0.094]
	MultiViewICA	**0.114**	[0.099 , 0.128]
forrest	Chance	0.000	[0.000 ,0.000]
	CanICA	0.181	[0.169 , 0.193]
	PCA + ConcatICA	0.072	[0.054 , 0.090]
	ConcatICA	0.081	[0.062 , 0.099]
	PermICA	0.098	[0.090 , 0.106]
	SRM	0.180	[0.168 , 0.193]
	MultiViewICA	**0.191**	[0.177 , 0.204]
raiders	Chance	0.000	[0.000 ,0.000]
	CanICA	0.136	[0.122 , 0.149]
	PCA + ConcatICA	0.063	[0.045 , 0.080]
	ConcatICA	0.062	[0.043 , 0.081]
	PermICA	0.107	[0.091 , 0.124]
	SRM	0.138	[0.121 , 0.154]
	MultiViewICA	**0.144**	[0.124 , 0.164]
sherlock	Chance	0.000	[0.000 ,0.000]
	CanICA	0.156	[0.141 , 0.172]
	PCA + ConcatICA	0.087	[0.065 , 0.108]
	ConcatICA	0.091	[0.070 , 0.112]
	PermICA	0.067	[0.055 , 0.078]
	SRM	**0.164**	[0.147 , 0.181]
	MultiViewICA	0.161	[0.142 , 0.180]

Table A.3: Reconstructing the BOLD signal of missing subjects: Summary of our quantitative results. We report the mean R2 score across cross-validation splits.

Method	Reconstruction error	1st and 3d quartiles
MultiViewICA	**0.0045**	[0.0039, 0.0052]
ConcatICA	0.1098	[0.0549, 0.1734]
PCA+ConcatICA	0.1111	[0.0760, 0.1502]
PermICA	0.0730	[0.0423, 0.1037]

Table A.4: Phantom MEG data: Summary of our quantitative results with 2 epochs. We report the median reconstruction error across cross-validation splits.

B SHICA

B.1 LEMMAS

Lemma 21. *Let $\mathbf{s} \in \mathbb{R}^k$ and $\mathbf{s}' \in \mathbb{R}^k$ have independent components among which g are Gaussian, and P a rotation matrix such that $\mathbf{s} = P\mathbf{s}'$. Then, $P = \Pi^{-1} O \Pi'$ where Π and Π' are sign and permutation matrices such that the first g components of $\Pi\mathbf{s}$ and $\Pi'\mathbf{s}'$ are Gaussian and O is a block diagonal matrix such that $O^{(g)}$, the first $g \times g$ block of O, is orthogonal and the other block is identity.*

Proof. From [38], Theorem 10: Assume $\mathbf{s} = P\mathbf{s}'$, if the column j of P has more than one non-zero element then s'_j is Gaussian.

Let us define permutations Π_1, Π'_1 such that the first g components of $\Pi_1\mathbf{s}$ and $\Pi'_1\mathbf{s}'$ are Gaussian and $P_1 = \Pi_1 P (\Pi'_1)^{-1}$. We can see that P_1 is orthogonal.

We have $\Pi_1 \mathbf{s} = P_1 \Pi'_1 \mathbf{s}'$. So the last $p - g$ columns of P_1 contain at most one non-zero element. Using orthogonality of P_1 this non-zero element has value 1 or -1 and is also the only one in its line. Let us focus on column $l > g$. Assume column l has its non-zero element at index $k \leqslant g$. Then line k in P_1 is only non-zero at index l and therefore $(\Pi_1\mathbf{s})_k$ (which is Gaussian) is equal to $(\Pi'_1\mathbf{s}')_l$ (which is not). Therefore column l can only have its non-zero element at an index greater than g. This shows that P_1 is block diagonal $P_1 = \begin{bmatrix} O_g & 0 \\ 0 & P_2 \end{bmatrix}$ where O_g is orthogonal and P_2 is a sign and permutation matrix.

$$\begin{bmatrix} O_g & 0 \\ 0 & P_2 \end{bmatrix} = \Pi_1 P (\Pi'_1)^{-1} \tag{B.1}$$

$$\iff \begin{bmatrix} O_g & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & P_2 \end{bmatrix} = \Pi_1 P (\Pi'_1)^{-1} \tag{B.2}$$

$$\iff \Pi_1^{-1} \begin{bmatrix} O_g & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & P_2 \end{bmatrix} \Pi'_1 = P \tag{B.3}$$

Therefore setting $\Pi' = \begin{bmatrix} I & 0 \\ 0 & P_2 \end{bmatrix} \Pi'_1$ and $\Pi = \Pi_1$ and $O = \begin{bmatrix} O_g & 0 \\ 0 & I \end{bmatrix}$ concludes the proof.

□

Lemma 22. *Assume that Assumption 2 holds for Σ_i, and that there is an orthogonal matrix P and diagonal matrices Σ'_i such that for all i, $\Sigma'_i = P \Sigma_i P^\top$. Then, P is a permutation matrix.*

Proof. The proof is in two parts. First, we show that there exist some coefficients $\alpha_1, \dots, \alpha_m$ such that the matrix $\sum_i \alpha_i \Sigma_i$ has distinct coefficients on the diagonal. Then, since we have $\sum_i \alpha_i \Sigma_i' = P\left(\sum_i \alpha_i \Sigma_i\right) P^\top$, and the diagonal $\sum_i \alpha_i \Sigma_i$ has distinct entries, we can invoke the unicity of the eigenvalue decomposition for symmetric matrices, which shows that P is necessarily a permutation matrix. Now, the only thing left is to prove is that Assumption 2 implies the existence of this linear combination.

We assume by contradiction that any linear combination of the Σ_i has two equal entries.

For $\alpha = [\alpha_1, \dots, \alpha_m]$, we let $\mathcal{S}(\alpha) = \mathrm{diag}(\sum_i \alpha_i \Sigma_i) \in \mathbb{R}^p$, where $\mathrm{diag}(\cdot)$ extracts the diagonal entries. The operator $\mathcal{S}$ is linear. We now define for $j, j' \leqslant p$ the linear form $\ell_{jj'}(\alpha) = \mathcal{S}(\alpha)_j - \mathcal{S}(\alpha)_{j'} \in \mathbb{R}$. The assumption on the linear combinations of Σ_i simply rewrites: For all $\alpha \in \mathbb{R}^m$, there exists $j, j' \leqslant p$ such that $\ell_{jj'}(\alpha) = 0$.

From a set point of view, this relationship writes

$$\bigcup_{j,j'} \mathrm{Ker}(\ell_{jj'}) = \mathbb{R}^m \quad .$$

Since the $\ell_{jj'}$ are all linear forms, the $\mathrm{Ker}(\ell_{jj'})$ are subspaces of dimensions m or $m-1$, and since their union is of dimension m, there exists j, j' such that $\mathrm{Ker}(\ell_{jj'}) = \mathbb{R}^m$, i.e. such that $\ell_{jj'} = 0$.

As a consequence, we have for all α, $\mathcal{S}(\alpha)_j = \mathcal{S}(\alpha)_{j'}$. This implies that the sequences $(\Sigma_{ij})_i$ and $(\Sigma_{ij'})_i$ are equal, which contradicts Assumption 2.

We have therefore shown that Assumption 2 implies the existence of a linear combination of the Σ_i that has distinct entries, which concludes the proof. □

Lemma 23. *Let us consider the following eigenvalue problem:*

$$\mathcal{A}z = \lambda \mathcal{B}z \tag{B.4}$$

$$\mathcal{A} = \begin{bmatrix} I+\Sigma_1 & I & \dots & I \\ I & I+\Sigma_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & I \\ I & \dots & I & I+\Sigma_m \end{bmatrix}$$

$$\mathcal{B} = \begin{bmatrix} I+\Sigma_1 & 0 & \dots & 0 \\ 0 & I+\Sigma_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & I+\Sigma_m \end{bmatrix}$$

where $\forall i,\ 1 \leqslant i \leqslant m,\ \Sigma_m \in \mathbb{R}^{p,p}$ *are positive diagonal matrices and I is the identity matrix. If the first* p *eigenvalues are distincts, the first* p *eigenvectors* $z^1, \dots, z^p, z^i \in \mathbb{R}^{mp}$ *have different first non-zero coordinates.*

Proof. We sort the eigenvectors in p groups of m vectors so that all vectors in group l have their l-th coordinate different from 0. Let $\boldsymbol{z}^{(l)}$ be an eigenvector in group l and let us call $\boldsymbol{w}_l \in \mathbb{R}^m$ the non-zero coordinates of this eigenvector: $\forall i \in \{1 \ldots m\}, w_{li} = z^{(l)}_{l+(i-1)p}$.

We have:

$$\mathcal{A}_l \boldsymbol{w}_l = \mathcal{B}_l \boldsymbol{w}_l \lambda_l \tag{B.5}$$

$$\mathcal{A}_l = \begin{bmatrix} 1+\Sigma_{1l} & 1 & \ldots & 1 \\ 1 & 1+\Sigma_{2l} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 1 \\ 1 & \ldots & 1 & 1+\Sigma_{ml} \end{bmatrix}$$

$$\mathcal{B}_l = \begin{bmatrix} 1+\Sigma_{1l} & 0 & \ldots & 0 \\ 0 & 1+\Sigma_{2l} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \ldots & 0 & 1+\Sigma_{ml} \end{bmatrix}$$

We now show that the biggest eigenvalue of (B.5) is strictly above 1 while all others are strictly below 1. The core of the proof comes from the study of the eigenvalues of a matrix modified by a rank 1 matrix. The reasoning we use here follows [57] (end of section 5).

Let us introduce $K^l = \mathrm{diag}(\Sigma_{1l} \ldots \Sigma_{ml})$ and $\mathbf{u} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$. Let us drop the index l in the notations for simplicity.

The problem can be rewritten

$$(\mathbf{u}\mathbf{u}^\top + K)\boldsymbol{w} = (I+K)\boldsymbol{w}\lambda \tag{B.6}$$

$$\iff (I+K)^{-1}(\mathbf{u}\mathbf{u}^\top + K)\boldsymbol{w} = \boldsymbol{w}\lambda \tag{B.7}$$

The characteristic polynomial is given by:

$$\mathcal{P}(\lambda) = \det((I+K)^{-1}K - \lambda I + (I+K)^{-1}\mathbf{u}\mathbf{u}^\top) \tag{B.8}$$

$$\propto \det(I + ((I+K)^{-1}K - \lambda I)^{-1}(I+K)^{-1}\mathbf{u}\mathbf{u}^\top) \tag{B.9}$$

where we implicitly focus here on eigenvalues λ such that $\det((I+K)^{-1}K - \lambda I) \neq 0 \iff \forall i, \lambda \neq \frac{k_i}{1+k_i}$.

We then use the following property: Let $A \in \mathbb{R}^{a,b}$ and $B \in \mathbb{R}^{b,a}$ we have $\det(I_a + AB) = \det(I_b + BA)$.

Let us call $\chi(\lambda) = \det(I + ((I+K)^{-1}K - \lambda I)^{-1}(I+K)^{-1}\mathbf{u}\mathbf{u}^\top)$ we have:

$$\chi(\lambda) = 1 + \mathbf{u}^\top((I+K)^{-1}K - \lambda I)^{-1}(I+K)^{-1}\mathbf{u} \tag{B.10}$$

$$= 1 + \sum_{i=1}^{m} \frac{1}{1+k_i} \frac{1}{\frac{k_i}{1+k_i} - \lambda} \tag{B.11}$$

where $k_i = \Sigma_{il} > 0$. Taking the derivative we get

$$\chi'(\lambda) = \sum_{i=1}^{m} \frac{1}{1+k_i} \frac{1}{(\frac{k_i}{1+k_i} - \lambda)^2} > 0 \tag{B.12}$$

Trivially, $\forall i$, $\frac{k_i}{1+k_i} < 1$. We also have

$$\chi(1) = 1 + \sum_{i=1}^{m} \frac{1}{1+k_i} \frac{1}{\frac{k_i}{1+k_i} - 1} = 1 - m < 0 \tag{B.13}$$

and $\lim_{\lambda \to +\infty} \chi(\lambda) = 1$ so as χ is continuous and strictly increasing on $[1, +\infty[$. Therefore, it reaches 0 only once on this interval (excluding 1 since we know $\chi(1) \neq 0$). Therefore the greatest eigenvalue λ^* is strictly above 1 while all other eigenvalues are strictly below 1.

Note that because $\chi' > 0$, λ^* is of multiplicity 1. In the analysis above we ignored those eigenvalues λ such that $\lambda = \frac{k_i}{1+k_i}$ for some i. However since $\frac{k_i}{1+k_i} < 1$, none of these eigenvalues can be the largest one.

Finally, the p first eigenvectors belong to different groups (the corresponding eigenvalues are all strictly above 1). This shows that these eigenvectors have different first non-zero coordinates.

□

B.2 IDENTIFIABILITY RESULTS FOR $m < 3$

We have a slightly weaker identifiability result when $m = 2$.

Proposition 24. *Let $m = 2$, and suppose that the scalars $(1 + \Sigma_{1j})(1 + \Sigma_{2j})$ for $j = 1 \dots p$ are all different. We let $\Theta' = (A_1', A_2', \Sigma_1', \Sigma_2')$ that also generates $\mathbf{x}_1, \mathbf{x}_2$. Then, there exists a permutation and scale matrix P such that $A_1' = A_1 P$ and $A_2' = A_2 P^{-\top}$.*

Proof. We let $P = A_1^{-1} A_1'$. Since $C_{12} = I_p$, it holds $A_2^{-1} A_2' = P^{-\top}$. Then, we have $I_p + \Sigma_1 = P(I_p + \Sigma_1')P^\top$. This means that there exists $U \in \mathcal{O}_p$ such that $P = (I_p + \Sigma_1)^{\frac{1}{2}} U (I_p + \Sigma_1')^{-\frac{1}{2}}$. Since $P^{-\top}(I_p + \Sigma_2')P^{-1} = I_p + \Sigma_2$, we find $U(I_p + \Sigma_1')(I_p + \Sigma_2')U^\top = (I_p + \Sigma_1)(I_p + \Sigma_2)$. By identification, U is a permutation matrix, and P is a scale and permutation matrix. □

As a consequence, when there are only two subjects, it is possible to recover the components and noise levels up to a scaling factor. When there is only one view, $m = 1$, there is a global rotation indeterminacy: $A_1(I_p + \Sigma_1)A_1^\top = A_1'(I_p + \Sigma_1)A_1'^\top$ for $A_1' = A_1(I_p + \Sigma_1)^{\frac{1}{2}} U (I_p + \Sigma_1)^{-\frac{1}{2}}$ where U is any orthogonal matrix. In this case, we lose identifiability.

B.3 EM E-STEP AND M-STEP FOR SHICA WITH GAUSSIAN COMPONENTS

B.3.1 *E-step*

The derivations are the same as in section B.4.1 but the sum over $\alpha \in \frac{1}{2}, \frac{3}{2}$ is replaced by just $\alpha = 1$.

B.3.2 *M-step*

The function to minimize in the M-step is then given by:

$$\mathcal{J} = \mathbb{E}[-\log p(\mathbf{x}, \mathbf{s})] \tag{B.14}$$

$$= \sum_{i=1}^{m} \log(|\Sigma_i|) + \tag{B.15}$$

$$\frac{1}{2}\operatorname{tr}(\Sigma_i^{-1}\left[\mathbb{E}[(\mathbf{y}_i - \mathbb{E}[\mathbf{s}|\mathbf{x}])(\mathbf{y}_i - \mathbb{E}[\mathbf{s}|\mathbf{x}])^\top] + \mathbb{V}[\mathbf{s}|\mathbf{x}]\right]) \tag{B.16}$$

$$+ c \tag{B.17}$$

where c does not depend on Σ_i

Therefore we get closed-form updates for Σ_i:

$$\Sigma_i \leftarrow \operatorname{diag}(\mathbb{E}[(\mathbf{y}_i - \mathbb{E}[\mathbf{s}|\mathbf{x}])(\mathbf{y}_i - \mathbb{E}[\mathbf{s}|\mathbf{x}])^\top] + \mathbb{V}[\mathbf{s}|\mathbf{x}]) \tag{B.18}$$

Plugging in the closed-form formula for $\mathbb{E}[\mathbf{s}|\mathbf{x}]$ and $\mathbb{V}[\mathbf{s}|\mathbf{x}]$ we get updates that only depends on the covariances $\hat{C_{ij}} = \mathbb{E}[\mathbf{x}_i \mathbf{x}_j^\top]$.

$$\begin{aligned}\Sigma_i \leftarrow \operatorname{diag}(&\hat{C_{ii}} \\ &- 2\mathbb{V}[\mathbf{s}|\mathbf{x}] \sum_{j=1}^{m} \Sigma_j^{-1}\hat{C}_{ji} \\ &+ \mathbb{V}[\mathbf{s}|\mathbf{x}] \sum_{j=1}^{m}\sum_{l=1}^{m} \left(\Sigma_j^{-1}\hat{C}_{jl}\Sigma_l^{-1}\right)\mathbb{V}[\mathbf{s}|\mathbf{x}] \\ &+ \mathbb{V}[\mathbf{s}|\mathbf{x}])\end{aligned}$$

B.4 EM E-STEP AND M-STEP FOR SHICA WITH NON-GAUSSIAN COMPONENTS

B.4.1 *E-step*

The complete likelihood is given by

$$p(\mathbf{x}, \mathbf{s}) = \prod_i p(\mathbf{x}_i|\mathbf{s}) p(\mathbf{s}) \tag{B.19}$$

$$= \prod_i p(\mathbf{x}_i|\mathbf{s}) \prod_j \sum_{\alpha \in \{0.5, 1.5\}} p(s_j|\alpha) \tag{B.20}$$

$$\tag{B.21}$$

where

$$p(s_j|\alpha) = \mathcal{N}(s_j; 0, \alpha) \tag{B.22}$$

We have

$$p(\mathbf{x}_i|\mathbf{s}) = |W_i|\mathcal{N}(\mathbf{y}_i; \mathbf{s}, \Sigma_i) \tag{B.23}$$

$$= |W_i| \prod_j \mathcal{N}(y_{ij}; s_j, \Sigma_{ij}) \tag{B.24}$$

where Σ_{ij} is the coefficient j, j of Σ_i and $\mathbf{y}_i = W\mathbf{x}_i$.

Let us introduce a first lemma:

Lemma 25.

$$\prod_{i=1}^{m} \mathcal{N}(x_i; u, v_i) = \prod_{i=1}^{m} \mathcal{N}(x_i; \bar{x}, v_i)\sqrt{2\pi\bar{v}}\mathcal{N}(\bar{x}; u, \bar{v})$$

where $\bar{v} = (\sum_{i=1}^{m} v_i^{-1})^{-1}$ *and* $\bar{x} = \frac{\sum_i v_i^{-1} x_i}{\sum_i v_i^{-1}}$.

Proof. We have that

$$\sum_i \frac{1}{v_i}(x_i - u)^2 = \sum_i \frac{1}{v_i}(x_i - u)^2 \tag{B.25}$$

$$= \sum_i \frac{1}{v_i}(x_i - \bar{x} + \bar{x} - u)^2 \tag{B.26}$$

$$= \sum_i \frac{1}{v_i}(x_i - \bar{x})^2 + \sum_i \frac{1}{v_i}(\bar{x} - u)^2 \tag{B.27}$$

and therefore

$$\prod_i (\frac{1}{\sqrt{2\pi v_i}} \exp(-\frac{1}{2v_i}(x_i - \mu)^2)) \tag{B.28}$$

$$= \prod_i \frac{1}{\sqrt{2\pi v_i}} \exp(\sum_i -\frac{1}{2}(\frac{1}{v_i}(x_i - \bar{x})^2 + \frac{1}{v_i}(\bar{x} - u)^2)) \tag{B.29}$$

$$= \prod_i \mathcal{N}(x_i, \bar{x}, v_i) \exp(-\frac{1}{2}(\sum_i \frac{1}{v_i})(\bar{x} - u)^2)) \tag{B.30}$$

$$\tag{B.31}$$

so the desired result follow. □

By Lemma 25, we have

$$\prod_i p(\mathbf{x}_i|\mathbf{s}) = \prod_i |W_i| \prod_j \mathcal{N}(y_{ij}; \bar{y}_j, \Sigma_{ij})\sqrt{2\pi\bar{\Sigma}_j}\mathcal{N}(\bar{y}_j; s_j, \bar{\Sigma}_j) \tag{B.32}$$

$$\tag{B.33}$$

where $\bar{y}_j = \frac{\sum_i \Sigma_{ij}^{-1} y_{ij}}{\sum_i \Sigma_{ij}^{-1}}$ and $\bar{\Sigma}_j = (\sum_i \Sigma_{ij}^{-1})^{-1}$. Hiding variable that do not depend on $\mathbf{s}$ we obtain

$$\prod_i p(\mathbf{x}_i|\mathbf{s}) \propto \prod_j \mathcal{N}(\bar{y}_j; s_j, \bar{\Sigma}_j) \tag{B.34}$$

$$\tag{B.35}$$

Then we get

$$p(\mathbf{x}, \mathbf{s}) \propto \prod_j \sum_{\alpha \in \{0.5, 1.5\}} \mathcal{N}(s_j; \bar{y}_j, \bar{\Sigma}_j) \mathcal{N}(s_j; 0, \alpha) \tag{B.36}$$

Let us now prove a second Lemma:

Lemma 26.

$$\mathcal{N}(x; y, \nu)\mathcal{N}(x, 0, \alpha) = \mathcal{N}(y; 0, \nu + \alpha)\mathcal{N}(x; \frac{\alpha y}{\alpha + \nu}, \frac{\nu\alpha}{\alpha + \nu})$$

Proof. We have

$$\mathcal{N}(x; y, \nu)\mathcal{N}(x, 0, \alpha) = \frac{\exp\left(-\frac{(x-y)^2}{2\nu}\right)}{\sqrt{2\pi\nu}} \frac{\exp\left(-\frac{x^2}{2\alpha}\right)}{\sqrt{2\pi\alpha}} \tag{B.37}$$

Then,

$$\exp\left(-\frac{(x-y)^2}{2\nu}\right) \tag{B.38}$$

$$= \exp\left(-\frac{\alpha(x-y)^2 + \nu x^2}{2\alpha\nu}\right) \tag{B.39}$$

$$= \exp\left(-\frac{\alpha(x^2 - 2xy + y^2) + \nu x^2}{2\alpha\nu}\right) \tag{B.40}$$

$$= \exp\left(-\frac{x^2(\alpha + \nu) - 2x(\alpha y) + \alpha y^2}{2\alpha\nu}\right) \tag{B.41}$$

$$= \exp\left(-\frac{x^2 - 2x\frac{\alpha y}{\alpha+\nu} + \frac{\alpha y^2}{\alpha+\nu}}{2\frac{\alpha\nu}{\alpha+\nu}}\right) \tag{B.42}$$

$$= \exp\left(-\frac{(x - \frac{\alpha y}{\alpha+\nu})^2 - (\frac{\alpha y}{\alpha+\nu})^2 + \frac{\alpha y^2}{\alpha+\nu}}{2\frac{\alpha\nu}{\alpha+\nu}}\right) \tag{B.43}$$

$$= \exp\left(-\frac{(x - \frac{\alpha y}{\alpha+\nu})^2}{2\frac{\alpha\nu}{\alpha+\nu}}\right) \exp\left(-\frac{-\alpha^2 y^2 + (\alpha + \nu)\alpha y^2}{2\alpha\nu(\alpha + \nu)}\right) \tag{B.44}$$

$$= \exp\left(-\frac{(x - \frac{\alpha y}{\alpha+\nu})^2}{2\frac{\alpha\nu}{\alpha+\nu}}\right) \exp\left(-\frac{\nu\alpha y^2}{2\alpha\nu(\alpha + \nu)}\right) \tag{B.45}$$

and

$$\frac{1}{\sqrt{2\pi\nu}\sqrt{2\pi\alpha}} = \frac{1}{\sqrt{2\pi(\nu + \alpha)}\sqrt{2\pi\frac{\nu\alpha}{\nu+\alpha}}} \tag{B.46}$$

so that the desired result follow. □

By Lemma 26, we have:

$$p(\mathbf{x}, \mathbf{s}) \tag{B.47}$$

$$\propto \prod_j \sum_{\alpha \in \{0.5, 1.5\}} \mathcal{N}(\bar{y}_j; 0, \bar{\Sigma}_j + \alpha)\mathcal{N}(s_j; \frac{\alpha \bar{y}_j}{\alpha + \bar{\Sigma}_j}, \frac{\bar{\Sigma}_j \alpha}{\alpha + \bar{\Sigma}_j}) \tag{B.48}$$

and therefore we get:

$$p(\mathbf{s}|\mathbf{x}) = \frac{p(\mathbf{s}, \mathbf{x})}{\int_s p(\mathbf{s}, \mathbf{x})} \tag{B.49}$$

$$= \prod_j \frac{\sum_{\alpha \in \{0.5, 1.5\}} \theta_\alpha \mathcal{N}(s_j; \frac{\alpha \bar{y}_j}{\alpha + \bar{\Sigma}_j}, \frac{\bar{\Sigma}_j \alpha}{\alpha + \bar{\Sigma}_j})}{\sum_{\alpha \in \{0.5, 1.5\}} \theta_\alpha} \tag{B.50}$$

where $\theta_\alpha = \mathcal{N}(\bar{y}_j; 0, \bar{\Sigma}_j + \alpha)$.

So we obtain the desired result:

$$\mathbb{E}[s_j|\mathbf{x}] = \frac{\sum_{\alpha \in \{0.5, 1.5\}} \theta_\alpha \frac{\alpha \bar{y}_j}{\alpha + \bar{\Sigma}_j}}{\sum_{\alpha \in \{0.5, 1.5\}} \theta_\alpha} \tag{B.51}$$

$$\mathbb{V}[s_j|\mathbf{x}] = \frac{\sum_{\alpha \in \{0.5, 1.5\}} \theta_\alpha \frac{\bar{\Sigma}_j \alpha}{\alpha + \bar{\Sigma}_j}}{\sum_{\alpha \in \{0.5, 1.5\}} \theta_\alpha} \tag{B.52}$$

B.4.2 *M-step*

The function to minimize in the M-step is then given by:

$$\mathcal{J} = \mathbb{E}[-\log p(\mathbf{x}, \mathbf{s})] \tag{B.53}$$

$$= \sum_{i=1}^{m} -\log(|W_i|) + \log(|\Sigma_i|) + \tag{B.54}$$

$$\frac{1}{2}\operatorname{tr}(\Sigma_i^{-1}\left[\mathbb{E}[(\mathbf{y}_i - \mathbb{E}[\mathbf{s}|\mathbf{x}])(\mathbf{y}_i - \mathbb{E}[\mathbf{s}|\mathbf{x}])^\top] + \mathbb{V}[\mathbf{s}|\mathbf{x}]\right]) + c \tag{B.55}$$

where c does not depend on Σ_i or W_i

Therefore we get closed-form updates for Σ_i:

$$\Sigma_i \leftarrow \operatorname{diag}(\mathbb{E}[(\mathbf{y}_i - \mathbb{E}[\mathbf{s}|\mathbf{x}])(\mathbf{y}_i - \mathbb{E}[\mathbf{s}|\mathbf{x}])^\top] + \mathbb{V}[\mathbf{s}|\mathbf{x}]) \tag{B.56}$$

We update W_i by performing a quasi-Newton step.
We use the relative gradient $\mathcal{G}^{W_i}$ and $\mathcal{H}^{W_i}$ defined by

$$\mathcal{J}(W_i + \varepsilon W_i) = \mathcal{J}(W_i) + \langle \varepsilon, \mathcal{G}^{W_i} \rangle + \frac{1}{2}\langle \varepsilon, \mathcal{H}^{W_i} \varepsilon \rangle \tag{B.57}$$

We get:

$$
\begin{aligned}
\mathcal{J}(W_i + \varepsilon W_i) = \sum_{i=1}^{m} \Bigg[&- \log(|W_i|) - \log(|I_k + \varepsilon|) \\
&- \mathbb{E}[\log(\mathcal{N}(y_i + \varepsilon y_i; s; \Sigma_i))] \Bigg] + \text{const} \qquad (B.58) \\
&= \mathcal{J}(W_i) - \mathrm{tr}(\varepsilon) + \frac{1}{2}\mathrm{tr}(\varepsilon^2) \\
&\quad + \frac{1}{2}\Big[\mathbb{E}[\langle \varepsilon y_i, (\Sigma_i)^{-1}(y_i - s)\rangle] + \\
&\quad \mathbb{E}[\langle (y_i - s), (\Sigma_i)^{-1}\varepsilon y_i\rangle] + \mathbb{E}[\langle \varepsilon y_i, (\Sigma_i)^{-1}\varepsilon y_i\rangle]\Big] \\
&\quad + o(\|\varepsilon\|^2) \qquad (B.59) \\
&= \mathcal{J}(W_i) - \sum_a \varepsilon_{a,a} + \frac{1}{2}\sum_{a,b}\varepsilon_{a,b}\varepsilon_{b,a} \\
&\quad + \sum_{a,b}\varepsilon_{a,b}\left[\mathbb{E}[(\Sigma_i)^{-1}(y_i - s)(y_i)^\top]\right]_{a,b} \\
&\quad + \frac{1}{2}\sum_{a,b}\varepsilon_{a,b}\left[\mathbb{E}[(\Sigma_i)^{-1}\varepsilon y_i (y_i)^\top]\right]_{a,b} \\
&\quad + o(\|\varepsilon\|^2) \qquad (B.60) \\
&= \mathcal{J}(W_i) - \sum_a \varepsilon_{a,a} + \frac{1}{2}\sum_{a,b}\varepsilon_{a,b}\varepsilon_{b,a} \\
&\quad + \sum_{a,b}\varepsilon_{a,b}\left[\mathbb{E}[(\Sigma_i)^{-1}(y_i - s)(y_i)^\top]\right]_{a,b} + \\
&\quad \frac{1}{2}\sum_{a,b,d}\varepsilon_{a,b}(\Sigma_i)^{-1}_{a,a}\varepsilon_{a,d}\left[\mathbb{E}[y_i(y_i)^\top]\right]_{d,b} \\
&\quad + o(\|\varepsilon\|^2) \qquad (B.61)
\end{aligned}
$$

So:

$$
\mathcal{G}^{W_i}_{a,b} = -\delta_{a,b} + \left[\mathbb{E}[(\Sigma_i)^{-1}(y_i - s)(y_i)^\top]\right]_{a,b} \qquad (B.62)
$$

and

$$
\mathcal{H}^{W_i}_{a,b,c,d} = \delta_{a,d}\delta_{b,c} + \delta_{a,c}\frac{\mathbb{E}[y_{ib}y_{id}]}{\Sigma_{ia}} \qquad (B.63)
$$

We approximate the Hessian by

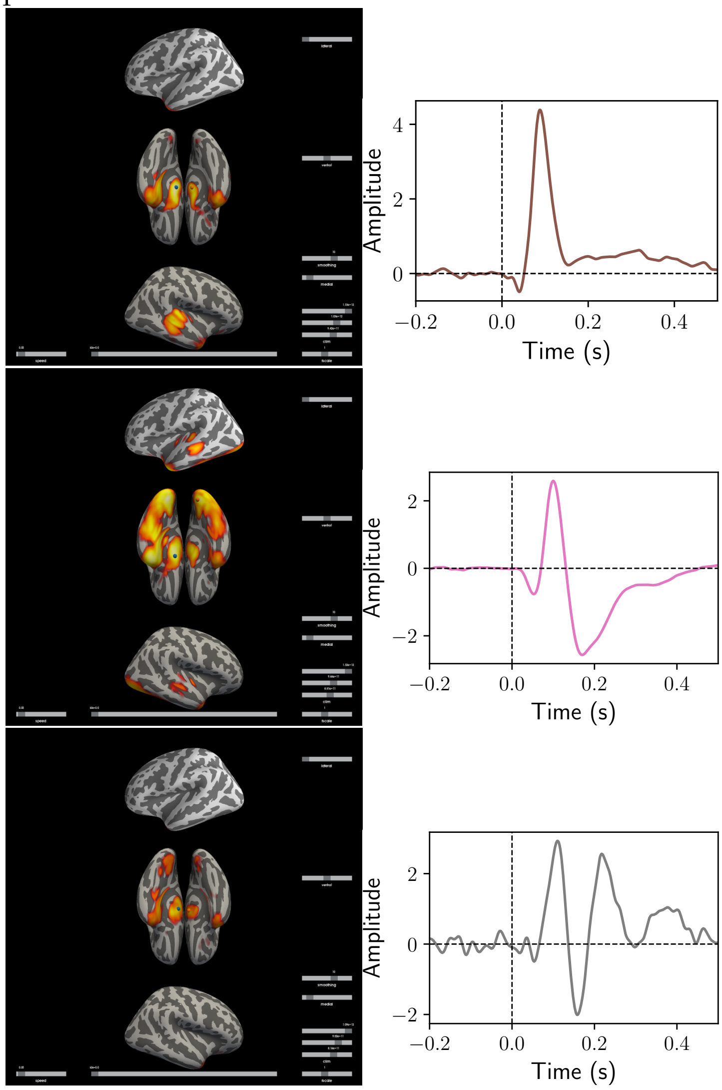
$$
\widehat{\mathcal{H}^{W_i}_{a,b,c,d}} = \delta_{ad}\delta_{bc} + \delta_{ac}\delta_{bd}\frac{\mathbb{E}[(y_{ib})^2]}{\Sigma_{ia}} \qquad (B.64)
$$

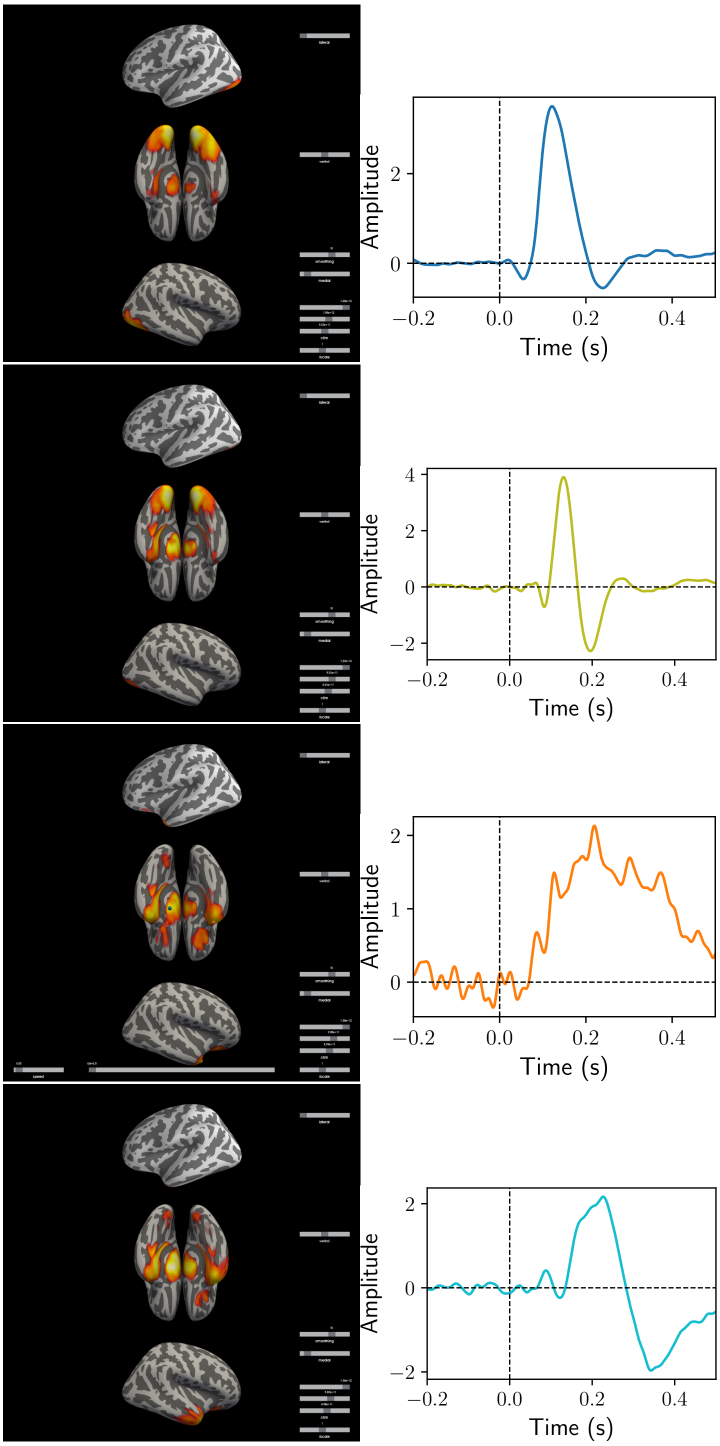
where the Hessian approximation is exact when the unmixed data have truly independent components.

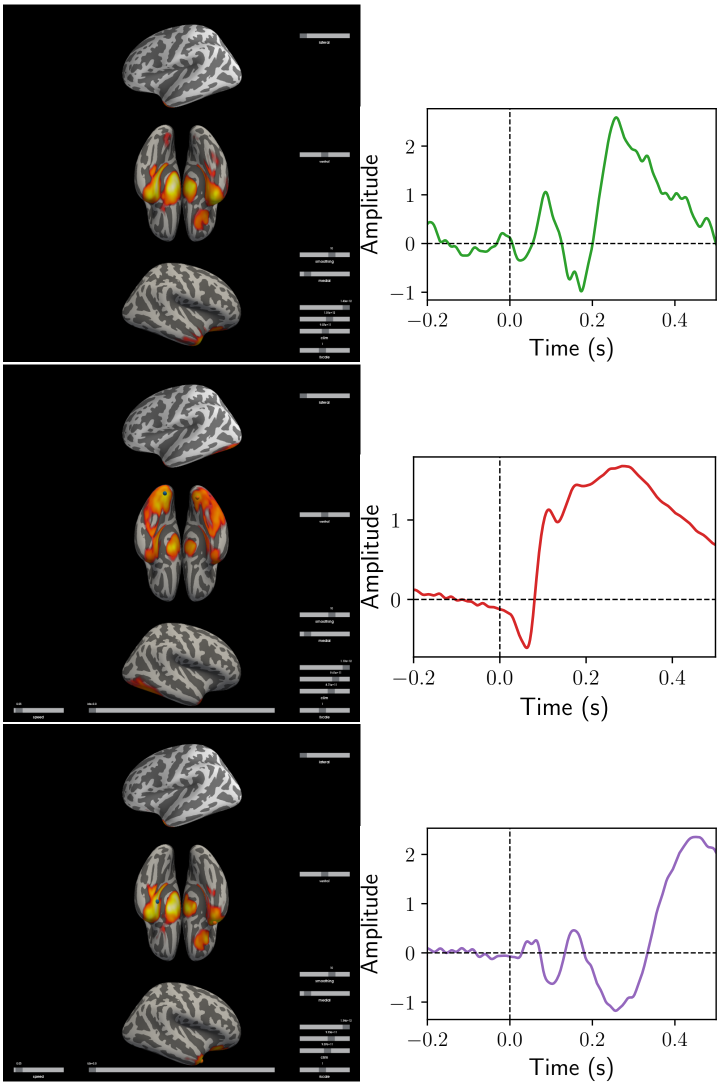
Updates for W_i are then given by $W_i \leftarrow (I - \rho(\widehat{\mathcal{H}^{W_i}})^{-1}\mathcal{G}^{W_i})W_i$, where ρ is chosen by backtracking line-search. We alternate between computing the statistics $\mathbb{E}[\mathbf{s}|\mathbf{x}]$ and $\mathrm{Var}[\mathbf{s}|\mathbf{x}]$ (E-step) and updates of parameters Σ_i and W_i for $i = 1 \dots m$ (M-step).

B.5 CAMCAN SPATIAL MAPS

We use $m = 496$ subjects and fit ShICA-ML with $p = 10$ components. We localize the components of each subject using sLORETA [113]. Then components are registered to a common brain and averaged. Thresholded maps are displayed below along with the time courses of each component. Spatial maps obtained with ShICA-ML highlight the ventral visual cortex and auditive cortex. The results suggest that the response of the auditive cortex is faster and lasts a shorter time than the response of the ventral visual cortex.




Amplitude
Time (s)
−0.2 0.0 0.2 0.4
0 2
Amplitude
Time (s)
−0.2 0.0 0.2 0.4
−2 0 2 4
Amplitude
Time (s)
−0.2 0.0 0.2 0.4
0 1 2
Amplitude
Time (s)
−0.2 0.0 0.2 0.4
−2 0 2


Amplitude
Time (s)
Amplitude
Time (s)
Amplitude
Time (s)