NNT : 2017SACLS297

Amélioration de connectivité fonctionnelle par utilisation de modèles déformables dans l'estimation de décompositions spatiales des images de cerveau

Thèse de doctorat de l'Université Paris-Saclay
préparée à l'Université de Paris-Sud

École doctorale n°580 : sciences et technologies de l'information et de la communication (STIC)

Spécialité de doctorat: **Informatique**

Thèse présentée et soutenue à Gif/Yvette, le 26 septembre 2017, par

Elvis Dohmatob

Composition du Jury :

Marc Schoenauer, Directeur de recherche, TAO Team, INRIA, Saclay, France	Président
John Ashburner, Directeur de recherche, UCL, Londre, Royaume-Uni	Rapporteur
Gabriel Peyré, Directeur de recherche, CNRS et ENS, Paris, France	Rapporteur
Moritz Grosse-Wentrup Directeur de recherche, MPI, Tuebingen, Allemagne	Examinateur
Gael Varoquaux, Directeur de recherche, Parietal Team, INRIA, Saclay, France	Examinateur
Bertrand Thirion, Directeur de recherche, Parietal Team, INRIA, Saclay, France	Directeur de thèse

Thèse de doctorat

université PARIS-SACLAY

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

Titre : Amélioration de connectivité fonctionnelle par utilisation de modèles déformables dans l'estimation de décompositions spatiales des images de cerveau

Mots clés : connectivité fonctionnelle, recalage non-linièaire, modèle deformables, optimisation, machine learning, IRMf

Résumé : Cartographier la connectivité fonctionnelle du cerveau à partir des donnés d'IRMf est devenu un champ de recherche très actif. Cependant, le outils théoriques et pratiques sont limités et plusieurs tâches importantes, telles que le définition empirique de réseaux de connexion cérébrale, restent difficiles à de l'absence d'un cadre pour la modélisation statistique de ces réseaux. Nous proposons de développer au niveau des populations, des modèles joints de connectivité anatomique et fonctionnelle et l'alignement inter-sujets des structures du cerveau. Grâce à une telle contribution, nous allons développer des nouvelles procédures d'inférence statistique afin de mieux comparer la connectivité fonctionnelle entre différent sujets en présence du bruit (bruit scanner, bruit physiologique, etc.).

Title : Enhancement of functional brain connectome analysis by the use of deformable models in the estimation of spatial decompositions of the brain images

Keywords : functional connectivity, nonlinear registration, deformable models, optimization, machine learning, fMRI

Abstract : Mapping the functions of the human brain using fMRI data has become a very active field of research. However, the available theoretical and practical tools are limited and many important tasks like the empirical definition of functional brain networks, are difficult to implement due to lack of a framework for statistical modelling of such networks. We propose to develop at the population level, models that jointly perform estimation of functional connectivity and alignment the brain data across the different individuals / subjects in the population. Building upon such a contribution, we will develop new methods for statistical inference to help compare functional connectivity across different individuals in the presence of noise (scanner noise, physiological noise, etc.).

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France

Dedicated *to my parents: Florence & Henry Dohmatob.*
And to my sister and brothers: Kah, Kingsly, Valy.
–The most wonderful people I know!–

Acknowledgements

Many thanks to my PhD advisors Bertrand Thirion and Gael Varoquaux for their help and support during these 3 years preparing my PhD project. Special thanks to the Parietal team assistants Regine Bricquet and Tiffany Caristan for their generous help with administrative matters. I would also want to thank the rest of the Parietal team –past and present– for ensuring such a nice work environment. Finally, I am thankful to all the members of the jury: John Ashburner, Gabriel Peyré, Marc Schoenauer, and Moritz Grosse-Wentrup, for kindly accepting to evaluate my PhD project.

Contents

Notation

Throughout the manuscript, we shall use the following standard notation and terminology without further explanation.

Notation	Name / synopsis	Definition
$[\![k]\!]$	Integers from 1 to k (inclusive)	$\{1,\ 2, \ldots,\ k\}$
$\|\mathbf{v}\|_r$	ℓ_r norm for vectors in a finite-dimensional Euclidean space	$\begin{cases} \sqrt[r]{\sum_i \|v_i\|^r}, & \text{if } 1 \leq r < \infty, \\ \max_i \|v_i\|, & \text{if } r = \infty \end{cases}$
$\mathcal{H}$	Arbitrary Hilbert space with norm inner product $\langle .,. \rangle_{\mathcal{H}}$ and norm $\mathbf{v} \mapsto \|\mathbf{v}\|_{\mathcal{H}} := \sqrt{\langle \mathbf{v}, \mathbf{v} \rangle_{\mathcal{H}}}$	Normed space containing its Cauchy limits
$H(X)$	Entropy of a random variable $X \sim p_X$	$\sum_x p_X(x) \log(p_X(x))$
$MI(X_1, X_2)$	Mutual Information between two random variables $X_1 \sim p_{X_1}$ and $X_2 \sim p_{X_2}$	$H(X_1) - H(X_2 \| X_1)$
$NMI(X_1, X_2)$	Normalized Mutual Information between two random variables $X_1 \sim p_{X_1}$ and $X_2 \sim p_{X_2}$	$I(X_1, X_2) / \sqrt{H(X_1) H(X_2)}$
$\mathbb{B}_{r,n}$	Unit ball for the ℓ_r norm on $\mathbb{R}^n$	$\{\mathbf{x} \in \mathbb{R}^n \| \|\mathbf{x}\|_r \leq 1\}$
$\mathrm{tr}(\mathbf{A})$	Trace of a matrix	$\sum_i a_{ii}$
$\langle \mathbf{A}, \mathbf{B} \rangle_{\mathrm{Fro}}$	Frobenius / Hilbert-Schmidt inner-product of two matrices $\mathbf{A}$ and $\mathbf{B}$	$\mathrm{tr}(\mathbf{A}\mathbf{B}^T)$
$\|\mathbf{X}\|_{\mathbf{Fro}}$	Frobenius norm of a matrix	$\sqrt{\langle \mathbf{X}, \mathbf{X} \rangle_{\mathrm{Fro}}}$
$\|\mathbf{X}\|_2$	Spectral norm of matrix	$\sup\{\|\mathbf{X}\mathbf{u}\|_2 \text{ s.t } \|\mathbf{u}\|_2 \leq 1\}$
$\|\mathbf{X}\|_{r,s}$	Mixed-norm of matrix $\mathbf{X} \in \mathbb{R}^{n \times m}$	$\left\| [\|\mathbf{X}_1\|_r, \|\mathbf{X}_2\|_r, \ldots, \|\mathbf{X}_n\|_r] \right\|_s$
$\mathbf{I}_n$	Identity matrix of size n	$I_{ij} = \delta_{ij},\ \forall 1 \leq i, j \leq n$
$\mathbf{1}_n$	Vector of ones of size n	$1_i = 1,\ \forall 1 \geq i \geq n$
$\mathbf{X}^\dagger$	Moore-Penrose pseudoinverse	Generalized inverse matrix
$\mathbf{A} \otimes \mathbf{B}$	Kronecker product of matrices $\mathbf{A}$ and $\mathbf{B}$	
$\mathbf{A} \circ \mathbf{B}$	Outer product of matrices $\mathbf{A}$ and $\mathbf{B}$	
$\mathrm{vec}(\mathbf{A})$	Vectorization of a matrix	Concatenation of the columns of a matrix into a single giant column vector
i_C	Indicator function of C	$i_C(\mathbf{x}) := \begin{cases} 0, & \text{if } \mathbf{x} \in C \\ \infty, & \text{otherwise} \end{cases}$
σ_C	Support function of C	$\sigma_C(\mathbf{x}) := \sup_{\mathbf{z} \in C} \mathbf{x}^T \mathbf{z}$
$\mathrm{dom}(f)$	Effective domain of $f : \mathcal{H} \to (-\infty, +\infty]$	$\{\mathbf{x} \in \mathbb{R}^n \| f(\mathbf{x}) < +\infty\}$

$\partial f(x)$	Subdifferential of f at $\mathbf{x}$	$\partial f(\mathbf{x}) := \{\mathbf{v} \in \mathcal{H} \mid f(\mathbf{z}) \geq f(\mathbf{x}) + \mathbf{v}^T(\mathbf{z}-\mathbf{x})\ \forall \mathbf{z} \in \mathcal{H}\}$
$\text{prox}_f(x)$	Proximal operator of f at $\mathbf{x}$	$\arg\min_{\mathbf{p}} \frac{1}{2}\|\mathbf{p}-\mathbf{x}\|_2^2 + f(\mathbf{p})$
$\text{proj}_C(x)$	Orthogonal projection of $\mathbf{x}$ onto C	$\arg\min_{\mathbf{p}\in C} \frac{1}{2}\|\mathbf{p}-\mathbf{x}\|_2^2 = \text{prox}_{i_C}(\mathbf{x})$
f^*	Convex conjugate of f	$f^*(\mathbf{x}) := \sup_{\mathbf{y}} \mathbf{x}^T\mathbf{z} - f(\mathbf{z})$
L_F	Lipschitz constant of $F : \mathcal{H}_1 \to \mathcal{H}_2$	$\inf\{C \geq 0 \mid \|F(\mathbf{x}) - F(\mathbf{y})\|_{\mathcal{H}_2} \leq C\|\mathbf{x}-\mathbf{y}\|_{\mathcal{H}_1}\ \forall \mathbf{x}, \mathbf{y} \in \mathcal{H}_1\}$
∇	Discrete spatial gradient operator. This defines a linear operator from $\mathbb{R}^p$ to $\mathbb{R}^{3p}$, where p is the number of voxels in the image	At a voxel j, the spatial gradient of an image $\mathbf{w}$ is a vector $\nabla\mathbf{w}(j) := [\nabla_x\mathbf{w}(j), \nabla_y\mathbf{w}(j), \nabla_z\mathbf{w}(j)], \forall \mathbf{w} \in \mathbb{R}^p$
Δ	Discrete spatial image Laplacian operator	$-\nabla^T\nabla \in \mathbb{R}^{p\times p}$
∇_ρ	The identity-augmented version of the discrete spatial gradient operator	$\nabla_\rho\mathbf{w} := [(1-\rho)\nabla\mathbf{w}, \rho\mathbf{w}] \in \mathbb{R}^{4p}, \forall \mathbf{w} \in \mathbb{R}^p$
Δ_ρ	Laplacian operator corresponding to the identity-augmented spatial gradient operator ∇_ρ . This defines a linear operator from $\mathbb{R}^p$ to $\mathbb{R}^{4p}$	$\rho^2\mathbf{I} + (1-\rho)^2\Delta \in \mathbb{R}^{p\times p}$
$\text{Lap}(\mathbf{w})$	Laplacian regularization of a 3D image $\mathbf{w}$	$\frac{1}{2}\|\nabla\mathbf{w}\|_{\text{Fro}}^2 = \frac{1}{2}\sum_{j=1}^p (\nabla_x\mathbf{w})_j^2 + (\nabla_y\mathbf{w})_j^2 + (\nabla_z\mathbf{w})_j^2$
$\|\mathbf{w}\|_{\text{TV}}$	Isotropic Total-Variation (TV) regularization	$\|\nabla\mathbf{w}\|_{2,1} = \sum_j \sqrt{(\nabla_x\mathbf{w})_j^2 + (\nabla_y\mathbf{w})_j^2 + (\nabla_z\mathbf{w})_j^2}$
$\|\mathbf{w}\|_{\text{SV}}$	Sparse Variation regularization	$\|\nabla_\rho\mathbf{w}\|_{2,1} = \sum_j \sqrt{\rho^2 w_j + (1-\rho)^2\|\nabla\mathbf{w})_j\|_2^2}$
$\|\mathbf{w}\|_{\text{AnisoTV}}$	Anisotropic TV regularization	$\|\nabla\mathbf{w}\|_{1,1} = \sum_j \|(\nabla_x\mathbf{w})_j\| + \|(\nabla_y\mathbf{w})_j\| + \|(\nabla_z\mathbf{w})_j\|$

Table 1: Notations

Part

I – Introduction

General overview of the thesis

Contents

1.1 Context

A MAJOR GOAL of the neurosciences is to understand the structure, function, and variability of the human brain, and how these give rise to the complex high-level behavior of human beings. One is typically interested in questions such as:

- *Which parts of the brain are in-charge of processing mathematical formulae as opposed to ordinary natural language ?*
- *Which parts of the brain increase/decrease their activity when the brain is at rest ?*
- *What are the neuro-biological markers of neurological or psychiatric mental illness ?*
- *How does the brain structure (sulci, gyri, etc.) and function change during aging ? etc.*
- *How do the language-responsive regions of one subject compare with that of another ? Can they be registered anatomically ?*
- *How are the different motor or cognitive functions (language, emotion, etc.) distributed over the brain, in terms of regions and networks of regions ?*
- *How are numbers represented and manipulated in the brain ?*
- *How does the brain and behavior change under the attack of a disease (e.g schizophrenia or a neuro-degenerative disease)*

Note that this list is by no way exhaustive.

In the last three decades, mapping brain functional connectivity from functional Magnetic Resonance Imaging (MRI) data has become a very active field of research. However, analysis tools are limited and many important tasks, such as the empirical definition of brain networks, remain

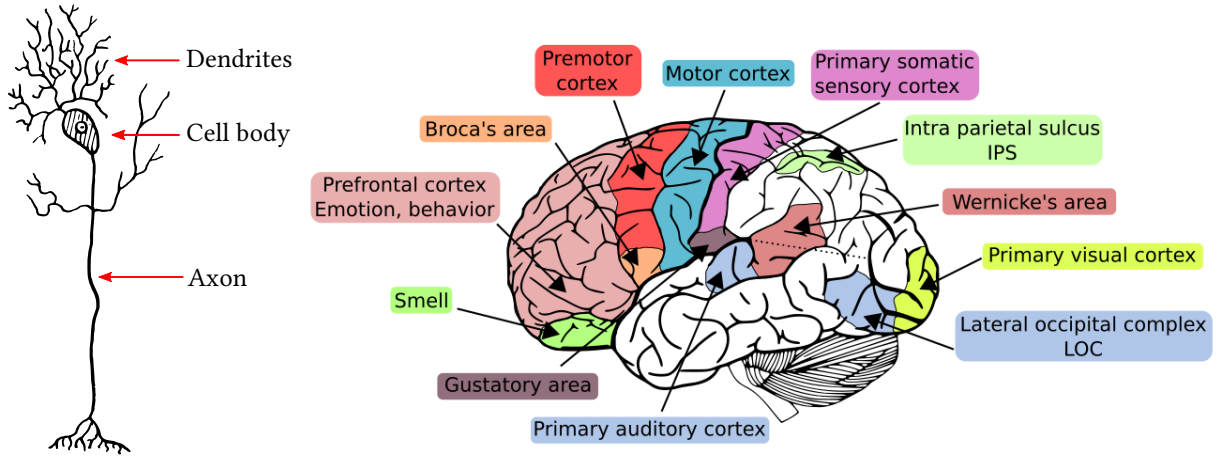


Figure 1.1: **Views of the brain** at different levels of detail. The brain is composed of (spatially connected) regions and such regions are in turn composed of populations of neurons. **Left:** Simplified view of a neuron. A neuron (there are many types) has a cell body called the *soma*, many regions for receiving information from other neural cells called *dendrites*, and often an *axon* (nerve fiber) for transmitting information to other cells (an axon can be longer than 1 meter in humans). The information in the axon is transmitted through an electrical signal called action potential, which is based on the electrical properties of the neuronal membrane. Adapted from `http://commons.wikimedia.org/`. **Right:** Each region is associated with a particular function such as sensory areas (e.g. visual cortex, auditory cortex) that receive and process information from sensory organs, motors areas (e.g. primary motor cortex, premotor cortex) that control the movements of the subject, and associative areas (e.g. Broca's area, Wernicke's area) that process the high-level information related to language production and understanding or the Intra Parietal Sulcus –IPS– that processes spatial information. Adapted from `http://agaudi.files.wordpress.com/`.

difficult due to the lack of a good framework for the statistical modeling of the data used to define these networks.

Objectives. The goal of this PhD thesis is to develop new statistical methods for studying inter-subject variability (eg. amplitude of activation, size of activation clusters, topography of activation maps, etc.), the prime goal being to improve the analysis of functional connectivity in the human brain at the population level. It turns out that these concerns naturally lead to problems related to data-driven extraction of functional atlases, multivariate models for brain decoding and segmentation, and inter-subject registration of functional MRI images.

1.2 Sketch of contributions

During the preparation of this PhD project, I have authored and co-authored a number of papers in conferences and journals (including NIPS, ICASSP, MICCAI, Frontiers in Neuroscience, etc.). A complete least of my publications can be found on my Google scholar page `https://scholar.google.fr/citations?user=FDWgJY8AAAAJ&hl=fr`. In figures,

- Total citations $\geq$ 194.
- Total papers (including co-authored papers) $\geq$ 15.
- h index $\geq$ 4.
- 110 index $\geq$ 3.

Below, I have roughly classified my main contributions under their respective sub-fields of relevance. Viz,

- Sparsity and spatial regularization: [Dohmatob et al., 2014], [Dohmatob et al., 2015b], [Abraham et al., 2014], [Eickenberg et al., 2015], [Pellé et al., 2016]
- Registration of brain images: [Dohmatob et al., 2016a]
- Optimization: [Dohmatob et al., 2015a], [Varoquaux et al., 2015], [Dohmatob, 2016]
- Modeling inter-subject functional variability: [Dohmatob et al., 2016b]
- Neuroscience: [Rahim et al., 2015], [Thirion et al., 2014]

There are also a number of preprints currently being prepared for journal publication:

- Sparsity and spatial regularization: *"Structured penalties for brain decomposition and decoding: a unified view"*
- *"Inter-subject registration of functional images: do we need anatomical images ?"*
- *"Enhanced prediction of task-based activation maps from resting-state data"*

1.3 Organization of the manuscript

In this report, I shall present a selection [1] of the work I have done during the preparation of my PhD project. This selection will be centered around

- **Part I:** General preliminaries on neurosciences and neuro-imaging methodology $\implies$ chapter 2.
- **Part II:** Structured penalties for brain decoding $\implies$ chapters 3, 4, 5, 6, 7.
- **Part III:** Functional inter-subject variability $\implies$ chapters 9, 10, 8, 11
- **Conclusion:** Summary and concluding remarks $\implies$ chapter 12.

My precise contributions in these domains will be comprehensively outlined as we proceed.

[1] For example, I shall not talk on excursional work I did on algorithmic non-cooperative game theory [Dohmatob, 2016].

Bibliography

Alexandre Abraham, Elvis Dohmatob, Bertrand Thirion, Dimitris Samaras, and Gael Varoquaux. Region segmentation for sparse decompositions: better brain parcellations from rest fMRI. In *Sparsity Techniques in Medical Imaging*, 2014.

Elvis Dohmatob. A simple algorithm for computing Nash-equilibria in incomplete information games. In *OPT2016 – NIPS workshop on optimization for machine learning*, 2016.

Elvis Dohmatob, Alexandre Gramfort, Bertrand Thirion, and Gaël Varoquaux. Benchmarking solvers for TV-l1 least-squares and logistic regression in brain imaging. In *PRNI*. IEEE, 2014.

Elvis Dohmatob, Michael Eickenberg, Bertrand Thirion, and Gael Varoquaux. Local Q-Linear Convergence and Finite-time Active Set Identification of ADMM on a Class of Penalized Regression Problems. In *ICASSP 2016*, 2015a.

Elvis Dohmatob, Michael Eickenberg, Bertrand Thirion, and Gaël Varoquaux. Speeding-up model-selection in GraphNet via early-stopping and univariate feature-screening. In *Pattern Recognition in NeuroImaging (PRNI), 2015 International Workshop on*. IEEE, 2015b.

Elvis Dohmatob, , Gaël Varoquaux, and Bertrand Thirion. Inter-subject highres EPI-to-EPI direct nonlinear registration outperforms classical T1-based method. In *Annual meeting of the Organization for Human Brain Mapping - 2016*, 2016a.

Elvis Dohmatob, Arthur Mensch, Gaël Varoquaux, and Thirion Bertrand. Learning brain regions via large-scale online structured sparse dictionary-learning. In *NIPS*, 2016b.

Michael Eickenberg, Elvis Dohmatob, Bertrand Thirion, and Gaël Varoquaux. Total Variation meets Sparsity: statistical learning with segmenting penalties. In *MICCAI*. 2015.

Hubert Pellé, Philippe Ciuciu, Mehdi Rahim, Elvis Dohmatob, Patrice Abry, and Virginie Van Wassenhove. Multivariate Hurst exponent estimation in fMRI. Application to brain decoding of perceptual learning. In *13th IEEE International Symposium on Biomedical Imaging*, 2016.

Mehdi Rahim, Bertrand Thirion, Alexandre Abraham, Michael Eickenberg, Elvis Dohmatob, Claude Comtat, and Gael Varoquaux. Integrating Multimodal Priors in Predictive Models for the Functional Characterization of Alzheimer's Disease. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer International Publishing, 2015.

Bertrand Thirion, Gaël Varoquaux, Elvis Dohmatob, and Jean-Baptiste Poline. Which fMRI clustering gives good brain parcellations? *Frontiers in neuroscience*, 8, 2014.

Gaël Varoquaux, Michael Eickenberg, Elvis Dohmatob, and Bertand Thirion. FAASTA: A fast solver for total-variation regularization of ill-conditioned problems with application to brain imaging. *arXiv:1512.06999*, 2015.

Introduction: functional magnetic resonance imaging to study the human brain

Contents

Neuroimaging has emerged as a distinguished data-acquisition and set of analysis techniques for probing and observing brain activity. Acquisition of the data goes hand-in-hand with statistical analysis methods for analyzing the data, in view of making specific quantifiable claims. These techniques operate at a scale much coarser than that of the *neuron*: one is interested physiological effects which are ultimately aggregates of activity over large population of neurons (see Fig. 1.1).

In this introductory chapter, I review the relevant theory sufficient to situate my own work in a larger scientific context. Section 2.1 will focus on imaging the human brain and preprocessing of the collected data and also classical methodologies for analyzing the data. Section 2.2 will present another celebrated way of probing brain function, namely resting-state fMRI –or the study of background spontaneous brain activity at rest.

2.1 Functional magnetic resonance imaging

Human neuroimaging consists in acquiring ex-vivo (non-invasively) image data from normal and diseased human populations. Several types of functional imaging techniques have been developed. Electro-encephalography (EEG) and magneto-encephalography (MEG) measure the superficial corti-

cal neural activity of the brain with a high temporal resolution. Functional magnetic resonance imaging (fMRI) [Ogawa et al., 1990a,b] uses strong magnetic fields to measure changes in oxygen flow in the brain that correlates with synaptic activity in the brain. This technique yields information on brain structure, variability, and function at high spatial resolution. Finally, invasive techniques have been developed such as positron emission tomography (PET) that relies on a radioactive tracer to track glucose consumption.

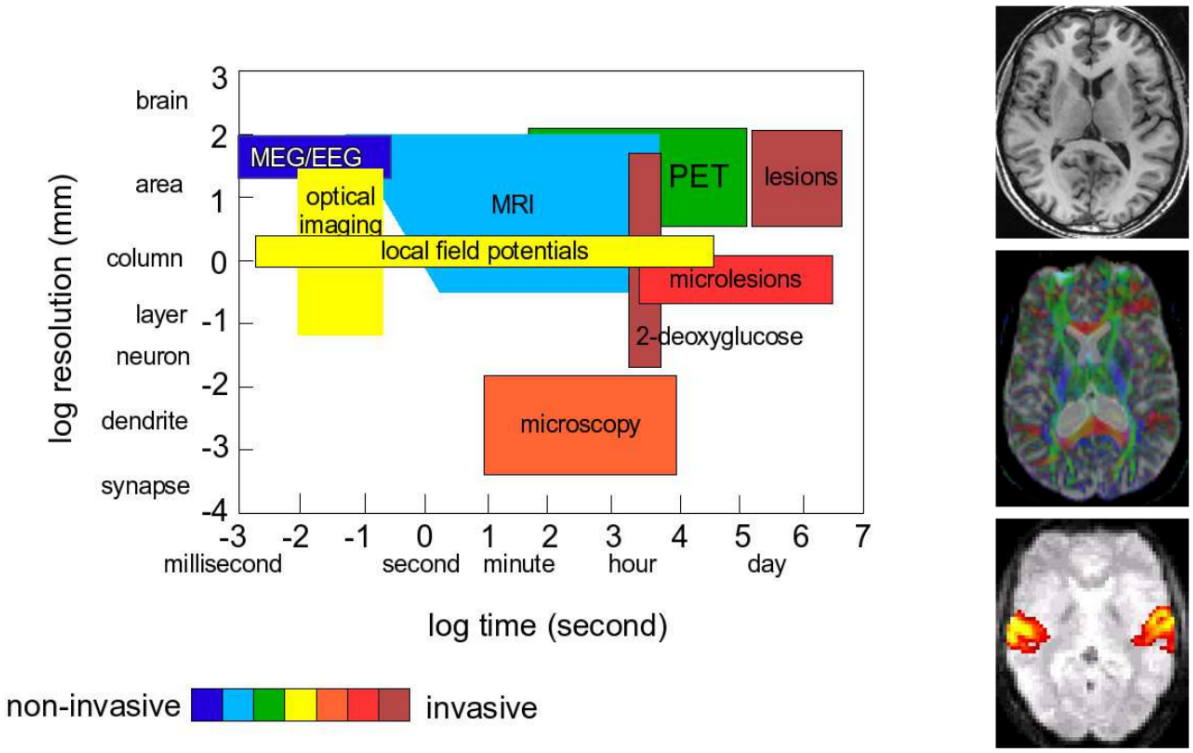


Figure 2.1: **Imaging modalities** for the brain. **Left:** The different imaging modalities for brain mapping. MRI and functional MRI have the unique property to yield high-resolution information while being minimally invasive. Unlike other modalities, MRI allows whole brain imaging. **Right:** Typical example of T1 / anatomical MRI (top), preprocessed Diffusion-Weighted (DW) MRI **middle** and fMRI **bottom** images, presented in axial views. These images are from the Neurospin 3T scanner. For the DW-MRI image, the main direction of water diffusion is color-coded: green for antero-posterior diffusion, red for lateral diffusion, blue for vertical diffusion. The functional image has been analyzed to yield the regions activated in an auditory task. Adapted with permission from [Thirion, 2009].

2.1.1 The BOLD signal

When a brain area is solicited, the brain fires chemical signals to *report* the consumption of oxygen and sugar. Nearby blood capillaries dilate to increase the quantity of flowing blood and provide these resources. This phenomenon is called the haemodynamic response. As a result, we expect a higher concentration of oxygenated hemoglobin in a given brain area soon after its activation. fMRI imaging can be used to measured this effect, called the BOLD (*Blood Oxygen-Level-Dependent*) signal [Ogawa et al., 1990a,b], at a spatial resolution of 1.5 to 3mm, and a temporal resolution of 1–3s, typically. This yields a spatially resolved image of brain functional networks

that can be modulated either by specific cognitive tasks or appear as networks of correlated activity. This method is subject to several physical and physiological noises. First, some artifacts may be induced by radio transmitters or other equipment. Then, spurious activations are naturally introduced by arteries present in the brain, heart beats and breathing movements. Finally, the brain can be shifted if the subject makes large movements in the scanner.

2.1.2 Preprocessing and analysis of fMRI data

Raw fMRI images are not intepretable with bare eyes. In particular because we are interested in small signal co-variation between voxels[1] and not by the values themselves. The human eye, however, is good at perceiving global artifacts in the data such as movements, ghost or scanner coils. Quality assessment of preprocessed fMRI data is done by eye and by relying on dedicated medical imaging software. In order to prepare the data for further statistical analysis, some preprocessing steps are required. Below, we outline the main ones. Viz,

[1] Voxel stands for *volume element*. It refers to a point in a 3D image, just as pixel refers to a point in a 2D image.

Data acquisition. The resolution of fMRI is usually between $1mm^3$ and $(3mm)^3$. In a single 3D scan, the brain represents 10^4 to 10^6 voxels. A run contains usually from 100 to 1000 scans. Functional MRI scans are acquired by slices, usually in the axial direction. The time required to acquire one slice is called echo time (TE) and is in the order of tens of milliseconds. The time required to acquire a whole 3D volume is called repetition time (TR) and is in the order of seconds. Typical values for a 3D volume of 60 slices are TE=33ms and TR=2s for a 3T (Tesla) scanner.

Motion-correction and coregistration to the anatomy. Head movement has a big impact on fMRI. A movement with an amplitude higher than the voxel resolution (i.e. 2 to 3mm) can shift the signal of the entire brain. Moreover, the worst impact of motion is inflow effects, i.e. artefactual signals. In the scanner, the head of the subject is fixed using cushion pads to avoid movements and the subject is asked to stay as still as possible. Yet, it is impossible to completely avoid head movement. In order to mitigate the effect of movement, the 3D scans are realigned on a reference scan –usually the one in the middle of the sequence– using rigid body transformation (translation and rotation, without change of scale). This is usually followed by an affine registration of the motion-corrected images to the anatomical (T1) image of the subject, in view of subsequent inter-subject preprocessing and analysis, like registration onto a group template (more on this later).

Slice-timing correction. As stated before, brain slices are not acquired at the same time. This introduces a shift in the haemodynamic response associated to each of them. The problem can be solved by interpolating the signal of each slice so that all of them can be considered as acquired at the same time. [Sladky et al., 2011] showed that depending on repetition time and paradigm design, slice-timing effects can significantly impair fMRI results and slice-timing correction methods can successfully compensate

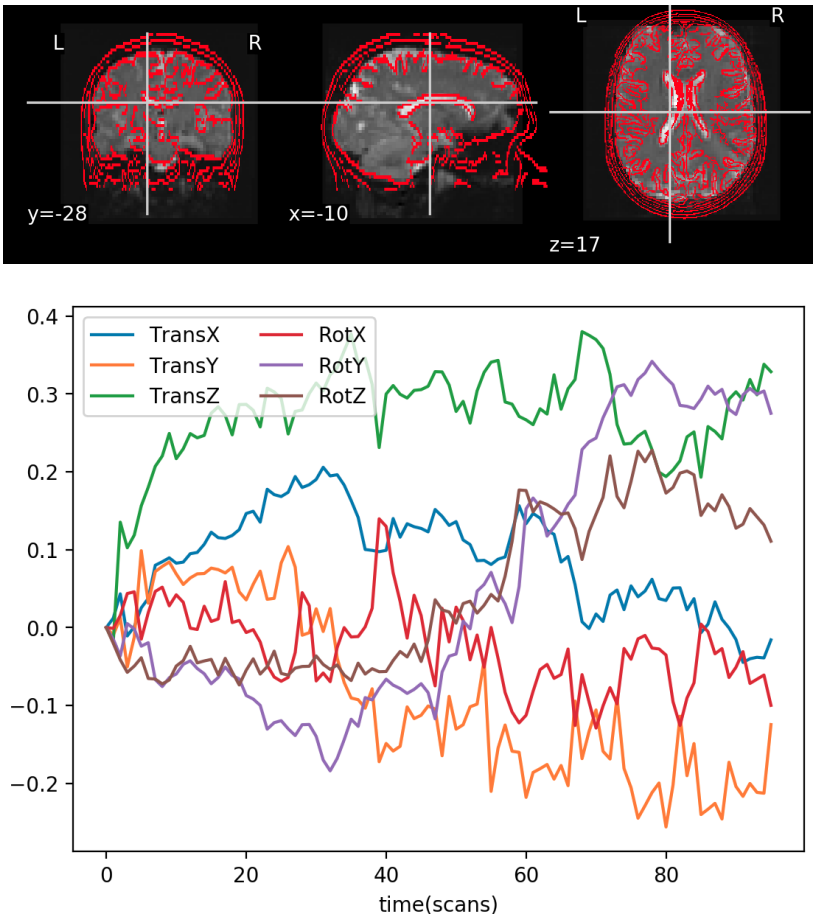


Figure 2.2: **Motion-correction and coregistration.** The **top** plot shows an overlay of a subjects anatomy onto their mean functional image (the background). Here the red contours are well aligned with the background image, indicative of a successful co-registration. Typical things that can go wrong include: lesions (missing brain tissue), bad orientation headers in the images, non-brain tissue in the images (e.g skull), etc. The **bottom** plot show estimated motion parameters for the subjects-head motion. Here all movements are well below 0.5mm, which is generally considered as fine. The preprocessing and plots displayed here were done using Pypreprocess `https://github.com/neurospin/pypreprocess`, an open-source Python wrapper built on standard toolkits like SPM [Friston et al., 1994].

for these effects and therefore increase the robustness of subsequent data analysis.

Registration of brain data into a common reference space. Each brain is of different size and shape. In order to compare brain activations across several individuals, we need to normalize them by registration to a common template [Friston et al., 1995, Ashburner and Friston, 2005, Ashburner, 2007, Klein et al., 2009]. This template can be a reference template used in the community (MNI for example). It is also possible to compute a template directly from the data. Once a template is chosen, for each subject, we perform two successive registrations. First, the anatomical scan acquired in the subject is registered to the MNI template. Then, the fMRI data are registered to the anatomical scan. After that, the two transformation matrices are combined in order to normalize the fMRI data to the template[2]. Estimation of the deformations necessary to warp a subject's brain anatomy onto a template is usually done alongside the classification of individual voxels into different classes: white matter (wm), grey matter (gm), and cerebro-spinal fluid (csf) producing so-called tissue probability maps (TPMs) [Ashburner and Friston, 2005].

[2] In chapter 8, we study the possibility of by-passing the anatomical image, when normalizing functional data.

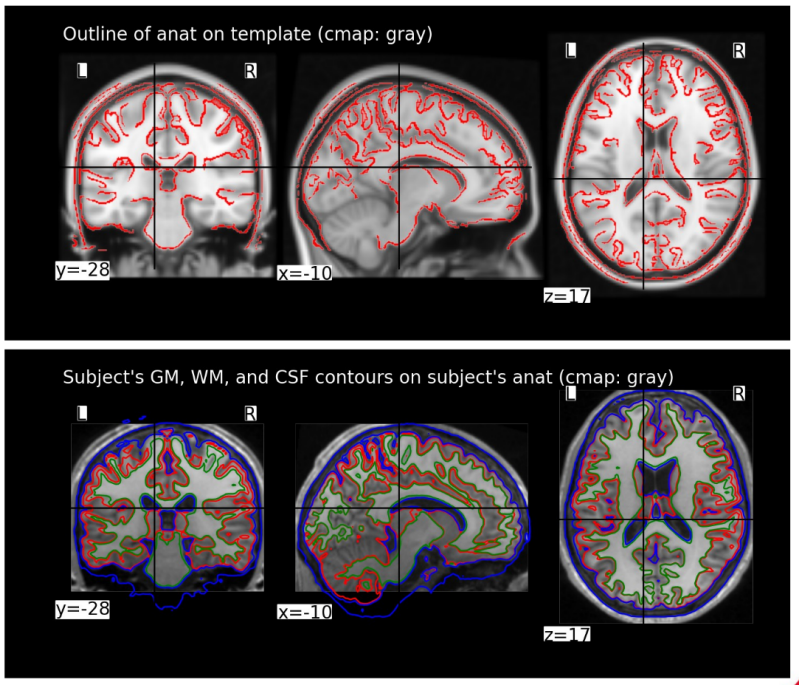


Figure 2.3: **Tissue segmentation and normalization.** Showing (**top**) outlines of a subject's anatomical / T1 image (foreground) projected onto an MNI template image (background) and also the tissue probability maps (TPMs), after registration to the latter. The contours should match the background image well. Typically impediments to correct registration include: lesions (missing brain tissue), corrupted image headers, non-brain tissue in anatomical image (i.e needs brain extraction), etc.

2.1.3 Statistical analysis of brain data

Forward inference made on fMRI data (e.g. prediction of brain activation from the stimuli) can be conceptualized as the *encoding* of perceptual, motor or cognitive parameters into brain signals. The inverse model, that predicts behavioral data from brain activation is called *decoding*, and will be the subject matter of chapter 3. Two main paradigms allow to experimentally study brain signals: either we study them in controlled condition on a particular

task –this is the task paradigm– or we study the spontaneous activity of the brain in order to uncover its organization: this is the resting-state paradigm.

The task paradigm and the general linear model

Using an experimental design, it is possible to relate the BOLD signal with specific tasks performed by the subject. For example, a sound can be played in the left or the right ear of the subject. By comparing brain activation between resting state and when the sound is playing, we can isolate the auditory cortex of the brain. Statistically, we do that by crafting a design matrix corresponding to the experiment: one column of the matrix represents an ideal response to one of the presented conditions [Friston et al., 1994]. Columns corresponding to known artifacts of the BOLD signal, such as heart beats or movements, can be added in the design matrix in order to regress out the part of the signal related to them. We then use a general linear model (GLM) to recover the brain maps corresponding to each of the columns in the design matrix $\mathbf{X} \in \mathbb{R}^{T \times k}$, where k is the number of conditions and T is the number of time points (times of repetition – TR). It is then supposed that for each voxel v, the measured BOLD signal $\mathbf{y}_v \in \mathbb{R}^T$ is a linear combination of the columns of $\mathbf{X}$, i.e of the experimental conditions, that is

$$\mathbf{y}_v = \mathbf{X}\boldsymbol{\beta}_v + \boldsymbol{\epsilon}_v, \tag{2.1}$$

where $\boldsymbol{\beta}_v \in \mathbb{R}^k$ are regression coefficients and $\boldsymbol{\epsilon}_v = (\epsilon_{v,1}, \ldots, \epsilon_{v,T}) \in \mathbb{R}^T$ is a non-iid vector of normally distributed noise. Such a problem is well-posed and weighted least-squares (WLS) are used to obtain a solution [3], to obtain $\hat{\boldsymbol{\beta}}_v = \mathbf{X}^\dagger \mathbf{y}_v$. Stacking these coefficients across all voxels per-brain correspond to k so-called $\boldsymbol{\beta}$-maps [Friston et al., 1994]. For a given combination of experimental conditions [4] $\mathbf{c} = (c_1, c_2, \ldots, c_k) \in \mathbb{R}^k$, one can compute a statistic

$$\hat{t}_{v,\mathbf{c}} := \frac{\mathbf{c}^T \hat{\boldsymbol{\beta}}_v}{\sqrt{var(\boldsymbol{\epsilon}_v)\mathbf{c}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{c}}}. \tag{2.2}$$

Under the null hypothesis that the effect were are interested in is zero, i.e

$$H_0 : \mathbf{c}^T \boldsymbol{\beta}_v = 0, \tag{2.3}$$

the above statistic is student-t distributed with $T - k$ degrees of freedom, and one can analytically obtain p-values and confidence intervals for inference. Projecting these values unto the brain (one value per voxel) yields a so-called *activation map*. Such maps are the main output of any forward analysis in task-based fMRI studies.

Subsequent statistical inference suffers from heavy *multiple comparison* issues in these so-called *mass-univariate* methods. The problem is further confounded by the fact that there are correlations between neighboring voxels, leading to situation where the Bonferoni and similar correction procedures, usually used to deal with these issues, may be too conservative and destroy the the sought-for effects. An alternative is to use *multi-variate* methods which directly model the spatial interactions between the voxels. Such methods will be the subject of chapter 3.

[3] There are usually more time points than experimental conditions, and so the design matrix is full-rank.

[4] For example, for $k = 3$ conditions, one may be interested in take $\mathbf{c} = (1, -1, 0)$, meaning we wish the find the effect of the first condition relative to the second, or $\mathbf{c} = (1, -1/2, -1/2)$ corresponding to the effect of the first condition w.r.t the average effect.

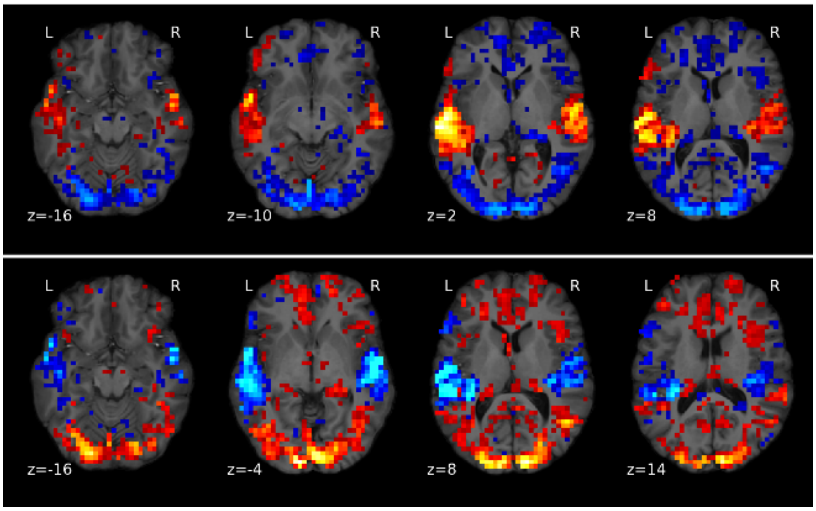


Figure 2.4: **Subject-level Activation maps** for auditory versus visual and visual versus auditory conditions. Here, we show an axial via (z-coordinate) of the Z-values corresponding the size of the effect, in each voxel. Values range from -13 (light blue) to +13 (light red). Analysis was done using the *Nistats* open-source Python library `https://github.com/nistats/nistats`.

2.2 Resting-state fMRI, brain networks, and functional connectivity

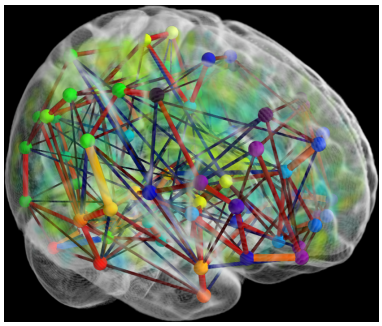


Figure 2.5: **Functional connectivity** patterns extracted from resting state data. The nodes are regions of the brain, and the thickness of the edges represent the relative strength average signal between the two corresponding regions.

Resting-state fMRI –or rsfMRI for short– uses the same acquisition method as task fMRI. However, instead of giving a particular task to the subject, they are asked to let their mind wander without sleeping. By studying this background activity of the brain, it is possible to uncover its underlying organization [Raichle, 2010]. Unlike the techniques described previously where the aim was to localize regions of activation for a given set of conditions, in functional connectivitly analysis were are interested in infering connections between such regions.

Depending on the protocol, the subject can be asked to keep eyes closed or to contemplate a fixation cross. The fixation cross prevents random eye movements and helps the subject not to sleep. In rsfMRI, we do not study the signal of each voxel itself but the interactions between the brain voxels. In particular, we study the functional connectivity of the brain, i.e. the similarity of activation patterns between brain regions that share a common functional role. Since there is no design matrix in rest fMRI, one must be careful to properly regress out physiological noises or spurious correlations may appear between brain regions, in particular longitudinally [Power et al., 2012, Van Dijk et al., 2012]. A first approach of functional connectivity is the voxel-to-voxel approach in which the similarity is measured between each pair of voxels. This method is not only computationally expensive, given the number of voxels in the brain, but it is also unfounded from the statistical standpoint: it requires the estimation of millions of parameters (one for each voxel pair), much more than the number of observations supports. As a consequence, some form of dimensionality reduction –a feature selection or extraction– is necessary to study connectivity.

2.3 Inter-subject functional variability

As noted in [Thirion et al., 2007, Thyreau et al., 2012, Xu et al., 2009], the inter-subject variability in GLM results (see Fig. 2.6) is not due to misregis-

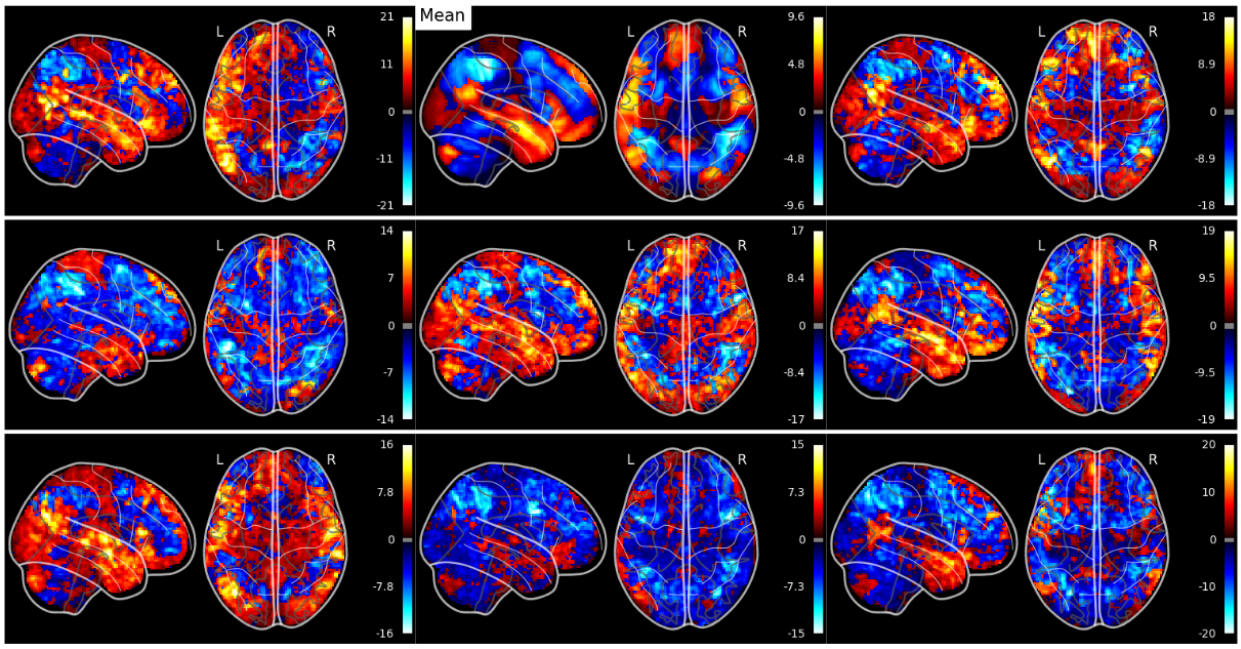


Figure 2.6: **Inter-subject functional variability.** Showing Z maps across different subjects for activation in Story vs Math language condition of the HCP –Human Connectome Project– dataset [van Essen et al., 2012]. The across-subject mean activation (top row, middle column) is also shown. Notice how the activations differ across subjects both in magnitude and spatial location.

tration, but intrinsic subject differences with a more physiological nature: the size of effects and the anatomical localization are subject-specific. Also, [Tavor et al., 2016] used dual regression [Filippini et al., 2009] to provide quantitative evidence that inter-subject differences in task-based brain activations are largely physiological –in contrast to being driven by subjects' brain morphological differences.

Chapters 9 and 11 will present generative models for understanding inter-subject variability at the functional level.

Bibliography

J. Ashburner. A fast diffeomorphic image registration algorithm. *Neuroimage*, 38, 2007.

J. Ashburner and K.J. Friston. Unified segmentation. *Neuroimage*, 26, 2005.

N. Filippini, B.J. MacIntosh, M.G. Hough, G.M. Goodwin, G.B. Frisoni, S.M. Smith, P.M. Matthews, C.F. Beckmann, and C.E. Mackay. Distinct patterns of brain activity in young carriers of the APOE-ε4 allele. *Proceedings of the National Academy of Sciences*, 106, 2009.

Karl J Friston, Andrew P Holmes, Keith J Worsley, J-P Poline, Chris D Frith, and Richard SJ Frackowiak. Statistical parametric maps in functional imaging: a general linear approach. *Human brain mapping*, 2(4):189–210, 1994.

Karl. J. Friston, J. Ashburner, C. D. Frith, J.-B. Poline, J. D. Heather, and R. S. J.

Frackowiak. Spatial registration and normalization of images. *Human Brain Mapping*, 3(3):165–189, 1995. ISSN 1097-0193. doi: 10.1002/hbm.460030303. URL http://dx.doi.org/10.1002/hbm.460030303.

A. Klein, J. Andersson, B. A. Ardekani, J. Ashburner, B. Avants, M. C. Chiang, G. E. Christensen, D. L. Collins, J. Gee, P. Hellier, J. H. Song, M. Jenkinson, C. Lepage, D. Rueckert, P. Thompson, T. Vercauteren, R. P. Woods, J. J. Mann, and R. V. Parsey. Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. *Neuroimage*, 46(3): 786–802, Jul 2009.

S. Ogawa, T. M. Lee, A. R. Kay, and D. W. Tank. Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proc Natl Acad Sci U S A*, 87(24):9868–9872, 1990a.

Seiji Ogawa, Tso-Ming Lee, Asha S. Nayak, and Paul Glynn. Oxygenation-sensitive contrast in magnetic resonance image of rodent brain at high magnetic fields. *Magnetic Resonance in Medicine*, 14(1):68–78, 1990b.

J. D. Power, K. A. Barnes, A. Z. Snyder, B. L. Schlaggar, and S. E. Petersen. Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. *Neuroimage*, 59(3):2142–2154, 2012.

Marcus E. Raichle. Two views of brain function. *Trends in Cognitive Sciences*, 14(4):180–190, 2010.

Ronald Sladky, Karl J. Friston, Jasmin Tröstl, Ross Cunnington, Ewald Moser, and Christian Windischberger. Slice-timing effects and their correction in functional MRI. *Neuroimage*, 58(2-2):588–594, Sep 2011.

I Tavor, O Parker Jones, RB Mars, SM Smith, TE Behrens, and S Jbabdi. Task-free MRI predicts individual differences in brain activity during task performance. *Science*, 352(6282):216–220, 2016.

Bertrand Thirion. *Structural and probabilistic methods for group analysis in functional neuroimaging*. PhD thesis, École normale supérieure de Cachan-ENS Cachan, 2009.

Bertrand Thirion, Philippe Pinel, Sébastien Mériaux, Alexis Roche, Stanislas Dehaene, and Jean-Baptiste Poline. Analysis of a large fMRI cohort: Statistical and methodological issues for group analyses. *Neuroimage*, 35 (1):105–120, 2007.

B. Thyreau, Y. Schwartz, B. Thirion, V. Frouin, E. Loth, S. Vollstadt-Klein, T. Paus, E. Artiges, P. J. Conrod, G. Schumann, R. Whelan, and J. B. Poline. Very large fMRI study using the IMAGEN database: sensitivity-specificity and population effect modeling in relation to the underlying anatomy. *Neuroimage*, 61(1):295–303, May 2012.

K.R.A. Van Dijk, M.R. Sabuncu, and R.L. Buckner. The influence of head motion on intrinsic functional connectivity MRI. *Neuroimage*, 59, 2012.

D.C. van Essen et al. The human connectome project: A data acquisition perspective. *NeuroImage*, 62(4), 2012.

Lei Xu, Timothy D. Johnson, Thomas E. Nichols, and Derek E. Nee. Modeling inter-subject variability in fMRI activation location: A Bayesian hierarchical spatial model. *Biometrics*, 65(4):1041–1051, Dec 2009.

Part

II – Multi-variate priors for analyzing brain data: models and algorithms

Structured priors for brain decoding and functional segmentation: the models

Contents

3.1 Introduction to brain decoding

As already discussed in chapter 2, functional brain imaging provides a distinctive opportunity to study brain functional architecture, while being minimally invasive, and is thus well-suited for the challenging study of the spatial layout of neural coding. Different modalities exist, each one having specific spatial and temporal resolutions; among them Functional Magnetic Resonance Imaging (fMRI) [Ogawa et al., 1990a,b] has emerged as a fundamental modality for brain imaging, striking a good balance between spatial and temporal resolution. fMRI images are pre-processed, and modeled through a general linear model (GLM), that takes into account the different experimental conditions and the dynamics of the haemodynamic response in the design matrix. The resulting model parameters, a.k.a. activation maps, represent the influence of the different experimental conditions on local fMRI signals.

The standard approach. The standard approach used for analyzing these activation maps is called classical inference, and relies on a mass-univariate statistical tests (one for each voxel), yielding the so-called statistical parametric maps (SPMs) [Friston et al., 1994]. Such maps are of particular interest in cognitive neuroscience, as they open the door to localizing the voxels that are significantly active for any combination of experimental conditions, and thus are probably implied in the underlying neural code of the cognitive processes. However, this classical inference suffers from multiple comparisons issues. Also, it does not take into account the multivariate structure of the fMRI data.

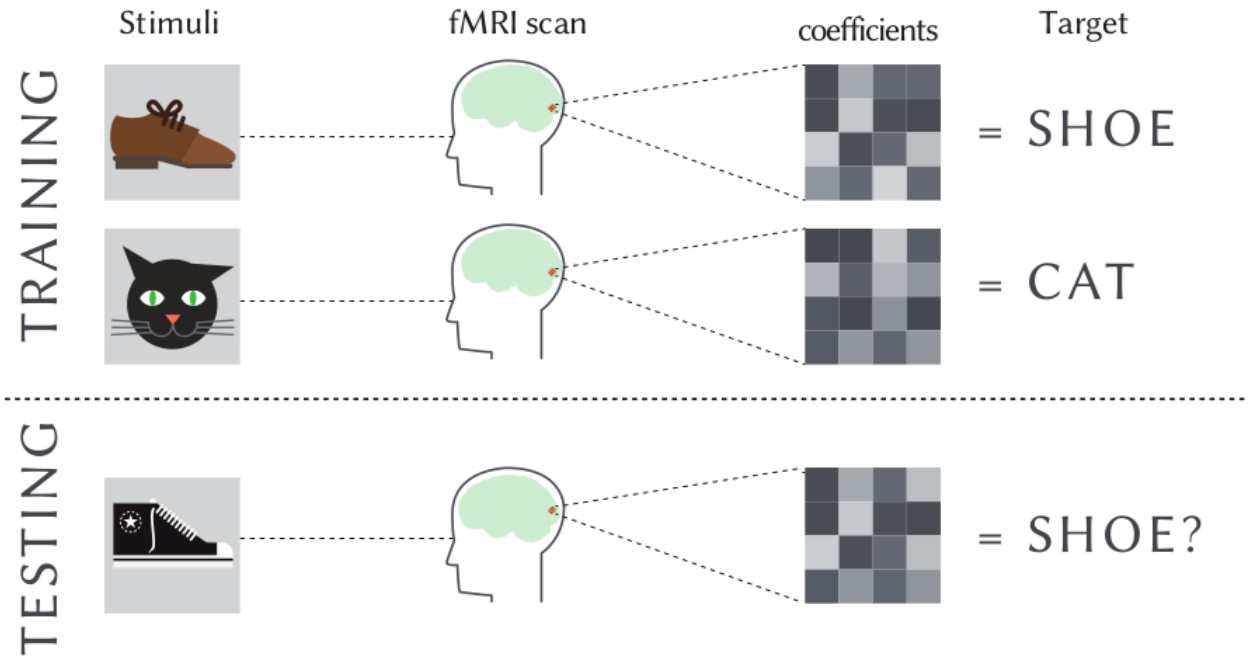


Figure 3.1: **Decoding** models mine patterns of activity to discriminate between cognitive states [Dehaene et al., 1998]. Different activation patterns reflect different mental states. For example, those associated with different images viewed by the subject. In a training phase, the classifier will learn to discriminate between brain activity measured under different cognitive states. In the testing phase the generalization performance of the trained model is quantified by evaluating the classifier on the testing set and comparing the output of the classifier with the true labels associated with the stimuli. The prediction accuracy of the model is used as a measure of the quantity of information about the cognitive task shared by the voxels. Adapted from [Pedregosa-Izquierdo, 2015]

.

Inverse inference (or "brain reading"). An alternative approach called *inverse inference* (or "brain-reading") [Dehaene et al., 1998, Cox and Savoy, 2003], has been proposed in order to cope with the limitations of the aforementioned classical inference. Inverse inference relies on a pattern recognition, and aims at decoding the neural code by using machine-learning methods. Based on a set of brain activation maps, inverse inference builds a predictive model that can be used for predicting a behavioral target (age, sex, IQ, etc.) for a new set of images. The prediction accuracy of the model is used as a measure of the quantity of information about the cognitive task

shared by the voxels. By construction, this approach is multivariate, and can provide more sensitive analysis than standard statistical parametric mapping procedure [Kamitani 05, Haynes 06] Several methods have been tested for classification or regression of activation images (Linear Discriminant Analysis – LDA, Support Vector Machines – SVM, Lasso, Elastic net regression, and many others), but, in this problem, the major bottleneck remains the localization of predictive regions within the brain volume. Additionally, we have to deal with the curse of dimensionality, as the number of features (voxels, regions) is much larger ($\sim 10^5 - 10^6$) than the numbers of sample (images) ($\sim 10^2 - 10^3$), the latter being limited by the cost of acquisition. Thus the prediction method may overfit the training set and thus not generalize well to new samples.

3.2 Sparsity and structure-inducing priors: towards intepretable multi-variate models

To cope with the high dimensionality of the data, the learning method has to be regularized. However, the spatial structure of the image is not taken into account in standard regularization methods, so that the extracted features are often hard to interpret. Sparsity and spatial smoothness inducing priors can be used to perform jointly the prediction of a target variable and region segmentation in multivariate analysis settings. Sparsity can be enforced by penalizing the (sum of) absolute values of the regression coefficients, leading to the so-called Lasso model. Smoothness can achieved in penalizing the spatial gradient of the regression coefficients, to enforce smooth regions ("blobs"). The Total-variation (TV) [Rudin et al., 1992] penalty has proven to be particularly powerful for realizing such effects. Laplacian regularization is an easier means to this end (because in leads to a differentiable problem), but have sub-optimal rates for noisy signal recovery [Sadhanala et al., 2016], and the visual effect is less appealing.

In the context of neuro-imaging, sparsity and smoothness have been compiled to yield regression coefficients which are faithful to known neurobiological organization of the brain, while alleviating the risk of over-fitting due to inherently small-sample settings. Specifically, it has been shown that one can employ priors like TV-ℓ_1 [Baldassarre et al., 2012, Gramfort et al., 2013], TV-ElasticNet [Dubois et al., 2014], and GraphNet [Grosenick et al., 2013] (aka Smooth-Lasso [Hebiri and van de Geer, 2011]) to regularize regression and classification problems in brain imaging. TV has also been employed to enhance the estimation of the voxel-wise *Hurst exponent* [1], as a measure of temporal self-similarity in brain dynamics [Pellé et al., 2016].

[1] $H := \lim_{N\to\infty} \frac{\mathbb{E}(r_N/\sigma_N)}{\log N}$, where r_N is the empirical range (i.e max value minus min value) of the first N values in a time-series, and σ_N is their standard deviation. For example in 1D, white noise has $H = -1/2$.

Notation. We denote by $\mathbf{y} \in \mathbb{R}^n$ the targets to be predicted (age, sex, IQ, etc.); the *design matrix* $\mathbf{X} \in \mathbb{R}^{n\times p}$ are the masked (see Fig. 3.2) brain images related to the presentation of different stimuli, or other brain acquisition (e.g gray-matter concentration maps from anatomy, etc.). The integer p is the number of voxels, and n the number of samples (images). In brain imaging, $n \ll p$; typically, $p \sim 10^3 - 10^6$ (in full-brain analysis), while $n \sim 10 - 10^3$ (n being limited by the cost of acquisition, etc.). ∇_x will denote the discrete spatial gradient operator along the x-axis, ∇_y along the y-axis, etc.

Thus, at a voxel j, the spatial gradient of an image $\mathbf{w}$ is a vector $\nabla\mathbf{w}(j) := [\nabla_x\mathbf{w}(j), \nabla_y\mathbf{w}(j), \nabla_z\mathbf{w}(j)]$, $\forall\mathbf{w} \in \mathbb{R}^p$. This defines a linear operator $\nabla \in \mathbb{R}^{3p\times p}$ (the discrete 3D spatial gradient operator) from $\mathbb{R}^p$ to $\mathbb{R}^{3p}$. For a mixing constant $\rho \in [0, 1]$, $\nabla_\rho \in \mathbb{R}^{4p\times p}$ will denote the identity-augmented version of ∇, defined by $\nabla_\rho\mathbf{w} := [(1-\rho)\nabla\mathbf{w}, \rho\mathbf{w}] \in \mathbb{R}^{4p}$.

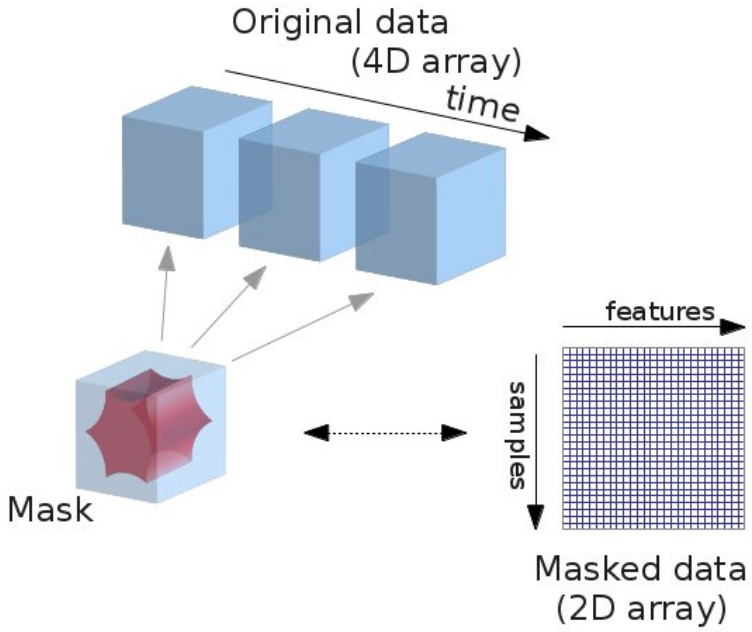


Figure 3.2: **Masking** of volumic brain data (4D = 3D space + 1 **time** or **samples**) to produce a design matrix required in standard machine learning (clustering, classification, regression, etc.). Each 3D volume considered is a sample point. The values of the voxels in this volume that lie in the mask are collected into a **feature vector**. All these vectors are vertically stacked to produce an n-by-p design matrix $\mathbf{X}$, where p is the number of voxels in the mask. The mask can be just the region of the 3D cube occupied by the brain, or a subset of such. In the latter case, this typically corresponds to Region-of-Interests (ROIs) deemed to be interesting for an experiment. The former case is referred to as "full brain", and the mask typically contains up to p = millions of voxels. See [Abraham et al., 2014] for more details.

Prerequisites. Given that this chapter and many others will be quite heavy on proximal calculus and sparse modelling, we would like to suggest the following references as a good starting point for the non-expert reader on these subjects:

- Proximal calculus [Combettes and Wajs, 2005, Beck and Teboulle, 2009, Combettes and Pesquet, 2011].
- Sparse modelling [Mairal et al., 2014, Bach et al., 2012].

That notwithstanding, we shall endeavour to develop the material bottom-up assuming as much as possible only a bare minimum prerequisite knowledge on very specialized topics.

3.3 SpaceNet: sparse structured models for brain data

We now describe the family of structured models which have been proposed for enhanced multivariate analysis in neuro-imaging, namely: Total-Variation (TV) [Michel et al., 2011], TV-ℓ_1 [Baldassarre et al., 2012, Gramfort et al., 2013], GraphNet [Grosenick et al., 2013, Hebiri and van de Geer, 2011], TV-ElasticNet [Dubois et al., 2014], Sparse Variation [Eickenberg

et al., 2015], and social-sparsity [Kowalski et al., 2013, Varoquaux et al., 2016]. These can all be synthesized into a general framework, referred to as *SpaceNet* [Dohmatob et al., 2015b], as follows

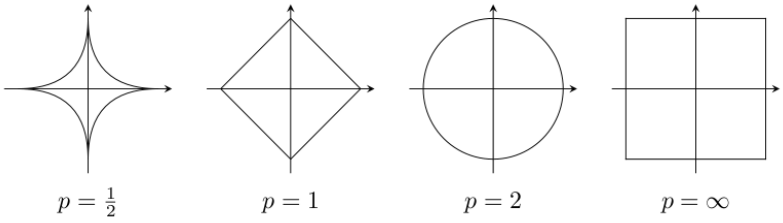


Figure 3.3: ℓ_p unit ball for various values of p. The kinks in the cases $0 \leq p \leq 1$ impose sparsity. One notes however that, the cases $0 \leq p < 1$ lead to non-convex intractable optimization problems.

$$\underset{\mathbf{w}\in\mathbb{R}^p}{\text{minimize}}\ E(\mathbf{w}) := \ell(\mathbf{y}, \mathbf{X}\mathbf{w}) + \alpha\mathcal{P}(\mathbf{w}). \tag{3.1}$$

The coefficients $\mathbf{w}$ define a spatial map in over the brain (one value per voxel). The term $\ell(\mathbf{y}, \mathbf{X}\mathbf{w})$ is the loss / data-fit term. Popular choices include:

$$\ell(\mathbf{y}, \mathbf{X}\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} \begin{cases} \frac{1}{2}(\mathbf{X}_i^T\mathbf{w} - y_i)^2, & \text{for least-squares regression} \\ \log(1 + \exp(-y_i\mathbf{X}_i^T\mathbf{w})), & \text{for logistic regression,} \\ (1 - y_i\mathbf{X}_i^T\mathbf{w})_+, & \text{for hinge loss (used in SVMs)} \\ \vdots \end{cases}$$

In the above general model, $\mathcal{P}(\mathbf{w})$ is the penalty term, which simultaneously imposes both sparsity and structure (blobs). The different spatial regularization methods that have appeared in neuro-imaging literature can be cast into this correspond to different choices of the convex penalty $\mathcal{P}$ acting on the extended gradient of the coefficients $\mathbf{w}$. Viz,

$$\mathcal{P}(\mathbf{w}) = \begin{cases} \rho\|\mathbf{w}\|_1 + \frac{1}{2}(1-\rho)\|\nabla\mathbf{w}\|_{\text{Fro}}^2 = \sum_{j\in[\![p]\!]} \rho|w_j| + \frac{1}{2}(1-\rho)\|(\nabla\mathbf{w})_j\|_2^2, & \text{for GraphNet ,} \\ \|\nabla_\rho\mathbf{w}\|_{1+2,1} = \rho\|\mathbf{w}\|_1 + \|\nabla\mathbf{w}\|_{2,1} = \sum_{j\in[\![p]\!]} \rho|w_j| + (1-\rho)\|(\nabla\mathbf{w})_j\|_2, & \text{for isotropic TV-}\ell_1\text{ ,} \\ \|\nabla_\rho\mathbf{w}\|_{1,1} = \rho\|\mathbf{w}\|_1 + (1-\rho)\|\nabla\mathbf{w}\|_{1,1} = \sum_{j\in[\![p]\!]} \rho|w_j| + (1-\rho)\|(\nabla\mathbf{w})_j\|_1, & \text{for anisotropic TV-}\ell_1\text{ ,} \\ \|\nabla_\rho\mathbf{w}\|_{2,1} = \sum_{j\in[\![p]\!]} \|(\nabla_\rho\mathbf{w})_j\|_2, & \text{for Sparse Variation ,} \\ \vdots \end{cases} \tag{3.2}$$

where

- $\alpha > 0$ is a regularization parameter controls the total amount of regularization;
- ρ ($0 < \rho \leq 1$) is a mixing constant between the sparsity-inducing ℓ_1 part and the cluster-promoting part of the penalty term. The particular case $\rho = 1$ corresponds to the usual Lasso. Vanilla TV [Michel et al., 2011] corresponds to TV-ℓ_1 with $\rho = 0$.
- The matrix ∇_ρ is the extended discrete gradient operator defined in Table 1.

Bayesian interpretation of SpaceNet models. The penalties $\mathcal{P}(\mathbf{w})$ in (3.1) admit a Bayesian interpretation as a prior on the distribution of the coefficients $\mathbf{w}$

$$p_{\alpha,\rho}(\mathbf{w}) \propto \exp(-\mathcal{P}(\mathbf{w})). \tag{3.3}$$

For example, the GraphNet [Grosenick et al., 2013, Hebiri and van de Geer, 2011], the penalties $\mathcal{P}(\mathbf{w})$ penalty corresponds to

$$p_{\alpha,\rho}(\mathbf{w}) \propto \prod_{j=1}^{p} \exp(-\alpha\rho|w_j|) \prod_{j=1}^{p} \exp\left(-\alpha(1-\rho) \sum_{l\sim\text{neigh}(j)} w_j \Delta_{j,l} w_l\right). \tag{3.4}$$

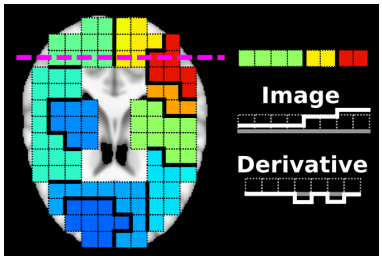


Figure 3.4: A cartoon showing a sparse and blobby (step-wise constant / cartoon-like) brain map, as would be sought for by Total-Variation regularization (9.2).

SpaceNet models (3.1) result in brain maps which are both sparse (i.e regression coefficients $\mathbf{w}$ are zero everywhere, except at predictive voxels) and structured (blobby). See Fig. 3.4. The superiority of such methods over methods without structured priors like the Lasso, ANOVA, Ridge, SVM, etc. for yielding more intepretable maps and improved prediction scores is now well established. See for example [Baldassarre et al., 2012, Gramfort et al., 2013]. These priors are fast becoming popular for brain decoding and segmentation. Indeed, they leverage a feature-selection function (since they limit the number of active voxels), and also a structuring function (since they penalize local differences in the values of the brain map). For example, see Fig. 3.6. Also, such priors produce state-of-the-art methods for automatic extraction of functional brain atlases [Abraham et al., 2013].

Submodular interpretation of TV. We note that anisotropic TV penalty (3.2) on an arbitrary (undirected) graph $G = (V, E)$ is the *Lovasz extension* of the *cut-function* $F : 2^V \to \mathbb{N}$, $S \mapsto$ "number of edges between S and $V \setminus S$", defined by $F^L(x) := \mathbb{E}_{\lambda\sim\mathcal{U}([0,1])}[F(\{v \in V | x_v \geq \lambda\})]$, for all $x \in [0,1]^{\#V}$. Thus anisotropic TV minimization can be seen as a graph-cut problem for which efficient algorithms exist [Bach, 2013, Landrieu and Obozinski, 2016].

3.4 Methods

The SpaceNet model leads to difficult non-smooth mathematical optimization problems making their implementation and practical usability challenging. [Dohmatob et al., 2014] benchmarked a rich variety of cutting-edge solvers for such problems, and gave crucial recommendations on how to effectively implement these algorithms in practice. In these benchmarks, the FISTA algorithm emerged as the go-to algorithm for the TV-L1 problem [Dohmatob et al., 2015a]. These hints have been carefully used in implementing SpaceNet. Also as a preprocessing step, we use univariate feature-screening (ANOVA) to eliminate voxels which are irrelevant to the learning problem, thus reducing the size of the problem. As a result the implementation of SpaceNet is fast, robust, and automatically sets its hyper-parameters

(internal cross-validation). All these technical details will be properly presented in the next few chapters.

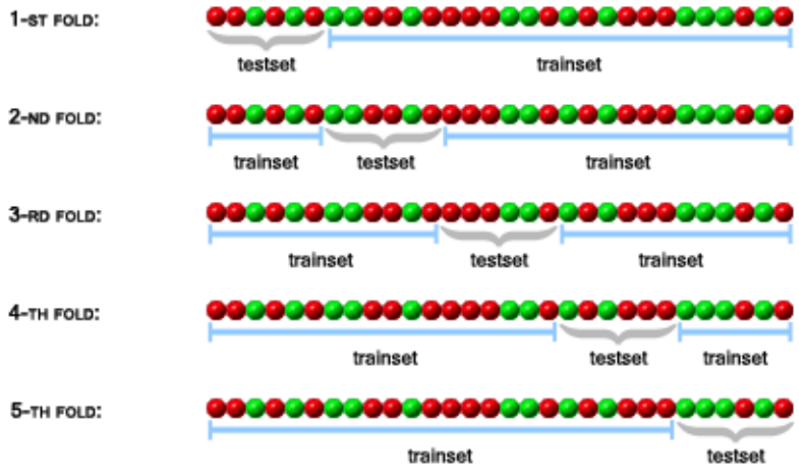
3.4.1 Cross-validation

Cross-validation (e.g see [Stone, 1974]) is a technique used to protect against overfitting in a predictive model, particularly in a case where the amount of data may be limited. In cross-validation, you make a fixed number of folds (or partitions) of the data, run the analysis on each fold, and then average the overall error estimate. This gives an (asymptotically) unbiased estimate of the true generalization error of the model. Two major types of cross-validation are K-Fold and Leave-One-Out (LOO).

K-Fold cross-validation. One iteration of the K-fold cross-validation is performed in the following way: First, a random permutation of the sample set is generated and partitioned into K subsets ("folds") of about equal size. Of the K subsets, a single subset is retained as the validation data for testing the model (this subset is called the "testset"), and the remaining K - 1 subsets together are used as training data ("trainset"). Then a model is trained on the trainset and its accuracy is evaluated on the testset. Model training and evaluation is repeated K times, with each of the K subsets used exactly once as the testset. The case of a 5-fold cross-validation with 30 samples is illustrated in Fig. 8.4.

Leave-One-Out cross-validation. As the name suggests, leave-one-out cross-validation involves using a single sample from the original sample set as the validation data, and the remaining samples as the training data. This is repeated such that each sample in the sample set is used exactly once as the validation data. This is the same as K-fold cross-validation where K is equal to the number of samples in the sample set. In LOO, there is no need in generating random permutations and in repeating it, because the training and validation datasets for each of the folds are always the same, and therefore the result of the accuracy estimation is determined.

Model-selection via cross-validation. One can instrument cross-validation to tune the hyper-parameters of a model like SpaceNet (7.1), by selected the configuration of model parameters with least cross-validation error. The number of models fitted is proportional to the size of the parameter grid –i.e exponential in the number of parameters to tune– and therefor can become prohibitive in case there are many free hyper-parameters in the model. Also, since some of the data has to be set aside for validation, cross-validation in very small sample settings (e.g e few tenths, as is the in some neuroimaging experiments) may be troublesome as the error estimates then have very high variance. A reasonable alternative in such situations are SURE (short for *Stein's Unbiased Risk Estimator* [Stein, 1981])-based methods, which are applicable whenever a procedure for obtaining (an unbiased estimate of) the number of degrees of freedom of the model is available. This is the case with the models like the ElasticNet and GraphNet [Hebiri and van de Geer, 2011]. Recently, [Deledalle et al., 2014] has proposed a SURE-like technique for structured models with many hyper-parameters.



(a) 1 iteration of 5-Fold cross-validation

(α_1, ρ_1)	(α_1, ρ_2)	$\dots$	(α_1, ρ_M)	
(α_2, ρ_1)	(α_2, ρ_2)	$\dots$	(α_2, ρ_M)	
(α_3, ρ_1)	(α_3, ρ_2)	$\dots$	(α_3, ρ_M)	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
(α_L, ρ_1)	(α_L, ρ_2)	$\dots$	(α_L, ρ_M)	

(b) Parameter grid

Figure 3.5: **Model-selection** via cross-validation. ***(a)*** K-Fold cross-validation, illustrated here for the case $K = 5$, involves taking the available data and partitioning it into K groups. Then $K - 1$ groups are used to train a set of models that are then evaluated on the remaining group. Adapted from `http://genome.tugraz.at/proclassify/help/pages/XV.html`. ***(b)*** $L \times M$ grid over which to search for optimal configuration in a model with two hyper-parameters α and ρ. For a model like SpaceNet (7.1), the grid is is constrained to verify $0 \leq \alpha_L < \dots < \alpha_1 = \alpha_{\max}$ and $0 \leq \rho_M < \dots < \rho_1 \leq 1$, with $L = 10$ and $M = 3$ typically, given a total of $LM = 30$ models to compare. In chapter 5, we show how *early-stopping* and other heuristics can be used to make the total cost much more effective than just fitting LM models in a CV loop.

3.4.2 How SpaceNet compares against classical unstructured models

Classification. We compared SpaceNet (TV-L1 and GraphNet / Smooth-Lasso priors) with an SVM (Support Vector Machine) on the visual-recognition dataset [Haxby et al., 2001]. This dataset consists of 6 subjects with 12 runs per subject. In each run, the subjects passively viewed images of eight object categories, grouped in 24-second blocks separated by intermittent rest periods. This experiment is a classification task: predicting the object category. The design matrix is made of time-series from the full-brain mask of $p = 23,707$ voxels over 216 TRs (Repetition Times), of a single subject (subj1). 126 TRs were used for training all the models, whilst testing was done on 90 left-out TRs. The results are depicted in Figures 3.6 and 3.7.

Regression. In [Gramfort et al., 2013], the authors compared several models on a dataset in which subjects were presented with mixed (gain/loss) gambles, and decided whether they would accept each gamble [Jimura and Poldrack, 2012]. No outcomes of these gambles were presented during scanning, but after the scan three gambles were selected at random and played for real money. The prediction task here is to predict the magnitude of the gain and thus a regression on a continuous variable. The full dataset of 16 subjects with 48 3D scans each, making up for a total of $n = 768$ samples with approximately $p = 3.3 \times 10^4$ voxels. The prediction here is inter-subject: the estimator learns on some subjects and predicts on left out

subjects. The results are shown in Fig. 3.8.

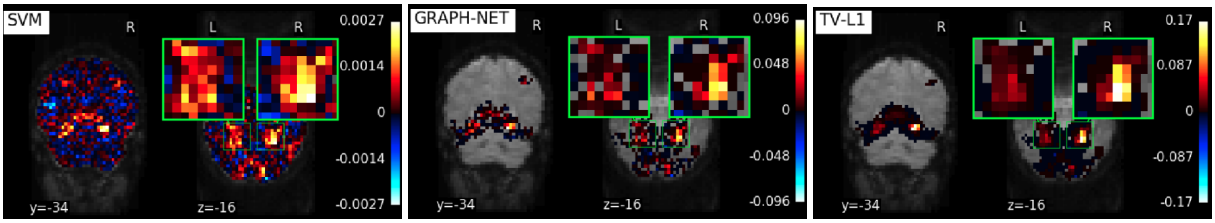


Figure 3.6: The figure shows results of comparing the SpaceNet models TV-ℓ_1 and Graph-Net against an SVM (Support Vector Machine) classifier on the visual-recognition dataset [Haxby et al., 2001] As can be seen from the figure, SpaceNet priors (TV-ℓ_1, GraphNet/Smooth-Lasso, etc.) yield stable and more intepretable maps by enforcing smoothness on the coefficients while segmenting predictive regions (blobs) from noisy background.

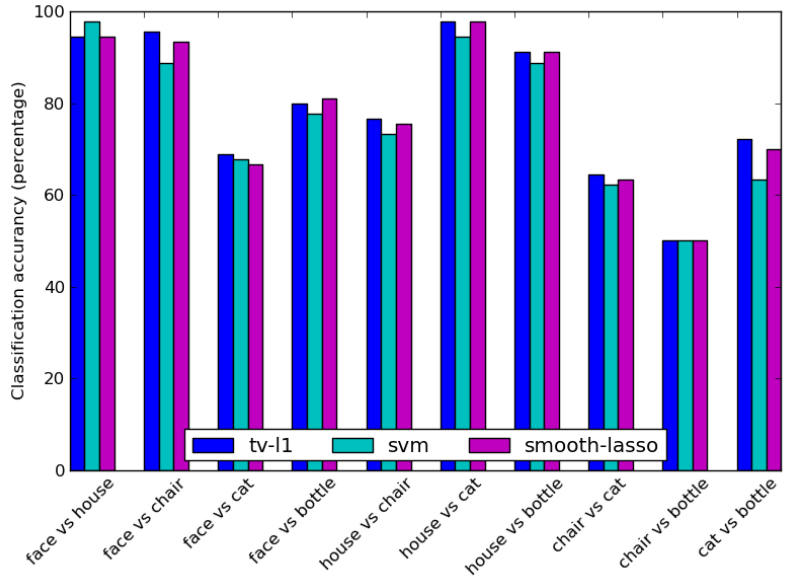


Figure 3.7: Bar chart showing percentage classification on left-out, for one-vrs-one classification on the visual recognition dataset [Haxby et al., 2001]. We see that the highly structured maps produced by SpaceNet models (3.1) (e.g see Fig. 3.6) are not at the expense of model accuracy.

3.5 Conclusion

We have presented SpaceNet, a family of priors for brain decoding that enforce both sparsity and structure, leading to better prediction scores and intepretable brain maps. We believe that such priors will become commonplace in future. In the next few chapters, we open the "black-box" and develop from ground-up, the details of such models, including their practical implementation on a computer.

Bibliography

Alexandre Abraham, Elvis Dohmatob, Bertrand Thirion, Dimitris Samaras, and Gael Varoquaux. Extracting brain regions from rest fMRI with Total-Variation constrained dictionary learning. In *MICCAI*. 2013.

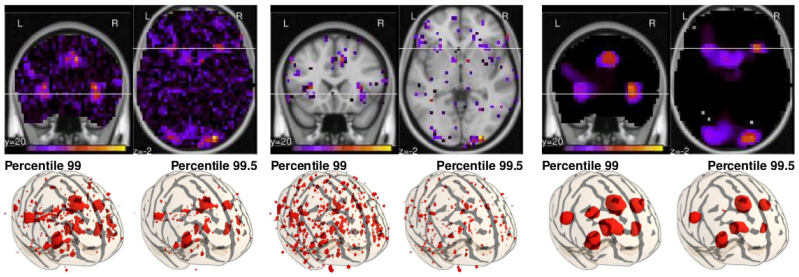


Figure 3.8: Results on fMRI data from [4] (from left to right F-test, ElasticNet and TV-L1). The TV-L1 regularized model segments neuroscientifically meaningful predictive regions in agreement with univariate statistics while the ElasticNet yields sparse although very scattered non-zero weights. Source: adapted from [Gramfort et al., 2013].

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. *Frontiers in neuroinformatics*, 8, 2014.

Francis Bach, Rodolphe Jenatton, Julien Mairal, Guillaume Obozinski, et al. Optimization with sparsity-inducing penalties. *Foundations and Trends® in Machine Learning*, 4(1):1–106, 2012.

Francis R. Bach. Learning with Submodular Functions: A Convex Optimization Perspective. *Foundations and Trends in Machine Learning*, 6(2-3): 145–373, 2013.

Luca Baldassarre, Janaina Mourao-Miranda, and Massimiliano Pontil. Structured sparsity models for brain decoding from fMRI data. In *PRNI*, 2012.

A. Beck and M. Teboulle. A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems. *SIAM Journal on Imaging Sciences*, 2, 2009.

Patrick L Combettes and Jean-Christophe Pesquet. Proximal splitting methods in signal processing. In *Fixed-point algorithms for inverse problems in science and engineering*. Springer New York, 2011.

Patrick L Combettes and Valérie R Wajs. Signal recovery by proximal forward-backward splitting. *Multiscale Modeling & Simulation*, 4, 2005.

D. D. Cox and R. L. Savoy. Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex. *Neuroimage*, 19, 2003.

Stanislas Dehaene, Gurvan Le Clec'H, Laurent Cohen, Jean-Baptiste Poline, Pierre-François van de Moortele, and Denis Le Bihan. Inferring behavior from functional brain images. *Nature neuroscience*, 1, 1998.

Charles-Alban Deledalle, Samuel Vaiter, Jalal Fadili, and Gabriel Peyré. Stein Unbiased GrAdient estimator of the Risk (SUGAR) for multiple parameter selection. *SIAM Journal on Imaging Sciences*, 7(4):2448–2487, 2014.

Elvis Dohmatob, Alexandre Gramfort, Bertrand Thirion, and Gaël Varoquaux. Benchmarking solvers for TV-l1 least-squares and logistic regression in brain imaging. In *PRNI*. IEEE, 2014.

Elvis Dohmatob, Michael Eickenberg, Bertrand Thirion, and Gaël Varoquaux. Speeding-up model-selection in GraphNet via early-stopping and univariate feature-screening. In *Pattern Recognition in NeuroImaging (PRNI), 2015 International Workshop on*. IEEE, 2015a.

Elvis Dohmatob, Michael Eickenberg, Bertrand Thirion, and Gael Varoquaux. SpaceNet: Multivariate brain decoding and segmentation. In *OHBM*, 2015b.

Mathieu Dubois, Fouad Hadj-Selem, Tommy Lofstedt, Matthieu Perrot, Clara Fischer, Vincent Frouin, and Edouard Duchesnay. Predictive support recovery with TV-Elastic Net penalty and logistic regression: an application to structural MRI. In *PRNI*. IEEE, 2014.

Michael Eickenberg, Elvis Dohmatob, Bertrand Thirion, and Gaël Varoquaux. Total Variation meets Sparsity: statistical learning with segmenting penalties. In *MICCAI*. 2015.

Karl J Friston, Andrew P Holmes, Keith J Worsley, J-P Poline, Chris D Frith, and Richard SJ Frackowiak. Statistical parametric maps in functional imaging: a general linear approach. *Human brain mapping*, 2(4):189–210, 1994.

Alexandre Gramfort, Bertrand Thirion, and Gaël Varoquaux. Identifying predictive regions from fMRI with TV-L1 prior. In *PRNI*, 2013.

Logan Grosenick, Brad Klingenberg, Kiefer Katovich, Brian Knutson, and Jonathan E Taylor. Interpretable whole-brain prediction analysis with GraphNet. *NeuroImage*, 72, 2013.

James V. Haxby, Ida M. Gobbini, Maura L. Furey, Alumit Ishai, Jennifer L. Schouten, and Pietro Pietrini. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293, 2001.

M. Hebiri and S. van de Geer. The Smooth-Lasso and other $\ell_1 + \ell_2$-penalized methods. *Electron. J. Stat.*, 5, 2011.

Koji Jimura and Russell A Poldrack. Analyses of regional-average activation and multivoxel pattern information tell complementary stories. *Neuropsychologia*, 50, 2012.

Matthieu Kowalski, Kai Siedenburg, and Monika Dörfler. Social Sparsity! Neighborhood Systems Enrich Structured Shrinkage Operators. *Signal Processing, IEEE Transactions on*, 61(10), 2013.

Loïc Landrieu and Guillaume Obozinski. Cut Pursuit: Fast Algorithms to Learn Piecewise Constant Functions. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, AISTATS 2016, Cadiz, Spain, May 9-11, 2016*, pages 1384–1393, 2016.

Julien Mairal, Francis R. Bach, and Jean Ponce. Sparse Modeling for Image and Vision Processing. *Foundations and Trends in Computer Graphics and Vision*, 8(2-3), 2014.

V. Michel, A. Gramfort, G. Varoquaux, E. Eger, and B. Thirion. Total Variation Regularization for fMRI-Based Prediction of Behavior. *Medical Imaging, IEEE Transactions on*, 30, 2011.

S. Ogawa, T. M. Lee, A. R. Kay, and D. W. Tank. Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proc Natl Acad Sci U S A*, 87(24):9868–9872, 1990a.

Seiji Ogawa, Tso-Ming Lee, Asha S. Nayak, and Paul Glynn. Oxygenation-sensitive contrast in magnetic resonance image of rodent brain at high magnetic fields. *Magnetic Resonance in Medicine*, 14(1):68–78, 1990b.

Fabian Pedregosa-Izquierdo. *Feature extraction and supervised learning on fMRI: from practice to theory*. PhD thesis, Université Pierre et Marie Curie-Paris VI, 2015.

Hubert Pellé, Philippe Ciuciu, Mehdi Rahim, Elvis Dohmatob, Patrice Abry, and Virginie Van Wassenhove. Multivariate Hurst exponent estimation in fMRI. Application to brain decoding of perceptual learning. In *13th IEEE International Symposium on Biomedical Imaging*, 2016.

Leonid I. Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. 60(1):259–268, 1992.

Veeranjaneyulu Sadhanala, Yu-Xiang Wang, and Ryan J Tibshirani. Total Variation Classes Beyond 1d: Minimax Rates, and the Limitations of Linear Smoothers. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 3513–3521. Curran Associates, Inc., 2016.

Charles M. Stein. Estimation of the mean of a multivariate normal distribution. *Ann. Statist.*, 9(6):1135–1151, 11 1981.

M. Stone. Cross-Validatory Choice and Assessment of Statistical Predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36 (2):111–147, 1974.

Gaël Varoquaux, Matthieu Kowalski, and Bertrand Thirion. Social-sparsity brain decoders: faster spatial sparsity. In *PRNI conference*, 2016.

Efficient optimization of sparsity and smoothness regularized models

Contents

Though the SpaceNet models introduced in equations (3.1) of chapter 3 lead to superior estimators compared to classical estimators (Ridge regression, SVM, etc.) without spatial penalization, they are considerably harder to optimize than these classical models. Indeed, the corresponding optimization problems is non-separable in the model coefficients, and except for the case of GraphNet [Hebiri and van de Geer, 2011, Grosenick et al., 2013] and social-sparsity [Kowalski et al., 2013, Varoquaux et al., 2016], the penalty term $\mathcal{P}(\mathbf{w})$ is neither smooth nor proximable[1]. For the penalty to fully exercise its structuring effect on the maps, this optimization problem must be solved to a good tolerance resulting in a computational challenge. Lack of good solver and explicit control of tolerance can lead to brain maps and conclusions that reflect properties of the solver more than of model coefficients, as illustrated in Fig. 4.1.

[1] A function f is said to be *proximable* if its operator $\text{prox}_{\gamma f}$ is easy to compute. This is the case for ℓ_p-norms (with $p \geq 1$, to ensure convexity) and indicator functions of simple closed convex sets like balls, simplexes, half-spaces, etc.

4.1 Solving TV-L1 regularized problems

The optimization problem (3.1) is very challenging: it is non-smooth (except in the case of Laplacian regularization), non-separable and heavily ill-conditioned. For the penalty to fully exercise its structuring effect on the maps, this optimization problem must be solved to a good tolerance resulting in a computational challenge. In [Dohmatob et al., 2014], we did an extensive study of all solvers applicable to the problem in TV-ℓ_1 special case (which happens to be the most difficult scenario). Our results outlined the best strategy: a double FISTA loop, where the inner loop computes the proximal operator of the penalty term, with approximate precision on the

duality-gap. This was further refined and implemented in [Varoquaux et al., 2015].

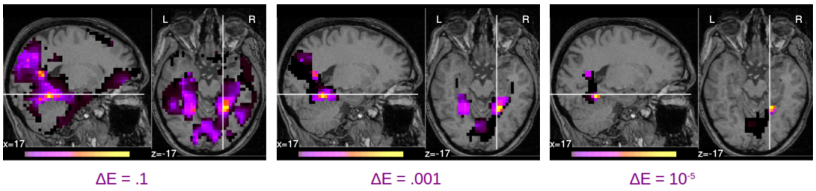


Figure 4.1: TV-ℓ_1 maps for the face-house discrimination task on the visual recognition dataset. Note that the stopping criterion is defined as a threshold on the energy decrease per one iteration of the algorithm, and thus differs from the tolerance displayed in figure 4.1. This figure shows the importance of convergence for problem (3.1), and motivates the need for fast solvers for SpaceNet priors, especially the non-smooth ones like TV-ℓ_1 and Sparse Variation. See [Dohmatob et al., 2014] for details.

4.1.1 The algorithms

ISTA/FISTA. ISTA [Daubechies et al., 2004], and its accelerated variant FISTA [Beck and Teboulle, 2009a], are proximal gradient approaches: the go-to methods for non-smooth optimization. In their seminal introduction of TV for fMRI, [Michel et al., 2011] relied on ISTA. The challenge of these methods for TV is that the proximal operator of TV cannot be computed exactly; we approximate it in an inner FISTA loop [Beck and Teboulle, 2009b, Michel et al., 2011]. Here, for all FISTA implementations we use the faster monotonous FISTA variant [Beck and Teboulle, 2009b]. We control the optimality of the TV proximal via its dual gap [Michel et al., 2011] and use a line-search strategy in the monotonous FISTA to decrease the tolerance as the algorithm progresses, ensuring convergence of the TV-ℓ_1 regression with good accuracy. See [Dohmatob et al., 2014, Varoquaux et al., 2015].

ISTA/FISTA with backtracking. A key ingredient in FISTA's convergence is the Lipschitz constant $L_{\nabla\ell}$, of the derivative of smooth part of the objective function . The tighter the upper bound used for this constant, the faster the resulting FISTA algorithm. In FISTA, the main use of $L_{\nabla\ell}$ is the fact that: for any stepsize $0 < t \leq 1/L_{\nabla\ell}$ and for any point $\mathbf{z}$,

$$\begin{aligned} \ell(\mathbf{p}_t(\mathbf{z})) \leq \ell(\mathbf{z}) + \mathbf{r}_t^T \nabla\ell(\mathbf{z}) + \frac{1}{2t}\|\mathbf{r}_t\|_2^2, \text{ where} \\ \mathbf{p}_t(\mathbf{z}) := \text{prox}_{\alpha t \mathcal{P}}(\mathbf{z} - t\nabla\ell(\mathbf{z})) \text{ and } \mathbf{r}_t := \mathbf{p}_t(\mathbf{z}) - \mathbf{z} \end{aligned} \tag{4.1}$$

In least-squares regression, $L_{\nabla\ell}$ is precisely the largest singular value of the design matrix $\mathbf{X}$. For logistic regression however, the tightest known upper bound for $L_{\nabla\ell}$ is $\|\mathbf{X}\|\|\mathbf{X}^T\|$, which performs very poorly locally (i.e, stepsizes $\sim 1/L_{\nabla\ell}$ are sub-optimal locally). A way to circumvent this difficulty is *backtracking line search* [Beck and Teboulle, 2009a], where one tunes the stepsize t to satisfy inequality (4.1) locally at point $\mathbf{z}$.

ADMM: Alternating Direction Method of Multipliers. ADMM is a Bregman Operator Splitting primal-dual method for solving convex-optimization problems by splitting the objective function in two convex terms which are functions of linearly-related auxiliary variables [Boyd et al., 2010]. ADMM is particularly appealing for problems such as TV regression: using the variable split $\mathbf{z} \leftarrow \nabla\mathbf{w}$, the regularization is a simple ℓ_1/ℓ_2 norm on $\mathbf{z}$ for which the proximal is exact and computationally cheap. However, in our settings, limitations of ADMM are:

- The $\mathbf{w}$-update involves the inversion of a large p-by-p ill-conditioned linear operator (precisely a weighted sum of $\mathbf{X}^T\mathbf{X}$, the laplacian Δ, and the identity operator).
- The dual stepsize parameter ν in the penalization of the split residual $\frac{1}{2}\nu\|\mathbf{z} - \nabla\mathbf{w}\|_2^2$ is hard to set (this is still an open problem), and though under mild conditions ADMM converges for any value of ν, the convergence rate depends on ν. In chapter 7, we study the rate of convergence of ADMM on the kinds of penalized least squares regression problem considered in this manuscript, and derive some theoretical results.

Primal-Dual algorithm of Chambolle and Pock [Chambolle and Pock, 2011]. This scheme is another method based on operator splitting. Used for fMRI TV regression by [Gramfort et al., 2013], it does not require setting a hyperparameter. However it is a first-order single-step method and is thus more impacted by the conditioning of the problem. Note that here we explore this primal-dual method only in the squared loss setting, in which the algorithm can be accelerated by precomputing the SVD of $\mathbf{X}$ [Gramfort et al., 2013] .

HANSO [Lewis and Overton, 2008]. a modified LBFGS scheme based on gradient sampling methods [Burke et al., 2005] and inexact line-search. For non-smooth problems as in our case, the algorithm relies on random initialization, to avoid singularities with high probability. Here, we used the original authors' implementation.

Uniform approximation by smooth convex surrogates. The ℓ_1 norm (resp. TV semi-norm) is differentiable everywhere with gradient $\left(w_j/|w_j|\right)_{j\in[\![p]\!]}$ (resp. $-\mathrm{div}(((\nabla\mathbf{w})_j/\|(\nabla\mathbf{w})_j\|_2)_{j\in[\![p]\!]})$), except when some voxels are inactive with $w_j = 0$ (resp. $(\nabla\mathbf{w})_j = 0$), corresponding to black spots (resp. edges). A convenient approach (see for example [Bobin et al., 2011, Nesterov, 2005a,b, Beck and Teboulle, 2012]) for dealing with such singularities is to uniformly approximate the offending function with smooth surrogates that preserve its convexity. Given a smoothing parameter $\mu > 0$, we define *smoothed* versions of ℓ_1 and TV:

$$\|\mathbf{w}\|_{1,\mu} := \sum_j \sqrt{\mathbf{w}_j^2 + \mu^2},\ \|\mathbf{w}\|_{\mathrm{TV},\mu} := \sum_j \sqrt{\|(\nabla\mathbf{w})_j\|_2^2 + \mu^2} \tag{4.2}$$

These surrogate upper-bounds are convex and everywhere-differentiable with gradients that are Lipschitz-continuous with constants $1/\mu$ and $\|\nabla\|^2(1/\mu) = 12/\mu$ respectively. They lead to smoothed versions of problem (3.1):

$$\hat{\mathbf{w}}_\mu := \arg\min_{\mathbf{w}} \{E_\mu(\mathbf{w}) := \ell(\mathbf{w}) + \alpha\mathcal{P}_{\text{TV-L1},\mu}(\mathbf{w})\}, \tag{4.3}$$

where $\mathcal{P}_{\text{TV-L1},\mu}(\mathbf{w}) := \rho\|\mathbf{w}\|_{1,\mu} + (1-\rho)\|\mathbf{w}\|_{\mathrm{TV},\mu}$.

To solve (3.1), we consider problems of the form (4.3) with $\mu \to 0^+$: we start with a coarse μ ($= 10^{-2}$, *e.g*) and cheaply solve the μ-smoothed problem (4.3) to a precision $\sim \mu$ using a fast iterative oracle like the LBFGS [Zhu et al., 1994]; we obtain a better estimate for the solution; then we decrease μ by a fixed factor, and restart the solver on problem (4.3) with this solution; and

so on, in a *continuation* process [Bobin et al., 2011] detailed in Alg. 1. This algorithm is not faster than $O(1/\epsilon)$: indeed a first-order algorithm for the sub-problem (4.3) has optimal worst-case iteration complexity $O(\sqrt{L_\mu/\epsilon})$ [Nesterov, 1983], and $L_\mu \sim 1/\mu \sim 1/\epsilon$. We believe that this bound is tight but a detailed analysis is beyond the scope of this manuscript.

Algorithm 1: LBFGS algorithm with continuation

Require: $\epsilon > 0$ the desired precision, β ($0 < \beta < 1$) the rate of decay of the smoothing parameter μ, and $\gamma > 0$ be a constant. Finally, let LBFGS: $(E_\mu, \mathbf{w}^{(0)}, \epsilon) \mapsto \mathbf{w}$ be an oracle which when warm-started with an initial guess $\mathbf{w}^{(0)}$, returns an ϵ-optimal solution (i.e $E_\mu(\mathbf{w}) - E^*_\mu < \epsilon$) for problem (4.3).

1: **Initialize** $0 < \mu^{(0)}$ ($= 10^{-2}$, *e.g*), $\mathbf{w}^{(0)} \in \mathbb{R}^p$, and $k = 0$.
2: **while** $\gamma\mu^{(k)} \geq \epsilon$ **do**
3: $\quad \mathbf{w}^{(k+1)} \leftarrow \text{LBFGS}(E_{\mu^{(k)}}, \mathbf{w}^{(k)}, \gamma\mu^{(k)})$
4: $\quad \mu^{(k+1)} \leftarrow \beta\mu^{(k)}$
5: $\quad k \leftarrow k + 1$
6: **end while**

4.1.2 Experiments on fMRI datasets

We now detail experiments done on publicly available data. All experiments were run full-brain without spatial smoothing.

Visual recognition. Our first benchmark dataset is a popular block-design fMRI dataset from a study on face and object representation in human ventral temporal cortex [Haxby et al., 2001]. It consists of 6 subjects with 12 runs per subject. In each run, the subjects passively viewed images of eight object categories, grouped in 24-second blocks separated by intermittent rest periods. This experiment is a classification task: predicting the object category. We use a two-class prediction target: $\mathbf{y}$ encodes faces versus houses. The design matrix $\mathbf{X}$ is made of time-series from the full-brain mask of $p = 23\,707$ voxels over $n = 216$ TRs, of a single subject (subj1).

Mixed Gambles. Our second benchmark dataset is a study in which subjects were presented with mixed (gain/loss) gambles, and decided whether they would accept each gamble [Tom et al., 2007]. No outcomes of these gambles were presented during scanning, but after the scan three gambles were elected at random and played for real money. The prediction task here is to predict the magnitude of the gain and thus a regression on a continuous variable [Jimura and Poldrack, 2012]. The data The are pulled from 16 subjects with 48 3D scans each, making up for a total of $n = 768$ samples with approximately $p = 3.3 \times 10^4$ voxels.

We study the convergence of the algorithms for parameters close to the optimal parameters set by 10-fold cross-validation to maximize prediction accuracy.

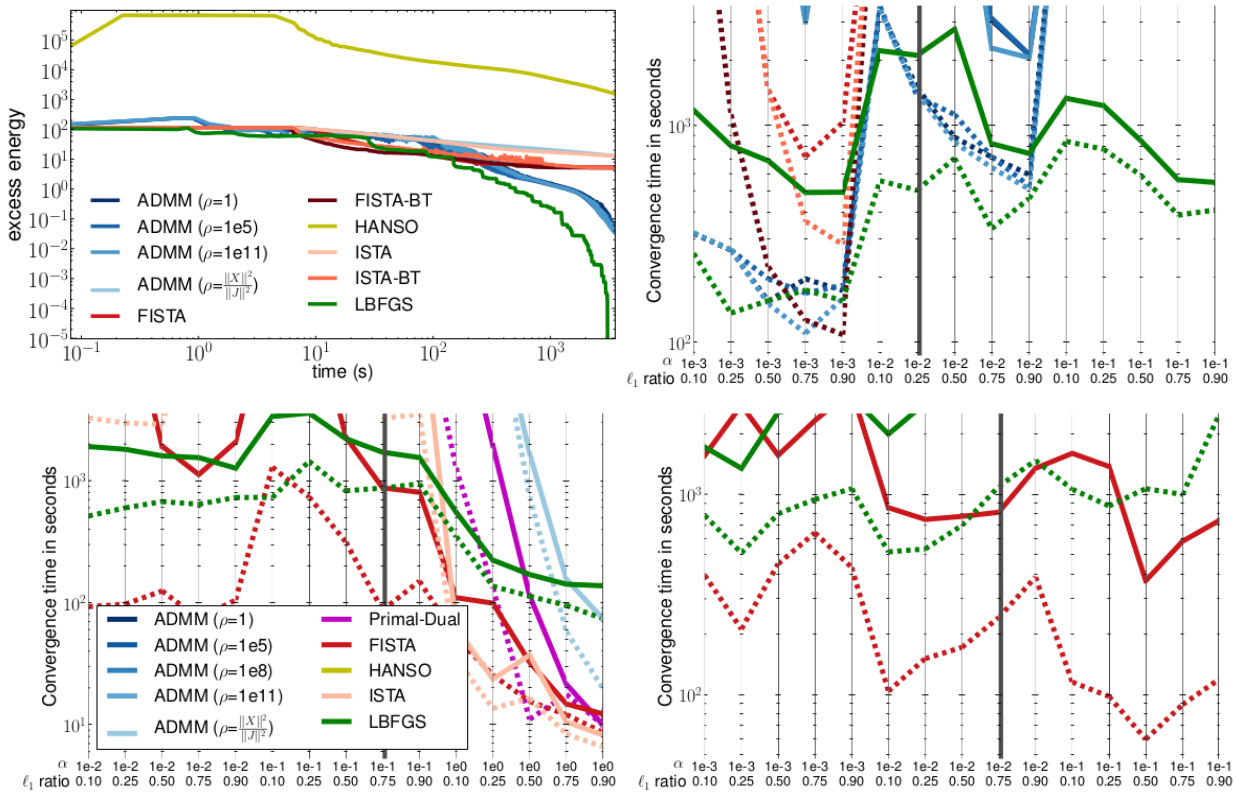


Figure 4.2: **Benchmarking** solvers for TV-ℓ_1 penalized models. **Top:** TV-ℓ_1 penalized Logistic Regression on the visual recognition face-house discrimination task. **Top Left:** excess energy $E(\mathbf{w}_t) - E(\mathbf{w}^*)$ as a function of time. **Top Right:** convergence time of the various solvers for different choice of regularization parameters. Broken lines correspond to a tolerance of 10^0 , whilst full-lines correspond to 10^{-2} . The thick vertical line indicates the best model selected by cross-validation. **Bottom:** TV-ℓ_1 penalized Least-Squares Regression. **Bottom Left:** on the visual recognition face-house discrimination task; **Bottom Right:** on the Mixed gambles dataset. The thick vertical line indicates the best model selected by cross-validation.

4.1.3 Results: convergence times

Here, we present benchmark results for our experiments. Figure 4.2 gives results for the logistic regression run on the visual recognition dataset: convergence plots of energy as a function of time show that all methods are asymptotically decreasing. The left part of Fig. 4.2 shows the time required to give a convergence threshold, defined as a given excess energy compared to the lowest energy achieved by all methods, for different choices of regularization parameters. Similarly, the right part of Fig. 4.2 shows convergence times for squared loss on both datasets. For these figures, each solver was run for a maximum of 1 hour per problem. Solvers that do not appear on a plot did not converge for the corresponding tolerance and time budget.

For logistic loss, the most serious contender is algorithm 1, LBFGS applied on a smooth surrogate, followed by ADMM, however ADMM performance varies markedly depending on the choice of ν (more on this in chapter 7). For the squared loss FISTA and algorithm 1 are the best performers, with FISTA achieving a clear lead for the larger mixed-gambles dataset. Note that in the case of strong regularization the problem is better conditioned, and first-order methods such as the primal-dual approach can perform well.

Bibliography

A. Beck and M. Teboulle. A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems. *SIAM Journal on Imaging Sciences*, 2: 183, 2009a.

Amir Beck and Marc Teboulle. Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems. *Trans. Img. Proc.*, 18:2419, 2009b.

Amir Beck and Marc Teboulle. Smoothing and First Order Methods: A Unified Framework. *SIAM J. OPTIM.*, 22(2):557–580, 2012.

Jérôme Bobin, Stephen Becker, and Emmanuel Candes. A Fast and Accurate First-order Method for Sparse Recovery. *SIAM J Imaging Sciences*, 2011.

Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. *Fundations and Trends in Machine Learning*, 2010.

J.V. Burke, A.S. Lewis, and M.L. Overton. A Robust Gradient Sampling Algorithm for Nonsmooth, Nonconvex Optimization. *SIAM J. Optimization*, 15:751–779, 2005.

Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *J. Math. Imaging Vis.*, 40, 2011.

I. Daubechies, M. Defrise, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Comm. Pure Appl. Math.*, 57, 2004.

Elvis Dohmatob, Alexandre Gramfort, Bertrand Thirion, and Gaël Varoquaux. Benchmarking solvers for TV-l1 least-squares and logistic regression in brain imaging. In *PRNI*. IEEE, 2014.

Alexandre Gramfort, Bertrand Thirion, and Gaël Varoquaux. Identifying predictive regions from fMRI with TV-L1 prior. In *PRNI*, 2013.

Logan Grosenick, Brad Klingenberg, Kiefer Katovich, Brian Knutson, and Jonathan E Taylor. Interpretable whole-brain prediction analysis with GraphNet. *NeuroImage*, 72, 2013.

James V. Haxby, Ida M. Gobbini, Maura L. Furey, Alumit Ishai, Jennifer L. Schouten, and Pietro Pietrini. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293, 2001.

M. Hebiri and S. van de Geer. The Smooth-Lasso and other $\ell_1 + \ell_2$-penalized methods. *Electron. J. Stat.*, 5, 2011.

Koji Jimura and Russell A Poldrack. Analyses of regional-average activation and multivoxel pattern information tell complementary stories. *Neuropsychologia*, 50, 2012.

Matthieu Kowalski, Kai Siedenburg, and Monika Dörfler. Social Sparsity! Neighborhood Systems Enrich Structured Shrinkage Operators. *Signal Processing, IEEE Transactions on*, 61(10), 2013.

A.S. Lewis and M.L. Overton. Nonsmooth Optimization via BFGS. 2008.

Vincent Michel, Alexandre Gramfort, Gael Varoquaux, Evelyn Eger, and Bertrand Thirion. Total variation regularization for fMRI-based prediction of behaviour. *IEEE Transactions on Medical Imaging*, 30:1328, 2011.

Yu. Nesterov. Smooth minimization of non-smooth functions. *Math. Program.*, Ser. A 103:127–152, 2005a.

Yu. Nesterov. Excessive Gap Technique in Nonsmooth Convex Minimization. *SIAM Journal on Optimization*, 16:235, 2005b.

Yurii Nesterov. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. In *Soviet Mathematics Doklady*, volume 27, pages 372–376, 1983.

Sabrina M. Tom, Craig R. Fox, Christopher Trepe, and Russell A. Poldrack. The Neural Basis of Loss Aversion in Decision-Making Under Risk. *Science*, 315(5811):515–518, 2007.

Gaël Varoquaux, Michael Eickenberg, Elvis Dohmatob, and Bertand Thirion. FAASTA: A fast solver for total-variation regularization of ill-conditioned problems with application to brain imaging. *arXiv:1512.06999*, 2015.

Gaël Varoquaux, Matthieu Kowalski, and Bertrand Thirion. Social-sparsity brain decoders: faster spatial sparsity. In *PRNI conference*, 2016.

Ciyou Zhu, Richard H. Byrd, Peihuang Lu, and Jorge Nocedal. L-BFGS-B– FORTRAN SUBROUTINES FOR LARGE-SCALE BOUND CONSTRAINED OPTIMIZATION. *NORTHWESTERN UNIVERSITY Department of Electrical Engineering and Computer Science*, 1994.

More speed via univariate feature-screening and early-stopping

Contents

5.1 Introduction

In our PRNI 2015 conference paper [Dohmatob et al., 2015], we developed some heuristics for speeding up the overall optimization process: (a) Early-stopping, whereby one halts the optimization process when the test score (performance on left-out data) for the internal cross-validation for model-selection stops improving, and (b) univariate feature-screening, whereby irrelevant (non-predictive) voxels are detected and eliminated before the optimization problem is entered, thus reducing the size of the problem. Empirical results with GraphNet on real MRI (Magnetic Resonance Imaging) datasets indicated that these heuristics are a winning strategy, as they add speed without sacrificing the quality of the predictions / classifications.

One notes that in the case of GraphNet, the penalty term of problem (3.1), the $\|\nabla \mathbf{w}\|_2^2$ sub-term is smooth (i.e differentiable) with *Lipschitz* gradient, whilst the ℓ_1 term though nonsmooth, is proximable by means of the *soft-thesholding* operator [Daubechies et al., 2004]. Thus problem (3.1) is amenable to the FISTA (Fast Iterative Shrinkage-Thresholding Algorithm) [Beck and Teboulle, 2009], with a provable $O(1/\sqrt{\epsilon})$ convergence rate. Our implementation of FISTA uses technical recommendations (line-searching, parametrization, etc.) which were provided in [Dohmatob et al., 2014], in the context of TV-L1 [Baldassarre et al., 2012, Gramfort et al., 2013]. The model parameters α and ρ in (3.1) are set by *internal* cross-validation.

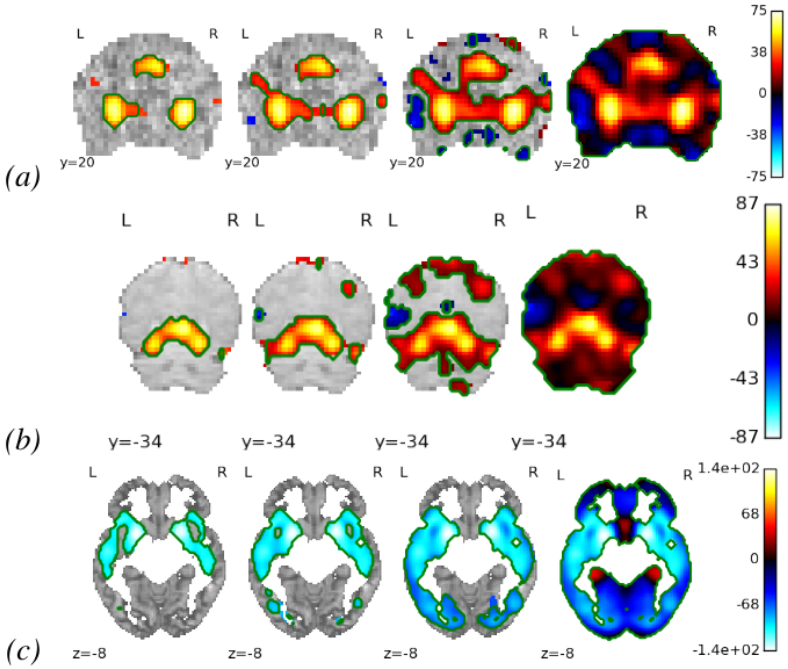


Figure 5.1: Univariate feature-screening for the GraphNet [Hebiri and van de Geer, 2011, Grosenick et al., 2013] problem (3.1) on different datasets. This figure shows spatial maps of $\mathbf{X}_j^T\mathbf{y}$, thresholded so that only voxels j with (from left to rightmost column) $|\mathbf{X}_j^T\mathbf{y}| \geq p_{10\%}(|\mathbf{X}^T\mathbf{y}|)$, $|\mathbf{X}_j^T\mathbf{y}| \geq p_{20\%}(|\mathbf{X}^T\mathbf{y}|)$, $|\mathbf{X}_j^T\mathbf{y}| \geq p_{50\%}(|\mathbf{X}^T\mathbf{y}|)$, and $|\mathbf{X}_j^T\mathbf{y}| \geq p_{100\%}(|\mathbf{X}^T\mathbf{y}|)$ (full-brain) respectively, survive. The green contours enclose the elite voxels which are selected by the screening procedure at the respective threshold levels. *(a)*: Mixed Gambles dataset [Jimura and Poldrack, 2012]. Weights maps obtained for the GraphNet model (3.1) with these different screening-percentiles are shown in Figure 5.4. *(c)*: OASIS dataset [Marcus et al., 2007] with VBM. See Figure 5.2 for weights maps and age predictions obtained using these different screening-percentiles.

5.2 Methods

5.2.1 Univariate feature-screening

In machine-learning, feature-screening aims at detecting and eliminating irrelevant (non-predictive) features thus reducing the size of the underlying optimization problem (here problem (3.1)). The general idea is to compute for each value of the regularization parameter, a *relevance measure* for each feature, which is then compared with a threshold (produced by the screening procedure itself). Features which fall short of this threshold are detected as irrelevant and eliminated. For the Lasso and similar models (including Group Lasso), *exact* screening techniques (i.e, techniques which don't mistakenly discard active predictive features) include those developed in [Ghaoui et al., 2010, Lee and Taylor, 2014, Liu et al., 2014, Wang et al., 2015]. Inexact screening techniques (e.g [Tibshirani et al., 2010]) have also been proposed in the literature.

Our proposed heuristic screening technique is inspired by the *Marginal screening* technique developed in Algorithm 1 of [Lee and Taylor, 2014], and operates as follows. The data $(\mathbf{X}, \mathbf{y})$ are standardized so that $\mathbf{y}$ has unit variance and zero mean, likewise each row of the design matrix $\mathbf{X}$. To ensure obtention of a smooth mask, a Gaussian-smoothed version of $\mathbf{X}$ is used in the screening procedure (but not in the actual model fit). For each voxel j (voxels are the features here) the absolute dot-product $|\mathbf{X}_j^T\mathbf{y}|$ of $\mathbf{y}$ with the jth column of $\mathbf{X}$ is computed. For a given screening-percentile $sp \in [0, 100]$, the spth percentile value of the vector $|\mathbf{X}^T\mathbf{y}| := (|\mathbf{X}_1^T\mathbf{y}|, ..., |\mathbf{X}_p^T\mathbf{y}|)$, denoted $p_{sp}(|\mathbf{X}^T\mathbf{y}|)$, is computed. The case $sp = 100$ corresponds to full-brain analysis with no screening. $sp = 25$ means we keep the quarter of the brain made of voxels with the highest $|\mathbf{X}_j^T\mathbf{y}|$ values, and so on. A brain-mask

is then formed containing only those voxels j for which $|\mathbf{X}_j^T \mathbf{y}| \geq p_{sp}(|\mathbf{X}^T \mathbf{y}|)$. Next, this brain-mask is morphologically eroded and then dilated, to obtain a more structured mask. Figure 5.1 shows results of applying this screening heuristic to various datasets, prior to model fitting.

5.2.2 Early-stopping

Optimization is a means to an end and not an end on its own. The only incentive for optimizing a model is to improve its generalization power: performance on unseen data. If this performs stops improving during training (statistical convergence), as measured on a left-out subset of data , then we may as well abrupt the optimization algorithm. We implement this principle heuristically as follows. In each train sub-sample of the internal cross-validation loop for setting the parameters of the GraphNet model (3.1), a pass is done on the 2-dimensional parameter grid (see Fig. 8.4) and each parameter pair (α, ρ) is scored according to its prediction / classification performance. For a fixed parameter pair (α, ρ), an instance of problem (3.1) is solved iteratively using FISTA Beck and Teboulle [2009]. At each iteration, the prediction / classification performance of the current (not yet optimal) solution $\hat{\mathbf{w}}_k$ in (3.1) is computed. If in a time-window of 5 iterations this score has not increased above an a priori fixed threshold, called the *early-stopping tolerance (es tol)*, then the optimization process is halted for the current model parameter pair (α, ρ) under inspection. This heuristic is motivated by the intuition that, for a particular problem, sub-optimal solutions $\hat{\mathbf{w}}_k$ can give the same score as an optimal solution $\hat{\mathbf{w}}$ (i.e "statistical convergence" happens before numerical convergence). By default we set this early-stopping tolerance to -10^{-4} for classification and -10^{-2} for regression problems. A value of $+\infty$ (in fact, any value above 10, say) corresponds to no early-stopping at all (i.e, solve problem (3.1) until numerical convergence).

5.3 Experiments

We experimented our early-stopping and (separately) feature-screening heuristics on different MRI datasets. All experiments were run using a single core of a laptop.

Regression. The OASIS dataset [Marcus et al., 2007] consists of a cross-sectional collection of 416 subjects aged 18 to 96. For each subject, 3 or 4 individual T1-weighted MRI scans obtained in single scan sessions are included. A natural regression problem for this dataset is to predict the age of a subject from their anatomical data. To this end, we segmented the gray-matter from the anatomical data of each subject (obtained from the T1 images), and used the gray-matter maps as features for predicting age. We split the 416 subjects into two equally-sized and age-balanced groups: a train set and a validation set. The GraphNet model [Hebiri and van de Geer, 2011, Grosenick et al., 2013] was fitted on the train set, with parameters (α and ρ in (3.1)) set internally via 8-fold cross-validation. The results for this experiment are shown in Figure 5.2.

Classification. The visual recognition dataset [Haxby et al., 2001] is a popular block-design fMRI dataset from a study on face and object representation in human ventral temporal cortex. It consists of 6 subjects with 12 sessions / runs per subject. In each run, the subjects passively viewed images of eight object categories, grouped in 24-second blocks separated by intermittent rest periods. This experiment is a classification task: predicting the object category $\mathbf{y}$. We use a *One-versus-Rest (OvR)* strategy. The design matrix $\mathbf{X}$ is made of time-series from the full-brain mask of $p = 23\,707$ voxels over $n = 216$ TRs, of a single subject (subj1). We divided the 12 runs into 6 runs for training and 6 other runs for validation. Leave-one-session-out [1] cross-validation was used for selecting the model parameters (α, ρ). The results are depicted in Figure 5.4.

[1] One session is held out and the other $S - 1$ sessions are used to train a model which is validated on the left-out session. This is repeated for all the sessions, yield an estimate –with error bars– on the generalization error of the model.

5.4 Results

We now summarize and comment the results of the experiments (refer to section 4.1.2). Figure 5.2 shows the effects of early-stopping heuristic and feature-screening heuristic on age prediction scores on the OASIS dataset [Marcus et al., 2007] (416 subjects). We see that in the internal cross-validation, stopping the optimization procedure for fixed (α, ρ) pair of regularization parameters, when test score increases by -10^{-2} or more is a good heuristic, and does just as good as running the optimization until numerical convergence. Also (and independently), one gets similar prediction scores using

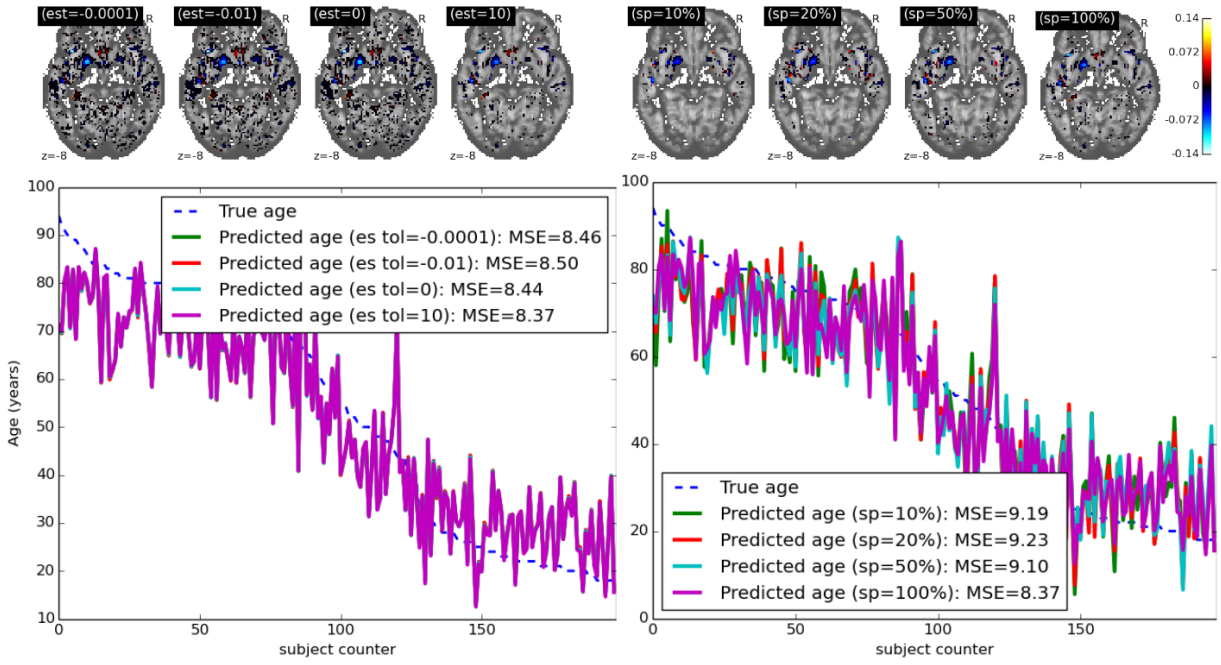


Figure 5.2: Predicting age from gray-matter concentration maps from the OASIS dataset [Marcus et al., 2007]. **Top**: Weights maps (solutions to problem (3.1)). **Bottom-left**: Mean Square Error (MSE) in age prediction, for different subjects of the validation set, for varying levels of the early-stopping tolerance ("es tol" for short), with the screening-percentile (sp) held constant at 100 (full-brain). **Bottom-right**: MSE in age prediction, for varying levels of the screening-percentile (sp).

as little as a fifth of the brain volume ($sp = 20$), compared to using the full-brain ($sp = 100$). Figure 5.4 reports similar results for classification on the visual recognition dataset [Haxby et al., 2001]. Overall, we see from Figures 5.4 and 5.2 that we can achieve up to 10-fold speedup with the proposed heuristics, with very little loss in accuracy. Also, we see that contiguous groups of bars are roughly flat at the top, with a sligh increase from lower to high screening-percentile values. The case "chair vs scramped" is an exception, where a slightly reverse tendency if observed. A possible explanation is that 20th percentile feature-screening already selects the right voxels (quasi-exact support recovery), and so including more voxels in the model can only hurt its performance.

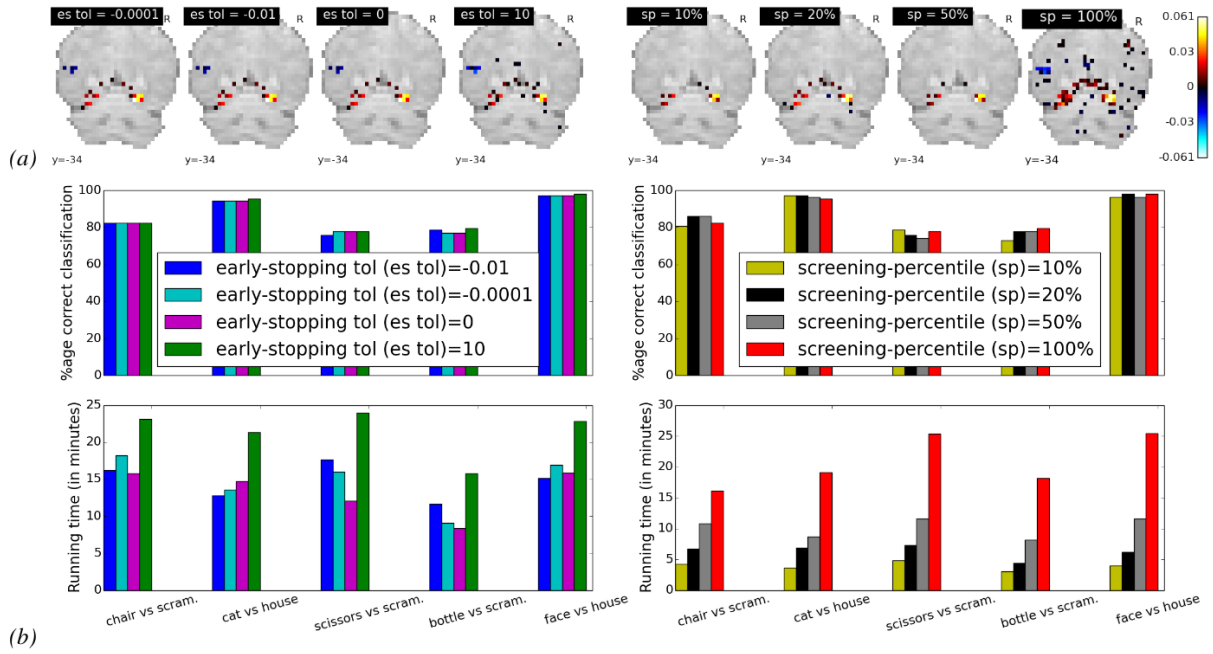


Figure 5.3: Predicting age from gray-matter concentration maps from the OASIS dataset [Marcus et al., 2007]. **Top**: Weights maps (solutions to problem (3.1)). **Bottom-left**: Mean Square Error (MSE) in age prediction, for different subjects of the validation set, for varying levels of the early-stopping tolerance ("es tol" for short), with the screening-percentile (sp) held constant at 100 (full-brain). **Bottom-right**: MSE in age prediction, for varying levels of the screening-percentile (sp). **Running times**: Increasing *est tol* (from -10^{-4} to 10): **100.2m, 171.4m, 188.8m, 289.6m**. For increasing *sp* (10 to 100): **44.2m, 81.3m, 186.5m, 341.3m**

Figure 5.4: Visual recognition dataset [Haxby et al., 2001]. ***(a)***: Weights maps for the Face vs House contrast, for different early-stopping and univariate feature-screening thresholds. One can see that the supports of these maps for different values of the thresholds are quite similar to cases involving no heuristic at all (the case where est = 10 and the where case sp = 100%). ***(b)*, top-left**: Prediction scores as a function of the early-stopping tolerance (est), for different task contrasts. ***(b)*, top-right**: Prediction scores as a function of the screening-percentile (sp), for different task contrasts. ***(b)*, bottom-row**: Running times in minutes for the different thresholds of the heuristics.

5.5 Conclusion

We have presented heuristics that provide speedups for optimizing GraphNet [Hebiri and van de Geer, 2011, Grosenick et al., 2013] in the difficult

context of brain data. These heuristics are a winning strategy as they add speed without sacrificing the quality of the predictions / classifications. In practice, we do a 20univariate feature-screening by default, which ensures a 5-fold speedup over full-brain analysis, and independently of an approximately 2-fold speedup obtained by the early-stopping heuristic, leading to an overall 10-fold speedup. Our results have been verified empirically on different MRI datasets. Our heuristics should be applicable to other hard-to-optimize models like TV-L1 [Baldassarre et al., 2012, Gramfort et al., 2013].

The result of these numerous contributions on optimizing the SpaceNet model (3.1) have been implemented as part of the *Nilearn* package [Abraham et al., 2014].

Bibliography

Alexandre Abraham, Fabian Pedregosa, Michael Eickenberg, Philippe Gervais, Andreas Mueller, Jean Kossaifi, Alexandre Gramfort, Bertrand Thirion, and Gael Varoquaux. Machine learning for neuroimaging with scikit-learn. *Frontiers in Neuroinformatics*, 2014.

Luca Baldassarre, Janaina Mourao-Miranda, and Massimiliano Pontil. Structured sparsity models for brain decoding from fMRI data. In *PRNI*, 2012.

A. Beck and M. Teboulle. A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems. *SIAM Journal on Imaging Sciences*, 2, 2009.

I. Daubechies, M. Defrise, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Comm. Pure Appl. Math.*, 57, 2004.

Elvis Dohmatob, Alexandre Gramfort, Bertrand Thirion, and Gaël Varoquaux. Benchmarking solvers for TV-l1 least-squares and logistic regression in brain imaging. In *PRNI*. IEEE, 2014.

Elvis Dohmatob, Michael Eickenberg, Bertrand Thirion, and Gaël Varoquaux. Speeding-up model-selection in GraphNet via early-stopping and univariate feature-screening. In *Pattern Recognition in NeuroImaging (PRNI), 2015 International Workshop on*. IEEE, 2015.

Laurent El Ghaoui, Vivian Viallon, and Tarek Rabbani. Safe Feature Elimination in Sparse Supervised Learning. *CoRR*, abs/1009.3515, 2010.

Alexandre Gramfort, Bertrand Thirion, and Gaël Varoquaux. Identifying predictive regions from fMRI with TV-L1 prior. In *PRNI*, 2013.

Logan Grosenick, Brad Klingenberg, Kiefer Katovich, Brian Knutson, and Jonathan E Taylor. Interpretable whole-brain prediction analysis with GraphNet. *NeuroImage*, 72, 2013.

James V. Haxby, Ida M. Gobbini, Maura L. Furey, Alumit Ishai, Jennifer L. Schouten, and Pietro Pietrini. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293, 2001.

M. Hebiri and S. van de Geer. The Smooth-Lasso and other $\ell_1 + \ell_2$-penalized methods. *Electron. J. Stat.*, 5, 2011.

Koji Jimura and Russell A Poldrack. Analyses of regional-average activation and multivoxel pattern information tell complementary stories. *Neuropsychologia*, 50, 2012.

Jason D Lee and Jonathan E Taylor. Exact post model selection inference for marginal screening. In *NIPS*, 2014.

Jun Liu, Zheng Zhao, Jie Wang, and Jieping Ye. Safe Screening With Variational Inequalities and Its Applicaiton to LASSO. In *ICML*, 2014.

Daniel S Marcus, Tracy H Wang, Jamie Parker, John G Csernansky, John C Morris, and Randy L Buckner. Open Access Series of Imaging Studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. *Journal of cognitive neuroscience*, 19, 2007.

Robert Tibshirani, Jacob Bien, Jerome Friedman, Trevor Hastie, Noah Simon, Jonathan Taylor, and Ryan J Tibshirani. Strong rules for discarding predictors in lasso-type problems. *J. R. Stat. Soc.: Series B*, 74, 2010.

Jie Wang, Peter Wonka, and Jieping Ye. Lasso Screening Rules via Dual Polytope Projection. *Journal of Machine Learning Research*, 2015.

On the equivalence of TV-L1 and iteratively-reweighted GraphNet

Contents

We present a result that shows that TV-L1 regularized regression problems (3.1) can also be solved through an iteratively reweighted GraphNet problem. Among other things, this provides a long-awaited statistical interpretation of the TV-L1 penalized models (3.1). The method dubbed *iGraphNet*, solves TV-L1 penalized model by considering modified GraphNet sub-problems corresponding to the minimization of the energy $E^{\boldsymbol{\gamma}}_{\text{GraphNet}}(\mathbf{w})$ defined in (6.5). These sub-problems are very well-conditioned and are quadratically easier to solve than TV-L1 itself. The limit of this sub-problems is solve the exact TV-L1 penalized problem.

This work follows the spirit of [Candes et al., 2007] which proposed an enhanced Lasso problem built iteratively from surrogate Ridge regression problems with inhomogeneous feature penalty parameters. However, unlike [Candes et al., 2007], we leave the Lasso part of the TV-L1 penalty (3.2) untouched and instead derive a surrogate on the TV part, which turns out to be a GraphNet problem with inhomogeneous penalty parameters. See Figure 6.1. Pending figures comparing (maps, scores, and runtime) GraphNet, iGraphNet, and the baseline TV-L1 implementation via double-FISTA implementations [Dohmatob et al., 2014, Varoquaux et al., 2015].

6.1 Derivation

Invoking the following well-known elementary result[1]

$$\forall u, w > 0, u \leq \frac{wu^2 + w^{-1}}{2}, \text{ with equality iff } w = u^{-1}, \tag{6.1}$$

we can rewrite the TV semi-norm as follows,

$$\|\mathbf{w}\|_{\text{TV}} := \sum_{j \in [\![p]\!], \|(\nabla \mathbf{w})_j\|_2 > 0} \|(\nabla \mathbf{w})_j\|_2 \leq \frac{1}{2} \sum_{j \in [\![p]\!], \|(\nabla \mathbf{w})_j\|_2 > 0} \gamma_j \|(\nabla \mathbf{w})_j\|_2^2 + \gamma_j^{-1}, \forall \boldsymbol{\gamma} \in \mathbb{R}^p_{++}, \tag{6.2}$$

[1] To prove it, one simply uses the fact that $w^2u^2 + 1 - 2wu = (wu - 1)^2 \geq 0$, with equality iff $wu = 1$.

with equality iff

$$\gamma_j = \|(\nabla \mathbf{w})_j\|_2^{-1}, \ \forall j \in [\![p]\!] \text{ s.t } \|(\nabla \mathbf{w})_j\|_2 > 0. \tag{6.3}$$

Thus,

$$\|\mathbf{w}\|_{\text{TV}} = \min_{\boldsymbol{\gamma} \in \mathbb{R}^p_{++}} \frac{1}{2} \sum_{j \in [\![p]\!], \|(\nabla \mathbf{w})_j\|_2 > 0} \gamma_j \|(\nabla \mathbf{w})_j\|_2^2 + \gamma_j^{-1}, \tag{6.4}$$

with the optimal scaling vector $\boldsymbol{\gamma} \in \mathbb{R}^p_{++}$ given by (6.3). Whence, the minimizers of the TV-L1 energy $E_{\text{TV-L1}}(\mathbf{w}) := \ell(\mathbf{y}, \mathbf{X}\mathbf{w}) + \alpha \mathcal{P}_{\text{TV-L1}}(\mathbf{w})$ coincide with the minimizers of the rescaled GraphNet energy

$$E^{\boldsymbol{\gamma}}_{\text{GraphNet}}(\mathbf{w}) = \ell(\mathbf{y}, \mathbf{X}\mathbf{w}) + \alpha\rho\|\mathbf{w}\|_1 + \frac{1}{2}\alpha(1-\rho) \sum_{j \in [\![p]\!], \|(\nabla w)_j\|_2 > 0} \gamma_j \|(\nabla \mathbf{w})_j\|_2^2 + \gamma_j^{-1}, \tag{6.5}$$

where the minimization is done both over regression coefficients $\mathbf{w}$ and the scaling parameters $\gamma_1, \ldots, \gamma_p > 0$.

Algorithm 2: iGraphNet: iteratively-reweighted GraphNet solver for the TV-L1 model

Require: Values for the model-tuning parameters $\lambda > 0$, and $0 \leq \rho \leq 1$; initial brain-map $\mathbf{w}^{(0)} \in \mathbb{R}^p$ (e.g, the zero vector); tolerance threshold $\epsilon > 0$ (say 10^{-5}); maximum number of outer iterations K.

Ensure: An optimal vector $\hat{\mathbf{w}}_{TV}$ of regressor coefficients (an approximation of) for the TV-L1 model.

1: **Initialize:** $k \leftarrow 0$; $\mu \leftarrow 10^{-4}$
2: **while** $\|\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\|_\infty \geq \epsilon$ **do**
3: **Recompute scaling:** $\gamma_j^{(k)} \leftarrow (\|(\nabla \mathbf{w}^{(k)})_j\|_2^2 + \mu^2)^{-\frac{1}{2}}$, for every voxel j
4: **Recompute coefficients:** $\mathbf{w}^{(k+1)} \leftarrow \arg\min_{\mathbf{w} \in \mathbb{R}^p} E^{\boldsymbol{\gamma}^{(k)}}_{\text{GraphNet}}(\mathbf{w})$, with energy tolerance $\sim \mu$. The solver for this sub-problem is warm-started with $\mathbf{w} = \mathbf{w}^{(k)}$.
5: **Goto next iteration:** $k \leftarrow k + 1$
6: **end while**

As a function of the regressor coefficients $\mathbf{w}$, the energy in (6.5) corresponds to a modified GraphNet model in which per-voxel penalty parameters $\alpha(1-\rho)\gamma_j$ given by (6.3) replace the constant $\alpha(1-\rho)$ factor in the pure GraphNet model (3.1), or equivalently the ∇ is pre-*whitened* by the diagonal matrix $\boldsymbol{\Gamma} := \text{diag}(\sqrt{\gamma_1}, \ldots, \sqrt{\gamma_p})$. This energy is optimized by an alternating scheme cyclically switching between optimizing w.r.t the regressor coefficients $\mathbf{w}$ and then w.r.t then rescaling parameters $\gamma_1, \ldots, \gamma_p$ (in closed form, via formula (6.3)). The algorithm so-obtained (detailed in section 6.2) alternates between minimization over the scaling parameters $\boldsymbol{\gamma}$ and minimization over the coefficients $\mathbf{w}$.

6.2 The algorithm: iGraphNet

We now present iGraphNet, an iteratively-reweighted scheme for solving the TV-L1 model, based on modified GraphNet (3.1) sub-problems corresponding to the minimization of the energy $E^{\boldsymbol{\gamma}}_{\text{GraphNet}}(\mathbf{w})$ defined in (6.5).

These sub-problems are very well-conditioned and are quadratically easier to solve than TV-L1 itself, and can be solved by a fast first-order method like FISTA or LARS. The algorithm is presented in Alg. 2.

Overall, for a tolerance $\epsilon > 0$, Alg. 2 converges in $O(1/\epsilon)$ basic iterations (i.e counting all the iterations run in a first-order method for solving the GraphNet sub-problem), though its observed runtime is in the order of about K times the time taken by a run of a solver for the GraphNet sub-problem. Practical details (like handling a brain mask, automatic model parameter selection via cross-validation and bagging, early-stopping, etc.) that go in the implementation of the optimization algorithms like the one just presented can be found in [Dohmatob et al., 2014].

Generalization to other complex non-smooth models. Similarly, one can show that Sparse-Variation [Eickenberg et al., 2015] can be solved via an IRLS (iteratively-reweighted Least Squares) scheme, where the weights are computed via (6.3), with the ∇ operator replaced with an identity-augmented version. Indeed, thanks to the inequality (6.1), it turns out that most complex rich non-smooth ℓ_p-norm-based models are just iteratively-reweighted versions of much simpler counterparts like Ordinary Least Squares, Lasso, ElasticNet, GraphNet, etc.

6.3 Experimental results

Preliminary experimental results are shown in Fig. 6.1. We run our iGraphNet procedure (Alg. 2) on data for the Face vs House condition of the visual recognition dataset [Haxby et al., 2001]. Model coefficients and accuracies on held-out data are shown. We monitor the evolution of the model as a function of the number of iGraphNet iterations $k = 0, 1, 2, \ldots$. We see that as more and more iterations of iGraphNet are run, the coefficients become more and more spatially denoised and localized (and therefore more intepretable), without deterioration of prediction accuracy.

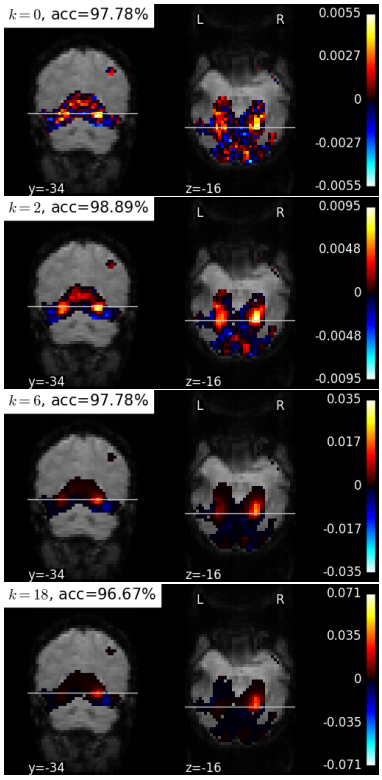


Figure 6.1: Estimated coefficients $\hat{\mathbf{w}}$ on the Face vs House condition of the visual recognition dataset [Haxby et al., 2001]. Classification accuracies on held-out data are shown in the legends. We monitor the evolution of the model as a function of the number of iGraphNet iterations $k = 0, 1, 2, \ldots$. We see that as more and more iterations of iGraphNet are run, the coefficients become more and more spatially denoised and localized (and therefore more intepretable), without deteriorating of the model accuracy.

Bibliography

Emmanuel J Candes, Michael B Wakin, and Stephen P Boyd. Enhancing Sparsity by Reweighted l1 Minimization. 2007.

Elvis Dohmatob, Alexandre Gramfort, Bertrand Thirion, and Gaël Varoquaux. Benchmarking solvers for TV-l1 least-squares and logistic regression in brain imaging. In *PRNI*. IEEE, 2014.

Michael Eickenberg, Elvis Dohmatob, Bertrand Thirion, and Gaël Varoquaux. Total Variation meets Sparsity: statistical learning with segmenting penalties. In *MICCAI*. 2015.

James V. Haxby, Ida M. Gobbini, Maura L. Furey, Alumit Ishai, Jennifer L. Schouten, and Pietro Pietrini. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293, 2001.

Gaël Varoquaux, Michael Eickenberg, Elvis Dohmatob, and Bertand Thirion. FAASTA: A fast solver for total-variation regulariza-

tion of ill-conditioned problems with application to brain imaging. *arXiv:1512.06999*, 2015.

A result on the rate of convergence of the ADMM algorithm

Contents

7.1 Introduction

The ADMM algorithm [Glowinski and Marroco, 1975, Gabay and Mercier, 1976, Eckstein and Bertsekas, 1992] is an operator-splitting optimization method which is easy to implement and well-adapted for large-scale optimization problems [Boyd et al., 2011]. ADMM can provide a distinctive advantage over proximal gradient methods such as FISTA [Beck and Teboulle, 2009] when there is no closed-form expression for the proximal operator. Indeed, ADMM can avoid this difficulty by introducing a "split" variable, for which the proximal operator results in updates computable in closed-form. This is typically the case in *analysis sparsity* regularization, that impose sparsity on a transformation of the optimization variable. However, the theory of the convergence rate of ADMM is not complete [Boyd et al., 2011].

In our ICASSP 2016 paper [Dohmatob et al., 2015], we studied the convergence of the ADMM (Alternating Direction Method of Multipliers) algorithm on a broad range of penalized regression problems including the Lasso, Group-Lasso and Graph-Lasso,(isotropic) TV-L1, Sparse Variation, and others, that can be written in the form

$$\underset{(\mathbf{w},\mathbf{z})\in\mathbb{R}^p\times\mathbb{R}^q}{\text{minimize}} \frac{1}{2}\|\mathbf{X}\mathbf{w}-\mathbf{y}\|^2+\lambda\Omega(\mathbf{z}) \text{ subject to } \mathbf{K}\mathbf{w}-\mathbf{z}=0, \tag{7.1}$$

where $\mathbf{X}\in\mathbb{R}^{n\times p}$ is the design matrix; $\mathbf{y}\in\mathbb{R}^n$ is a vector of measurements or classification targets; $\mathbf{K}\in\mathbb{R}^{q\times p}$ is linear operator; $\lambda>0$ is the regularization parameter; and $\Omega:\mathbb{R}^p\to(-\infty,+\infty]$ is the penalty, which is assumed to be a *closed proper convex* function. This is an instance of the SpaceNet model (3.1) presented in chapter 3. In signal processing literature, (7.1) is an example of what is referred to as a synthesis problem: the penalty Ω is imposed not directly on the image, but on a the output of a dictionary, $\mathbf{z}=\mathbf{K}\mathbf{w}$. $\mathbf{K}$ is referred to the analysis operator. The case $\mathbf{K}=\mathbf{I}$ corresponds to the *synthesis* setting.

7.1.1 The ADMM algorithms

Consider the ADMM algorithm [Glowinski and Marroco, 1975, Gabay and Mercier, 1976, Eckstein and Bertsekas, 1992, Boyd et al., 2011] applied to problem (7.1). Let $\boldsymbol{\mu}\in\mathbb{R}^q$ be the dual variable and $\nu>0$ be the penalty parameter on the splitting residual. The augmented Lagrangian is:

$$\mathcal{L}_\nu(\mathbf{w},\mathbf{z},\boldsymbol{\mu})=\frac{1}{2}\|\mathbf{X}\mathbf{w}-\mathbf{y}\|^2+\lambda\Omega(\mathbf{z})+\boldsymbol{\mu}^T(\mathbf{K}\mathbf{w}-\mathbf{z})+\frac{1}{2}\nu\|\mathbf{K}\mathbf{w}-\mathbf{z}\|^2.$$

Further, introducing the scaled dual variable $\mathbf{u}:=\nu^{-1}\boldsymbol{\mu}$, which we will use instead of $\boldsymbol{\mu}$ from here on, the ADMM iterates for problem (7.1) are given by the following equations:

$$\begin{aligned}
\mathbf{w}^{(n+1)} &\leftarrow \arg\min_{\mathbf{w}} \mathcal{L}_\nu(\mathbf{w},\mathbf{z}^{(n)},\mathbf{u}^{(n)}) = (\nu\mathbf{K}^T\mathbf{K}+\mathbf{X}^T\mathbf{X})^{-1}(\nu\mathbf{K}^T(\mathbf{z}^{(n)}-\mathbf{u}^{(n)})+\mathbf{X}^T\mathbf{y})\\
\mathbf{z}^{(n+1)} &\leftarrow \arg\min_{\mathbf{z}} \mathcal{L}_\nu(\mathbf{w}^{(n+1)},\mathbf{z},\mathbf{u}^{(n)}) = \text{prox}_{(\alpha/\nu)\Omega}(\mathbf{K}\mathbf{w}^{(n+1)}+\mathbf{u}^{(n)})\\
\mathbf{u}^{(n+1)} &\leftarrow \mathbf{u}^{(n)}+\mathbf{K}\mathbf{w}^{(n+1)}-\mathbf{z}^{(n+1)}.
\end{aligned} \tag{7.2}$$

Assumptions. We will assume that the matrix sum $\nu\mathbf{K}^T\mathbf{K}+\mathbf{X}^T\mathbf{X}$ is invertible. This assumption is equivalent to $\ker\mathbf{K}^T\mathbf{K}\cap\ker\mathbf{X}^T\mathbf{X}=\{0\}$ (see e.g [Piziak et al., 1999, Theorem 1]), which is reasonable in the context of regularization. Indeed, the idea behind this assumption is that, in high-dimensional problems ($n\ll p$), $\mathbf{X}$ typically has a large kernel, and so one would naturally choose $\mathbf{K}$ to act on it.

7.1.2 Examples

Problem (7.1) covers a broad spectrum of problems encountered in pattern recognition and image processing. Here are a few:

Classical examples. We have $\Omega=\frac{1}{2}\|.\|^2$ for Ridge regression; $\Omega=\|.\|_1:z\mapsto\sum_{j\in[p]}|z_j|$ for Lasso and Fused-Lasso [Tibshirani et al., 2005]. For all but the last of these examples, we have $\mathbf{K}=\mathbf{I}$. For Group-Lasso, we have

$\mathbf{K} = \mathbf{I}$, Ω = the *mixed-norm* $\ell_{2,1} = \|.\|_{2,1} : z \mapsto \sum_{j \in [\![d]\!]} \|z_{j:j+c-1}\|$, where there are $d \geq 1$ blocks $z_{j:j+c-1} := (z_j, z_{j+1}, ..., z_{j+c-1})$ each of size $c \geq 1$.

Isotropic TV-L1 and Sparse Variation. The different extensions of the TV penalty presented in chapter 3 can be posed in the form of the problem above. For example, Sparse Variation [Eickenberg et al., 2015] corresponds to taking $\mathbf{K} = \left[\rho\mathbf{I}, \ (1-\rho)\nabla\right]^T \in \mathbb{R}^{4p \times p}$, where ∇ is the discrete (refer to chapter 3) spatial gradient operator and $\rho \in [0, 1]$ is a mixing parameter. For TV-L1 [Baldassarre et al., 2012, Gramfort et al., 2013], the penalty is given by $\Omega(\mathbf{z}) = \sum_{j \in [\![p]\!]} |z_{j,1}| + \sum_{j \in [p]} \|z_{j,2:4}\|$ (i.e an ℓ_1 norm on the first p coordinates of z and an $\ell_{2,1}$ mixed-norm on the last $3p$ coordinates). In particular, the case $\rho = 1$ corresponds to the usual ℓ_1 norm, while $\rho = 0$ corresponds to the isotropic TV semi-norm.

In Sparse Variation [Eickenberg et al., 2015], the penalty is modified to simply be an $\ell_{2,1}$ mixed-norm on $d = p$ blocks of size $c = 4$ each, i.e $\Omega(\mathbf{z}) = \sum_{j \in [p]} \|z_{j,1:4}\|$. TV-L1 and Sparse Variation combine sparsity (due to the the ℓ_1-norm) and structure (due to the isotropic TV term) to extract local concentrations of spatially correlated features from the data.

7.2 Our contributions

7.2.1 Preliminaries

In the spirit of [Ghadimi et al., 2013], let us start with a simple lemma (proof omitted) which rewrites the ADMM iterates (7.2) as a Picard fixed-point process in terms of the $(\mathbf{z}, \mathbf{u})$ pair of variables.

Lemma 1. *Define the following objects:*

$$\mathbf{G}_\nu := \mathbf{K}(\mathbf{K}^T\mathbf{K} + \nu^{-1}\mathbf{X}^T\mathbf{X})^{-1}\mathbf{K}^T, \ \ \mathbf{A}_\nu := \left[\mathbf{G}_\nu \ \ \mathbf{I} - \mathbf{G}_\nu\right],$$
$$\mathbf{b}_\nu := \nu^{-1}\mathbf{K}(\mathbf{K}^T\mathbf{K} + \nu^{-1}\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}, \ \tilde{\mathbf{A}}_\nu := \mathbf{A}_\nu(.) + \mathbf{b}_\nu,$$
$$\Lambda_\nu := \left(\text{prox}_{(\alpha/\nu)\varphi} \circ \tilde{\mathbf{A}}_\nu, (\mathbf{I} - \text{prox}_{(\alpha/\nu)\varphi}) \circ \tilde{\mathbf{A}}_\nu\right).$$

Then the z and u updates in the ADMM iterates (7.2) *can be jointly written as a Picard fixed-point iteration for the operator Λ_ν, i.e*

$$(\mathbf{z}^{(n+1)}, \mathbf{u}^{(n+1)}) \leftarrow \Lambda_\nu(\mathbf{z}^{(n)}, \mathbf{u}^{(n)}). \tag{7.3}$$

In the special case where $\text{prox}_{(\alpha/\nu)\varphi}$ is a linear transformation –as in Ridge regression or the nonnegative Lasso, for example– the operator Λ_ν is linear so that the fixed-point iteration (7.3) is a linear dynamical system. Moreover, in such cases one can derive closed-form formulae for the spectral radius $r(\Lambda_\nu)$ of Λ_ν as function of ν, and thus recover the results of [Ghadimi et al., 2013] and [Boley, 2013]. In the latter simple situations, a strategy for speeding up the ADMM algorithm is then to choose the parameter ν so that the spectral radius of the linear part of the then affine transformation Λ_ν is minimized. The following Corollary is immediate, whose proof is obtainable via the *Spectral Mapping Theorem.*

Corollary 1. *Let $\mathbf{G}_\nu$, $\mathbf{A}_\nu$, $\tilde{\mathbf{A}}_\nu$, and Λ_ν be defined as in Lemma 1. Then the following hold:*

(a) $\max(\|\mathbf{G}_\nu\|, \|\mathbf{I}-\mathbf{G}_\nu\|) \leq 1$, $\nu_{\min^*}(\mathbf{A}_\nu) \geq 1/\sqrt{2}$, *and* $\|\mathbf{A}_\nu\| \leq 1$ *with equality in the last inequality iff at least one of* $\mathbf{G}_\nu$ *and* $\mathbf{I}-\mathbf{G}_\nu$ *is singular.*

(b) Λ_ν *is* $\|\mathbf{A}_\nu\|$*-Lipschitz. That is,* $\forall(\mathbf{x}_1, \mathbf{x}_2) \in \mathbb{R}^{q+q} \times \mathbb{R}^{q+q}$,

$$\|\Lambda_\nu(\mathbf{x}_1) - \Lambda_\nu(\mathbf{x}_2)\| \leq \|\mathbf{A}_\nu\|\|\mathbf{x}_1 - \mathbf{x}_2\|. \tag{7.4}$$

In particular, if $\|\mathbf{A}_\nu\| < 1$, *then* Λ_ν *is a contraction and the ADMM iterates* (7.2) *converge globally Q-linearly to a solution of* (7.1). *Moreover, this solution is unique.*

According to Corollary 1, Λ_ν is an $\|\mathbf{A}_\nu\|$-contraction in case $\|\mathbf{A}_\nu\| < 1$, and so we have global Q-linear convergence of the ADMM iterates (7.2) at the rate $\|\mathbf{A}_\nu\|$. This particular case is analogous to the results obtained in [Nishihara et al., 2015] when the loss function or the penalty is strongly convex. But what if $\|\mathbf{G}_\nu\| = \|\mathbf{I}-\mathbf{G}_\nu\| = \|\mathbf{A}_\nu\| = 1$? Can we still have Q-linear convergence, –at least locally ? These questions are answered in the sequel.

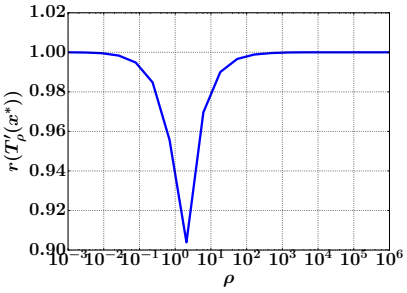


Figure 7.1: Rate of convergence $r(\Lambda'_\nu(\mathbf{x}^*))$ as a function of ν for a Lasso problem with column-rank deficient design matrix $\mathbf{X}$. Taking ν too small leads to badly conditioned problem (as $\mathbf{I} + (1/\nu)\mathbf{X}^T\mathbf{X}$ is then almost singular), and thus a slow rate of convergence (near 1). On the other hand, the figure suggests that taking ν "too large" is also detrimental. Most remarkable, one notices that the basin of "good" ν values is rather tight, and so care must be taken in choosing the ν parameter.

7.2.2 Behavior of ADMM around fixed-points

Henceforth, we consider problem (7.1) in situations where the penalty φ is an $\ell_{2,1}$ mixed-norm. Note that the ℓ_1-norm is a special case of the $\ell_{2,1}$ mixed-norm with $c = 1$ feature per block, and corresponds to the anisotropic case. The results presented in Theorem (1) carry over effortlessly to the case where the φ is the concatenation of $\ell_{2,1}$ norms, for example as in the the TV-L1 semi-norm. The following theorem –inspired by a careful synthesis of the arguments in [Holmes, 1973] and [Bayram and Selesnick, 2010]– is our main result.

Our main results are summarized in Theorem 1 of the aforementioned paper, which we now state.

Theorem 1. *Consider the ADMM algorithm* (7.2) *on problem* (7.1), *where* Ω *is an* $\ell_{2,1}$ *mixed-norm on* $d \geq 1$ *blocks each of size* $c \geq 1$, *for a total of* $q = d \times c$ *features. Let the operators* $\mathbf{A}$, $\tilde{\mathbf{A}}$, *and* Λ *be defined as defined above, with the* ν *subscript dropped for ease of notation. Let For* $\mathbf{x} = (\mathbf{z}, \mathbf{u}) \in \mathbb{R}^{q+q}$, *let* $\Lambda_1(\mathbf{x}) \in \mathbb{R}^q$ *denote the first* q *coordinates of* $\Lambda(\mathbf{x})$, *i.e its* z*-part. Define*

- $\operatorname{supp}(\mathbf{z}) := \{j \in [\![d]\!] \mid z_{j:j+c-1} \neq 0\}$;
- $\mathcal{A}_1(\mathbf{z}) := \{\mathbf{z}' \in \mathbb{R}^q \mid \operatorname{supp}(\mathbf{z}') = \operatorname{supp}(\mathbf{z})\}$, *and* $\mathcal{A}(\mathbf{x}) := \mathcal{A}_1(\mathbf{z}) \oplus \mathbb{R}^q$;
- $\tilde{\mathbf{x}} := (\tilde{\mathbf{x}}_j)_{j\in[\![d]\!]} := \tilde{\mathbf{A}}\mathbf{x}$, $\kappa := \alpha/\nu$, $\epsilon(\mathbf{x}) := \min_{j\in[\![d]\!]} |\|\tilde{\mathbf{x}}_j\| - \kappa| \geq 0$.

Then the following hold:

(a) ***Attractivity of supports.*** *For all* $\mathbf{x} \in \mathbb{R}^{q+q}$, *we have*

$$\Lambda(\bar{\mathbb{B}}_{2q}(\mathbf{x}, \epsilon(\mathbf{x})/\|\mathbf{A}\|)) \subseteq \bar{\mathbb{B}}_{2q}(\Lambda(\mathbf{x}), \epsilon(\mathbf{x})) \cap \mathcal{A}(\Lambda(\mathbf{x})).$$

In particular, if $\mathbf{x}^*$ *is a fixed-point of the operator* Γ, *then*

$$\Lambda(\bar{\mathbb{B}}_{2q}(\mathbf{x}^*, \epsilon(\mathbf{x}^*)/\|\mathbf{A}\|)) \subseteq \bar{\mathbb{B}}_{2q}(\mathbf{x}^*, \epsilon(\mathbf{x}^*)) \cap \mathcal{A}(\mathbf{x}^*).$$

(b) ***Fréchet-differentiability.*** *If* $\mathbf{x} \in \mathbb{R}^{q+q}$ *with* $\epsilon(\mathbf{x}) > 0$, *then* Λ *is Fréchet-differentiable at* $\mathbf{x}$ *with derivative*

$$\Lambda'(\mathbf{x}) = \mathbf{F}_{\mathbf{x}}\mathbf{A} \in \mathbb{R}^{2q\times 2q}, \tag{7.5}$$

where $\mathbf{F}_{\mathbf{x}} := [\mathbf{D}_{\mathbf{x}} \;\; \mathbf{I} - \mathbf{D}_{\mathbf{x}}]^T$ *and* $\mathbf{D}_{\mathbf{x}} \in \mathbb{R}^{q\times q}$ *is a block-diagonal matrix with block* $\mathbf{D}_{\mathbf{x},j} \in \mathbb{R}^{c\times c}$ *given by*

$$\mathbf{D}_{\mathbf{x},j} = \begin{cases} \mathbf{I} - \frac{\kappa}{\|\tilde{\mathbf{x}}_j\|} P_{\langle \tilde{\mathbf{x}}_j \rangle^\perp}, & \text{if } j \in \operatorname{supp}(\Lambda_1(\mathbf{x})), \\ 0, & \text{otherwise.} \end{cases} \tag{7.6}$$

In particular, when $c = 1$, *each* $\mathbf{D}_{\mathbf{x},j}$ *reduces to a bit* $\in \{0, 1\}$ *which indicates whether the* j*th feature is active, and* $\mathbf{D}_{\mathbf{x}}$ *reduces to a diagonal projector matrix with only 0s and 1s.*

(c) *Let* $\mathbf{x}^* = (\mathbf{z}^*, \mathbf{u}^*) \in \mathbb{R}^{q+q}$ *be any fixed-point of* Γ.

(1) ***Finite-time identification of active set.*** *If the closed ball* $\bar{\mathbb{B}}_{2q}(\mathbf{x}^*, \epsilon(\mathbf{x}^*)/\|\mathbf{A}\|)$ *contains any point of the sequence of iterates* $x^{(n)}$, *then the active set* $\mathcal{A}(\mathbf{x}^*)$ *is identified after finitely many iterations, i.e*

$$\exists N_{\mathbf{x}^*} \geq 0 \; s.t \; \mathbf{x}^{(n)} \in \mathcal{A}(\mathbf{x}^*) \forall n \geq N_{\mathbf{x}^*}. \tag{7.7}$$

In particular, (7.7) *holds if* $\mathbf{x}^{(n)}$ *converges to* $\mathbf{x}^*$.

(2) ***Local Q-linear convergence.*** *If* $\epsilon(\mathbf{x}^*) > 0$ *and* $r(\Lambda'(\mathbf{x}^*)) < 1$, *then the iterates* $x^{(n)}$ *converge locally Q-linearly to* $\mathbf{x}^*$ *at the rate* $r(\Lambda'(\mathbf{x}^*))$.

(3) ***Optimal rates in the anisotropic case.*** *If* $c = 1$ *(as in anisotropic TV deconvolution) and* ν *is large, then the optimal rate of convergence rate is the cosine of the Friedrichs angle between* $\operatorname{Im} \mathbf{K}$ *and* $\operatorname{Im} \mathbf{D}_{\mathbf{x}^*} \simeq \mathcal{A}_1(\mathbf{z}^*)$. *If in addition* $K = \mathbf{I}$ *(as in synthesis inverse problems like the Lasso, sparse Spike-deconvolution, etc.), then the whole algorithm converges in a finite number of iterations.*

Proof. See our ICASSP paper [Dohmatob et al., 2015]. □

7.3 Relation to prior work

Recently, there have been a number of results on the local linear convergence of ADMM on particular classes of problems. Below, we outline the corresponding major works.

7.3.1 Ridge, QP, and nonnegative Lasso

On problems like Ridge regression, quadratic programming (QP), and nonnegative Lasso, [Ghadimi et al., 2013] demonstrated local linear convergence of ADMM under certain rank conditions which are equivalent to requiring that the p.s.d matrix $\mathbf{G}_\nu$ (defined in (7.3)) be invertible. The same paper prescribed explicit formulae for optimally selecting the tuning parameter ν for ADMM on these problems. We note that these results can be recovered from our Lemma 1 and Corollary 1 as they correspond to the case where $\operatorname{prox}_{(\alpha/\nu)\varphi}$ is a linear operator. Using similar spectral arguments, [Boley, 2013] demonstrated similar local convergence results for quadratic and linear QP problems.

7.3.2 Fréchet-differentiable nonlinear systems

In the SISTA algorithm [Bayram and Selesnick, 2010], the authors linked the rate of convergence of their multi-band ISTA (refer to [Daubechies et al., 2004] and the references therein, for the original ISTA algorithm) scheme to the spectral radius of a certain Jacobian matrix related to the problem data and dependent on the fixed-point [Bayram and Selesnick, 2010, Propositions 6 and 7], provided this spectral radius is less than 1. Most importantly, the authors show [Bayram and Selesnick, 2010, Proposition 8] how their algorithm can be made as fast as possible by choosing the shrinkage parameter per sub-band to be "as large as possible". Finally, analogous to our Theorem 1*(a)*, Lemma 2 of [Bayram and Selesnick, 2010] shows that the SISTA iteration projects points sufficiently close to fixed-points onto the support of these fixed-points.

7.3.3 Partly-smooth functions and Friedrichs angles

In the recent work [Liang et al., 2014] which focuses on Douglas-Rachford/ADMM, and [Liang et al., 2015] which uses the same ideas as in [Liang et al., 2014] but with a forward-backward scheme [Combettes and Wajs, 2005], the authors consider a subclass PSS (refer to definition 2.2 of [Liang et al., 2015]) of the class of so-called partly-smooth (PS) penalties and general C^2 loss functions with Lipschitz gradient. Under nonlinear complementarity requirements analogous to the non-degeneracy assumption "$\epsilon(\mathbf{x}^*) > 0$" of Theorem 1*(b)*, and rank constraints analogous to the requirement that the Jacobian matrix $\Lambda'(\mathbf{x}^*)$ have spectral radius less than 1 (in Theorem 1*(c2)*), the authors of [Liang et al., 2014, 2015] prove finite-time activity identification and local Q-linear convergence at a rate given in terms of *Friedrichs angles*, via direct application of [Bauschke et al., 2014, Theorem 3.10]. The authors show that their arguments are valid for a broad variety of problems, for example the *anisotropic* TV penalty. Still in the framework of partly-smooth penalties, [Demanet and Zhang, 2013] showed local Q-linear convergence of the Douglas-Rachford algorithm on the Basis Pursuit problem.

Detailed comparison with [Liang et al., 2014, 2015]. The works which are most comparable to ours are [Liang et al., 2014] and [Liang et al., 2015], already presented above. Let us point out some similarities and differences between these papers and ours. First, though our constructions are entirely different from the techniques developed in [Liang et al., 2014, 2015], one notes that both approaches are ultimately rooted in the same idea, namely the work of B. Holmes [Holmes, 1973] on the smoothness of the euclidean projection onto convex sets, and other related functionals (Minkowski gauges, etc.). Indeed, Theorem 1 builds directly upon [Holmes, 1973], whilst, [Liang et al., 2015] and [Liang et al., 2014] are linked to [Holmes, 1973] via [Wright, 1993], which builds on [Fitzpatrick and Phelps, 1982], and the latter builds on [Holmes, 1973].

Second, part *(c1)* of Theorem 1 (finite-time identification of active set) of the theorem can be recovered as a consequence of the results established in [Liang et al., 2014, 2015]. However, the rest of our results, notably part *(c2)* (Q-linear convergence) cannot be recovered from the aforementioned

works, at least on models like isotropic TV-L1, Sparse Variation, etc., since these models are not PSS. Indeed, the convergence rates in [Liang et al., 2014, 2015] do not extend from anisotropic to isotropic TV, for example. Success in the former case is due to the fact that the anisotropic TV seminorm is polyhedral and therefore is of class PSS at each point. By contrast, our framework can handle isotropic TV and similar "entangled" penalty types like isotropic TV-L1, Sparse Variation, etc., but suffers complementary limitations; for example, we were unable to generalize it beyond the squared-loss setting and we can only handle penalties which are a composition of a $\ell_{2,1}$ mixed-norm (or a concatenation of such) and a linear operator. The recent work [Vaiter et al., 2016] on counting the degrees of freedom of general partly-smooth penalties is worth mentioning and may contain some key ideas to help bridge the "isotropicity gap " in the methods developed in [Liang et al., 2014, 2015], concerning rates of convergence.

Lastly, the convergence rates in [Liang et al., 2014, 2015] are tight and given in terms of Friedrichs angles [Bauschke et al., 2014], whilst our rates are given in terms of spectral radii, and will be suboptimal in certain cases. An exception are the anisotropic cases, where we proved in part *(c3)* of Theorem 1 that we recover the optimal rates obtained in [Liang et al., 2014, 2015] in terms of Friedrichs angles. Moreover, for the Lasso, we showed that the whole algorithm converges after only finitely many iterations.

7.4 Numerical experiments and results

Here, we present results for a variety of experiments. Each experiment is an instance of problem (7.1) with an appropriate choice of the linear operators $\mathbf{X}$, $\mathbf{K}$, and the penalty function φ which can be the ℓ_1-norm the $\ell_{2,1}$ mixed-norm, or a mixture of the two (as in TV-L1).

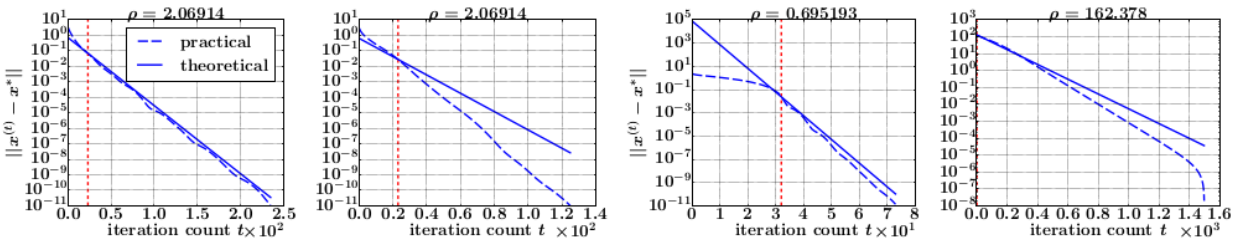
Setting. We use a grid of 20 values of ν, evenly spaced in log-space from 10^{-3} to 10^{6}. For each problem model (see below), the iteration process (7.3) is started with $\mathbf{x}^{(0)} = 0 \in \mathbb{R}^{q\times q}$, and iterated $N = 1500$ times. The final point $\mathbf{x}^{(N)}$ is approximately a fixed-point $\mathbf{x}^{(*)}$ of the operator Λ_ν. Now, the iteration process is run again (starting with the same initial $\mathbf{x}^{(0)}$) and the distance $\|\mathbf{x}^{(k)} - \mathbf{x}^{(N)}\|$ is recorded on each iteration k, producing a curve. This procedure is run for each value of ν from the aforementioned grid. Except otherwise stated, the n rows of design matrix X where drawn from a p-dimensional standard Gaussian. The measurements variable $\mathbf{y}$ is then computed as $\mathbf{y} = \mathbf{X}\mathbf{w}_0 + \text{noise}$, where $\mathbf{w}_0$ is the true signal.

Simple models. As discussed in section 7.3, the local Q-linear convergence of ADMM on a variety of particular problems has been studied in the literature (for example [Ghadimi et al., 2013, Nishihara et al., 2015, Liang et al., 2014, 2015]). We validated empirically our linear convergence results (Theorem 1) by reproducing experiments from [Liang et al., 2014, 2015]. For each of these experiments the regularization parameter α was set to 1. Viz,

(a) Lasso: Here the problem is an instance of (7.1) with $\mathbf{K} = \mathbf{I}$ and $\varphi = \|\cdot\|_1$; $n = 32$, $q = p = 128$, and w_0 is 8-sparse.

(b) Group-Lasso: Here $\mathbf{K} = \mathbf{I}$ and $\varphi = \|\cdot\|_{2,1}$, $n = 48$, $p = 128$, number of blocks $d = 32$, block size = $c = 4$, $q = d \times c = 128$, w_0 is has 2 non-zero blocks.

(c) Sparse spikes deconvolution: Here, $\mathbf{K} = \mathbf{I}$, $\mathbf{X}$ is a projector onto low Fourier frequencies (Dirichlet kernel) and the penalty φ is the ℓ_1-norm; $n = p = 200$ (with $\operatorname{rank} \mathbf{X} = 40$). The true signal w_0 is a 20-sparse vector (of length p), containing randomly distributed spikes with Gaussian values at a minimum pairwise distance of 5.



(a) Lasso (b) Group-Lasso (c) Sparse spike deconv. (d) Sparse Variation

Figure 7.2: Experimental results from ICASSP paper [Dohmatob et al., 2015]. showing local Q-linear convergence for ADMM on problem (7.1). The "theoretical" line is the exponential curve $t \mapsto \|\mathbf{x}^{(0)} - \mathbf{x}^*\| r(\Lambda'(\mathbf{x}^*))^t$. The red broken vertical line marks the instant the support of the fixed-point $\mathbf{x}^*$ is identified.

7.5 Concluding remarks

We have derived a fixed-point iteration which is equivalent to the ADMM iterates for a broad class of penalized regression problems (7.1). Exploiting the formulation so obtained, we have established detailed qualitative properties of the algorithm around solution points (Theorem 1). Most importantly, under mild conditions, local Q-linear convergence is guaranteed and we have provided an explicit formula for this rate of convergence. Finally, Theorem 1 –implicitly– opens the possibility of speeding up the ADMM algorithm on problem (7.1) by selecting the tuning parameter ν so as to minimize the spectral radius (an inverted mexican-hat-shaped curve, as ν varies from 0 to $+\infty$) of the Jacobian matrix $T'_\nu(\mathbf{x}^*)$.

Bibliography

Luca Baldassarre, Janaina Mourao-Miranda, and Massimiliano Pontil. Structured sparsity models for brain decoding from fMRI data. In *PRNI*, 2012.

Heinz H Bauschke, JY Cruz, Tran TA Nghia, Hung M Phan, and Xianfu Wang. Optimal rates of convergence of matrices with applications. *arXiv:1407.0671*, 2014.

Ilker Bayram and Ivan W Selesnick. A Subband Adaptive Iterative Shrinkage/Thresholding Algorithm. *Signal Processing, IEEE Transactions on*, 58, 2010.

A. Beck and M. Teboulle. A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems. *SIAM Journal on Imaging Sciences*, 2, 2009.

Daniel Boley. Local Linear Convergence of the Alternating Direction Method of Multipliers on Quadratic or Linear Programs. *SIAM Journal on Optimization*, 23, 2013.

Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. *Foundations and Trends in Machine Learning*, 3, 2011.

Patrick L Combettes and Valérie R Wajs. Signal recovery by proximal forward-backward splitting. *Multiscale Modeling & Simulation*, 4, 2005.

I. Daubechies, M. Defrise, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Comm. Pure Appl. Math.*, 57, 2004.

Laurent Demanet and Xiangxiong Zhang. Eventual linear convergence of the Douglas-Rachford iteration for basis pursuit. *CoRR*, abs/1301.0542, 2013.

Elvis Dohmatob, Michael Eickenberg, Bertrand Thirion, and Gael Varoquaux. Local Q-Linear Convergence and Finite-time Active Set Identification of ADMM on a Class of Penalized Regression Problems. In *ICASSP 2016*, 2015.

Jonathan Eckstein and Dimitri P Bertsekas. On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming*, 55, 1992.

Michael Eickenberg, Elvis Dohmatob, Bertrand Thirion, and Gaël Varoquaux. Total Variation meets Sparsity: statistical learning with segmenting penalties. In *MICCAI*. 2015.

S. Fitzpatrick and R. R. Phelps. Differentiability of the metric projection in Hilbert space. *Trans. Am. Math. Soc.*, 270, 1982.

Daniel Gabay and Bertrand Mercier. A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Computers & Mathematics with Applications*, 2, 1976.

Euhanna Ghadimi, André Teixeira, Iman Shames, and Mikael Johansson. Optimal parameter selection for the alternating direction method of multipliers (ADMM): Quadratic problems. *arXiv preprint arXiv:1306.2454*, 2013.

Roland Glowinski and A Marroco. Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problemes de Dirichlet non linéaires. *ESAIM: Mathematical Modelling and Numerical Analysis-Modélisation Mathématique et Analyse Numérique*, 9, 1975.

Alexandre Gramfort, Bertrand Thirion, and Gaël Varoquaux. Identifying predictive regions from fMRI with TV-L1 prior. In *PRNI*, 2013.

Richard B. Holmes. Smoothness of certain metric projections of Hilbert space. *Trans. Amer. Math. Soc.*, 184, 1973.

Jingwei Liang, Jalal Fadili, Gabriel Peyré, and Russell Luke. Activity Identification and Local Linear Convergence of Douglas–Rachford/ADMM under Partial Smoothness. *arXiv:1412.6858*, 2014.

Jingwei Liang, Jalal Fadili, and Gabriel Peyré. Activity Identification and Local Linear Convergence of Inertial Forward-Backward Splitting. *arXiv:1503.03703*, 2015.

Robert Nishihara, Laurent Lessard, Benjamin Recht, Andrew Packard, and Michael I Jordan. A General Analysis of the Convergence of ADMM. *arXiv:1502.02009*, 2015.

R. Piziak, P.L. Odell, and R. Hahn. Constructing projections on sums and intersections. *Pergamon, Computers and Mathematics with Applications*, 1999.

Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society Series B*, 2005.

Samuel Vaiter, Charles Deledalle, Jalal Fadili, Gabriel Peyré, and Charles Dossal. The degrees of freedom of partly smooth regularizers. *Annals of the Institute of Statistical Mathematics*, pages 1–42, 2016.

S.J Wright. Identifiable surfaces in constrained optimization. *SIAM Journal on Control and Optimization*, 31, 1993.

Part

III – Functional inter-subject variability

Direct EPI-to-EPI inter-subject non-linear registration

Contents

In this chapter, we turn our attention to another key problem in neuroimaging: registration of functional brain data from different subjects. After a concise review of the existing literature, we present a contribution of ours, namely direct registration of fMRI data without using the subject's anatomy as a proxy.

8.1 Introduction

Registering brain images from different subjects in a common space (for example, the MNI space [Collins et al., 1994, Mazziotta et al., 1995]), is an essential step in any multi-subject analysis pipeline [Friston et al., 1995]. Indeed, a voxel-to-voxel correspondence across subjects is needed for group-level statistics on brain maps to make sense. In addition, the use of a standard space opens the possibility to share results in a consistent fashion,

hence the comparison of experiments and meta-analysis [Wager et al., 2007, Gorgolewski et al., 2015]. This is especially true in fMRI (functional Magnetic Resonance Imaging) studies in which the activations might span just a few voxels in diameter.

Traditional indirect T1-based techniques for inter-subject registration of EPI data assume that the mismatch between a subject's T1 (i.e anatomical) image and associated EPI scan is only affine, i.e. it includes only pose differences related to slice orientation and field of view selection. such methods exploit the fact that learning a deformation from the subject's T1 to a template is easier, due to the relatively high anatomical contrast in T1 images, than learning a deformation from the subject's EPI image to the template. Thus the EPI and the T1 are affinely aligned in a primary step called *coregistration*, then one applies the transformation T1 $\rightarrow$ template to the EPI images to warp them from subject to template space.

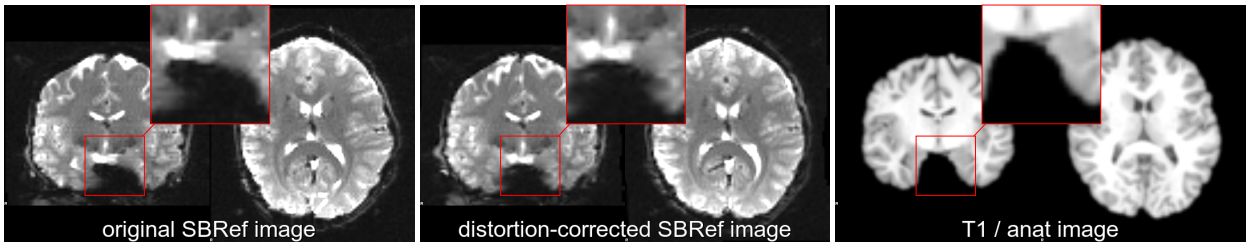


Figure 8.1: **Nonlinear mismatch** between EPI and T1-weighted image of the same subject of the HCP dataset [van Essen et al., 2012], before and after distortion-correction. **Left:** Single-band high-resolution EPI (SBRef) image of the same subject. Notice the large distortions along the Left-Right direction (inside the highlighted patches). **Center:** Distortion-corrected single-band EPI image. Here, the distortion-correction managed to undo most –but not all– of the distortions. Even after distortion correction, there are minor shape (nonlinear) differences between the EPI and the T1-weighted image of the subject (**Right**). The same native-space coordinates where used in all of the 3 plots.

A crucial assumption in these classical methods is that the T1 and EPI images of the same subject can be properly aligned to one another via an affine transformation. Thus one assumes, for example, that distortion correction is good enough that the EPI image can be realigned to the T1 with a rigid or affine transformation. However, high-resolution EPI sequences deviate from the aforementioned underlying assumptions of classical T1-based indirect inter-subject registration methods, namely the EPI sequences suffer from distortions that push them nonlinearly out-of-match relative to the T1-weighted image of the same subject.

As an example, Fig. 8.1 illustrates this issue (distortions) on the HCP (Human Connectome Project) [van Essen et al., 2012] dataset, a reference dataset that contains high-quality EPI data acquired using state-of-the-art sequences, yet with severe distortions [Wan et al., 1997a, Mangin et al., 2002, Zeng and Constable, 2002, Andersson et al., 2003]. Indeed, as discussed in the literature (e.g [Freire et al., 2002]), EPI distortions and signal loss related to B0 inhomogeneities cannot be separated with registration based techniques, which are compensatory operations. Consequently, the set up of efficient distortion correction method in EPI-based imaging is an open question. Moreover, sophisticated anatomy-based methods like Freesurfer's recon-all cannot scale to huge data sets like the 5,000 participants of the initial release of the UKBioBank dataset [Miller, 2016], due the long computation time that renders such approaches impractical.

The goal of this paper is to provide experimental evidence that the indirect T1-based inter-subject EPI registration explained above is no longer needed, if not sub-optimal in such settings. We also provide a computationally cheap pipeline based on publicly available tools, which bypasses the

need for a T1-weighted image, and do direct inter-subject registration of the EPI images.

8.2 Methods

8.2.1 An important note on normalization

Let us begin by stressing that the *normalization* problem (i.e registration to a standard template) is not addressed in our work. We concentrate on inter-subject (nonlinear) *registration*, since our goal is to show the benefits of using EPI images in place of anatomical images in pipelines. We also note that there is an increasing concern in the literature that in the future, normalization will be based on multi-modal atlases (tissue probability maps, functional parcellation maps, etc.) [Amunts et al., 2014].

8.2.2 General preprocessing procedures

Motion correction During acquisitions, subjects move their heads in the scanner. This head movement induces an approximately affine mismatch between different volumes acquired in the same acquisition run. Motion correction is done to remove this source of intra-subject variability. We used FSL's *flirt* tool [Smith et al., 2004] for motion correction.

Distortion correction Due to inhomogeneities in the ambient B0 field, the EPI images are distorted (i.e artifactually warped) along the *phase-encoding directions* (Left-Right / Right-Left in the case of HCP dataset [van Essen et al., 2012]). See Figure 8.1. In our experiments, distortion correction [Wan et al., 1997a, Mangin et al., 2002, Zeng and Constable, 2002, Andersson et al., 2003, Jezzard and Balaban, 1995, Wan et al., 1997b] was achieved using the methods described in [van Essen et al., 2012]. Both methods use FSL's *topup* tool [Smith et al., 2004] to estimate the deformation field due to B0 inhomogeneities (the distortions) [Glasser et al., 2013].

Deformation model We used ANTs' *Symmetric Normalization* (aka *SyN*) deformation model [Avants et al., 2008, 2011], which has been shown to be a state-of-the art method for nonlinear registration [Klein et al., 2009]. As done usually, we initialize a nonlinear registration algorithm with an affine (rigid-body) registration algorithm. The former is simply meant to estimate an alignment for the bounding boxes of the images (thus ensuring a sufficiently large region of overlap). Concretely, we stack a 2-level pyramidal[1] affine transformation model (as initialization) with a 3-level pyramidal SyN deformation model. *Mattes mutual information* [Mattes et al., 2003] is used as the loss function.

[1] *Pyramidal* means multiple passes are made by a registration algorithm on the input images, with finer and finer resolution (aka *pyramid*). In this speedup technique, each pass of the pyramid is initialized with the solution of the previous pass (this is known as *warmstarting*).

8.2.3 The pipelines

We now present constructions for the pipelines whose benchmark is the core of this work. All registration pipelines presented here were scripted in using command-line tools from FSL version 5.0 [Smith et al., 2004] for affine registration, distortion correction, motion correction, ANTs [Avants

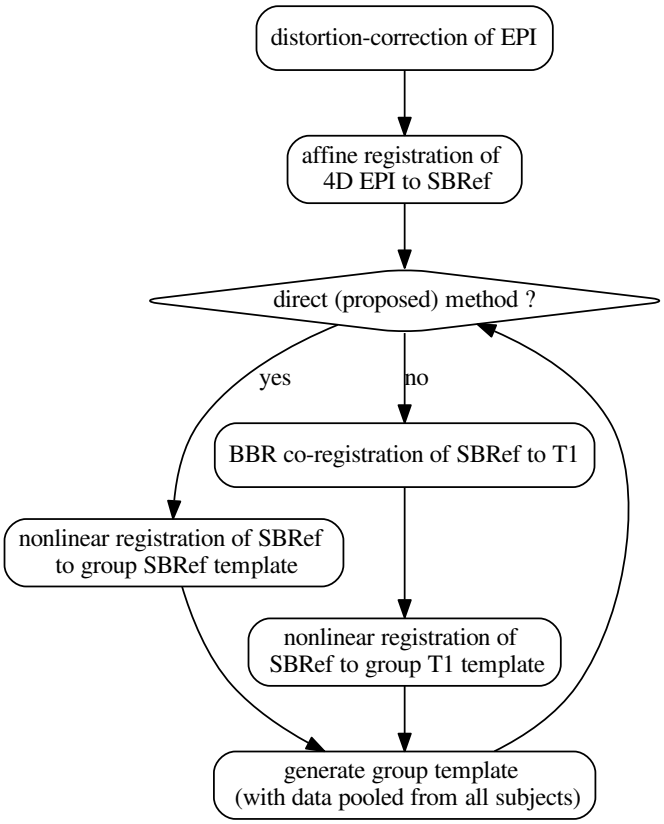


Figure 8.2: **The pipelines.** The template-generation step is done using ANTs [Avants et al., 2008, 2011]. It pools registered data from all subjects. **N.B.:** SBRef = single-band high-resolution 3D volume. As in [Glasser et al., 2013], all the transformations are postponed and the original 4D EPI is resampled at the end by applying the composition of these transformations in a single step.

et al., 2009] antsRegistration, antsApplyTranforms, and some custom scripts (for distortion correction) from the HCP scripts described in [Glasser et al., 2013], hosted on Github. Except stated otherwise, all affine registrations (motion correction, coregistration) were performed using FSL's *flirt* tool [Smith et al., 2004] with Normalized Mutual Information as the cost function (option: *-cost normmi*).

Classical indirect T1-based method

The classical indirect T1-based pipeline for registration of EPI images can be schematized as follows[2]:

$$\text{EPI} \rightarrow \text{templ.} = \underbrace{(\text{T1} \rightarrow \text{templ.})}_{\text{nonlinear}} \circ \underbrace{(\text{EPI}_0 \overset{\text{BBR}}{\rightarrow} \text{T1})}_{\text{linear}} \circ \text{DistCorr} \tag{8.1}$$

[2] The "$\circ$" symbol denotes composition of transformations.

in which a deformation of the subject's T1 is estimated and this deformation is then used to warp the same subject's EPI data. We implemented the pipeline as follows. Here, EPI_0 is any single-volume EPI image previously-coregistered with the 4D EPI sequence. Typical choices include: the middle volume of the film or the mean volume after motion correction. In our implementations, we used the former.

For the template, a subject is fixed and its T1-weighted image is used as the template. For each other subject, *(a)* distortion correction is used to learn a nonlinear undistorting warpfield, in a procedure already described in subsection 8.2.2 above. Then, *(b)* motion correction is done to realign the the subject's EPI data to the mean thereof. The subject's T1-weighted image is then aligned to this mean EPI image via coregistration (an affine transformation). We use BBR (boundary-based registration) [Greve and Fischl, 2009] for this coregistration step, for optimal results and fair comparison. BBR is a state-of-the-art functional-to-structural registration method driven by intensity difference across boundary (samples). It uses white-matter boundaries (via T1w segmentation). BBR need good structural images (with little contrast bias), and some anatomical contrast in the EPI image (which is the case of the single-band high-resolution reference images in the HCP dataset [van Essen et al., 2012]). The implementation we use is *epi_reg* script of FSL [Smith et al., 2004]. However, since BBR is an affine correction method, it still suffers from the limitations explained in the introductory section. In particular, it is not resiliant to distortions in the input EPI image.

(c) ANTs is used to learn a deformation of the T1 image to the template (which is a fixed subject). This produces a warped version of the T1-weighted image, together with the corresponding deformation (and its inverse too), for passing from the subject's space to the template space. Finally, *(d)* the deformation above (T1-based), and all the other postponed warpfields, affine transformations, etc., are then applied (in respective order) to all EPI data previously aligned (rigidly, via coregistration) with the T1-weighted image of the subject; these may include EPI images acquired on the same subject during another task, for instance. This one-step resampling procedure (see subsection 8.2.2) then produces a registered, motion-corrected, undistorted version of the input EPI data.

Then mean of all the registered T1-weighted images is computed, and becomes the template henceforth. This procedure is iterated a couple of

times.

Our proposed *direct* EPI-based non-linear inter-subject registration method

Our proposed pipeline operates just as the classical indirect T1-based pipeline described above in 8.2.3, except that the anatomical image is replaced with the single-band high-resolution EPI (the SBRef) image, which has more tissue contrast than the any volume of the 4D EPI film being registered [Glasser et al., 2013], and also does not suffer from multi-band artifacts. The anatomical image is not used anywhere in this pipeline. The pipeline can be schematized as follows:

$$\text{EPI} \rightarrow \text{templ.} = \underbrace{(\text{EPI}_0 \rightarrow \text{templ.})}_{\text{nonlinear}} \circ \text{DistCorr}, \tag{8.2}$$

where we take EPI_0 = Single-band high-resolution (SBRef) EPI image.

A note on image interpolation (resampling) To avoid degrading the images as they travel through a pipeline, we stack all intermediate transformations and postpone the resampling operations to the end of the pipeline. The transformations are then concatenated (i.e composed), and applied to the input image in a one-step resampling procedure based on the *ApplyTransforms* tool of the ANTs software [Avants et al., 2008, 2009]. For example, affine transformations estimated during the motion correction step are converted to warpfields using FSL's *convertwarp* tool [Smith et al., 2004]. FSL's *applywarp* tool [Smith et al., 2004] is then used to jointly apply this affine transformation warpfields and the warpfields corresponding to the deformations estimated by *topup* [Smith et al., 2004], which are then stacked with subsequent transformations. We use this strategy in both pipelines.

8.3 Relation to previous works

8.3.1 Direct EPI-to-EPI non-linear inter-subject registration

The idea of EPI-to-EPI registration has already been suggested in the literature. For example, the method in [Grabner et al., 2014] used high-resolution EPI (1.1mm isotropic) data for different subjects acquired at 7T to iteratively build a sequence of EPI-based study-specific templates of increasing quality / resolution [Grabner et al., 2006]. The finest of these templates shows a great deal of anatomical detail. Group-level activation patterns for a finger-tapping task were also shown to be very accurately localized on the posterior bank of the central sulcus. The authors concluded that high-resolution (7T) EPI images contain enough anatomical information for inter-subject registration, and so one can effectively by-pass the anatomical image of subjects in pipeline. This would for example allow one to avoid the classical coregistration step used to align the subject's EPI images to their anatomy. Our experiments confirm and extend the findings of [Grabner et al., 2014], but at an even lower resolution: 2mm resolution, obtained from 3T MRI, and

on a much larger dataset. Indeed, using a much larger bail of 110 subjects, from Human Connectome Project (HCP) dataset [van Essen et al., 2012], and a variety of different task contrasts, we show that registration with our pipeline increases the pairwise NMI of subjects, over the classical pipeline; crucially, this leads to a decrease in residual post-registration inter-subject misalignement.

In comparison, the pipeline we propose (refer to 8.2.3) is much lighter computationally as we bypass the potentially expensive and challenging step of generating a good template from EPI data [Grabner et al., 2006]. Of course, this economy is more of a compromise between complexity and accuracy, and might be potential limitation of our contribution. Finally, we note that the work in [Grabner et al., 2014] did not consider the distortion problem which are severe even at 3T [Andersson et al., 2003], as it is the case with the HCP data.

8.3.2 Non-linear EPI-to-structural coregistration

A recent work [Wang et al., 2017] has considered the possibility of replacing the classical linear EPI-to-structural coregistration step with a non-linear counterpart, and then running a non-linear structural-to-template registration as usual. They show that their method outperforms the method based on distortion correction and linear EPI-to-structural coregistration followed by structural-to-template registration as usual (see 8.2.3). In contrast, our proposed method (refer to 8.2.3) does not use the structural image at all.

8.4 Experiments

We now describe benchmarks done to compare the pipelines presented in this paper (subsection 8.2.3) on the task fMRI data of 110 subjects from the HCP dataset [van Essen et al., 2012]. The task fMRI data were acquired in an attempt to assess major domains that sample the diversity of neural systems, including: 1) visual, motion, somatosensory, and motor systems; 2) language processing (semantic and phonological processing); 3) social cognition (Theory of Mind); and 4) emotion processing. Due to time constraints, our benchmarks were run only on these 4 (out of a total of 7) tasks (i.e protocols). Also, only data for LR (left-right) phase-encoding direction [Chang and Fitzpatrick, 1992] runs were used. In all the non T1-based pipelines, the single-band high-resolution (SBRef) image of the motor task was used to learn deformations of the subject's brain into template space (a fixed subject of the same dataset).

The estimated deformations were then applied to warp EPI data (previously coregistered to same subject's motor SBRef) acquired on the same subject during different task conditions, into template space. GLMs (General Linear Models) [Friston et al., 1994] were run using *nipy* [Gorgolewski et al., 2011], open-source Python library for analysis of neuro-imaging data. For the purpose of reporting the results, the resulting maps were co-registered to MNI space a posteriori.

8.4.1 Evaluation metrics

The pipelines were evaluated using the following qualitative and quantitative metrics.

Normalized mutual information evaluation (NMI)

NMI (see Table 1 for definition) is a popular similarity metric used to assess the quality of registration between two images, i.e how well aligned the images are to one another (for example [Maes et al., 1997]). It is also the loss function minimized by many optimization algorithms in image registration. A detailed overview of the use of the NMI metric in medical image registration can be found in [Pluim et al., 2003]. In our experiments, FSL's *flirt* [3] tool [Smith et al., 2004] was used to compute NMI.

[3] With the "*-schedule*" option.

Inter-subject residual variance

In a good registration method, the residual subject-to-subject variance of the EPI image should be reduced. The aim of inter-subject registration is indeed to put subjects into spatial correspondence to facilitate later group analysis. To measure the quality of the different registration methods in this regards, we computed the coefficient of variation (CoV) across the different subjects after registration. This is defined by

$$\text{CoV} = \frac{\text{variance image across subjects}}{\text{mean image across subjects}}. \tag{8.3}$$

High values in this 3D image would outline regions of the brain which are not well registered across subjects.

Group-level statistics and functional brain network patterns

Finally, in a successful inter-subject registration procedure, we expect the functional activation patterns to be more localized in space and to have higher peaks. Or could this effect be masked by inter-subject variability in activation magnitude ? This will be discussed in detail in the discussion section 8.6.

8.4.2 How many (plausible) pipelines are there ?

It is worth noting that there are potentially hundreds of pipelines which could have considered for testing: should we do distortion correction ? And if yes, how ? Should we use linear or nonlinear model for the deformation field ? What degree should we use for the interpolating splines ? In fact as noted in [Poldrack et al., 2016], there are exponentially many pipelines that can be considered, based on the answers to the above choices. Of course some of these parameters have rule-of-thumb default values (for example, there is no doubt distortion correction is a good thing to do), but others are open to preferential choice. Thus our goal is not to consider all possible pipelines, but to look at a more focal question: does direct EPI-based inter-registration outperform the traditional indirect T1-based pipeline ?

8.5 Results

We now present results of experiments performed on the task fMRI protocols of the HCP dataset [van Essen et al., 2012]. Refer to section 8.4 for detailed information about the experiments we did. The different pipelines discussed in section 8.2.3 were used to register the data (inter-subject registration), and the quality of the registration was benchmarked using the different evaluation metrics discussed in Section 8.4.

Normalized Mutual Information (NMI)

The results comparing across-subject NMI for the pipelines are presented in Figure 8.3. We see that MNI is in most cases higher through our approach, which implies that our proposed direct EPI-based pipeline mildly outperforms the classical indirect T1-based pipeline.

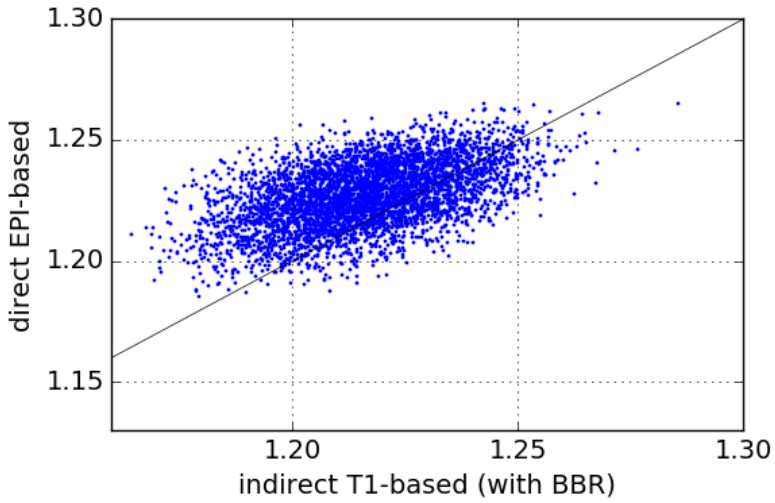


Figure 8.3: Normalized Mutual Information – NMI (higher values are better). Each point (x, y) on the plots such that x is the NMI of a given pair of subjects registered using the pipeline on the abscissa and y is the NMI of the same pair of subjects registered using the pipeline on the ordinate. From the one-sided We see that our proposed direct EPI-based pipeline significantly outperforms the classical indirect T1-based pipeline.

Residual inter-subject spatial variability

In Figure 8.4, we show across-subject histograms of across-subject per-voxel Coefficient of Variation (small is better). We see that our proposed direct method outperforms the classical indirect T1-based method, as the former leads to relatively more mis-aligned voxels across subjects, most concentrated on the outer edge of the cortex (see Figure 8.4 *(a)*).

Quality of estimated EPI group template

To compare the quality of the group template produced by either pipeline, a snapshot of the resulting mean image or template is displayed in Figure 8.5. Compared to the our proposed direct method, the mean image (across all subjects) from the indirect T1-based pipeline is blurry and has "ripples" on the cortical surface, indicative of residual mismatch between subjects after registration. The across-subject mean image post-registration with our direct EPI-based pipeline is the sharpest, showing that the subjects have

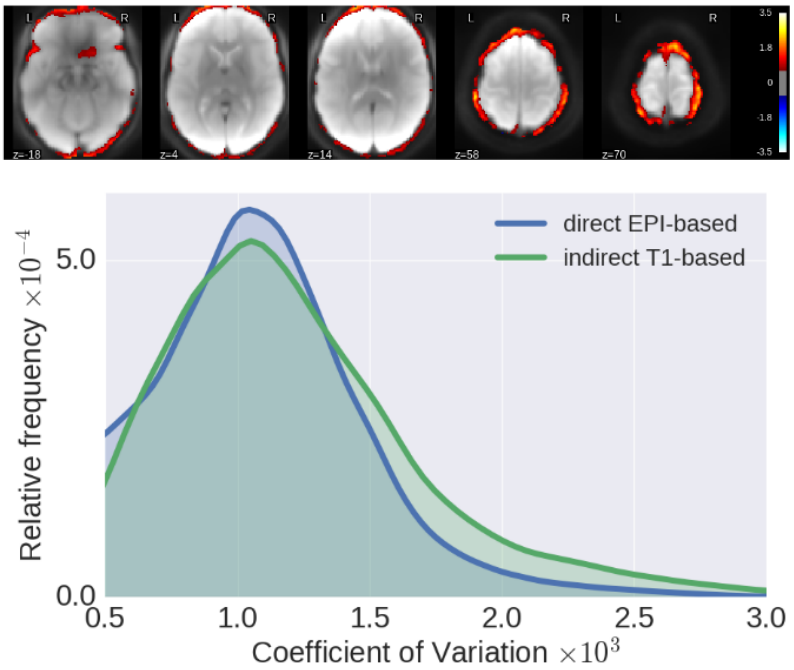


Figure 8.4: **Coefficient-of-Variation (CoV)** after registration. **Top:** Log10 of ratio of across-subject Coefficient of Variation (CoV) for indirect T1-based pipeline / direct EPI-based. We see that the gain of our proposed method is most pronounced along the cortical surface. **Bottom:** Histograms of CoV for both pipelines. Again, we see clearly that our proposed method reduces the inter-subject variability by a much larger margin, indicative of improved subject-to-subject alignment.

been matched extremely well. Also, one notices that the mean image from the indirect T1-based pipeline still has some residual distortion (here in the left-to-right direction), even though distortion correction was done as part of both pipelines.

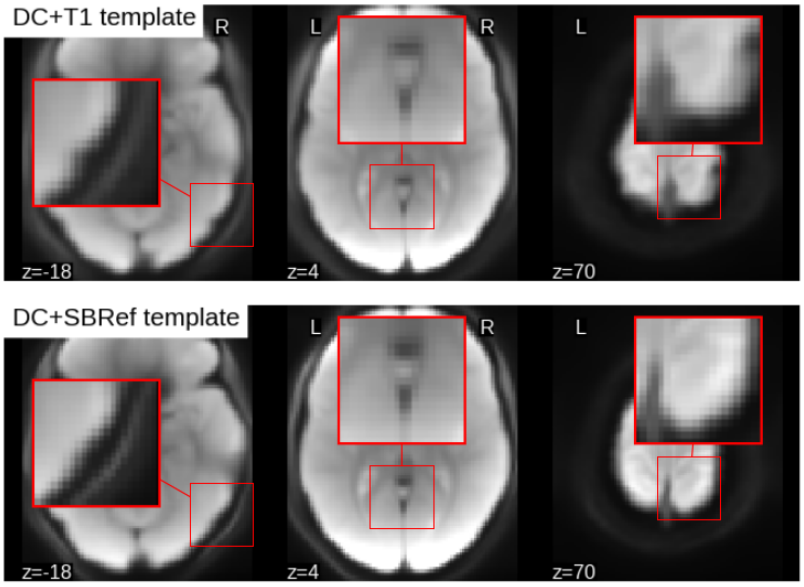


Figure 8.5: Mean EPI image across all subjects after registration (aka estimated population templates). Patches on the images have been zoomed to highlight details. The mean image from the indirect T1-based pipeline (**Left**) is more blurry (as seen here in the cerebellum), compared to our direct EPI-based pipeline post-registration across-subject mean image (**Right**) which is much sharper, indicative of a better inter-subject registration. Also the mean image from the T1-based pipeline has ripples on the cortical surface indicative of residual registration problems, which can be attributed residual EPI-distortions that could not be captured via coregistration.

Group-level statistics and Functional brain network patterns

As regards group-level GLM scores, we see from Figure 8.6 that our proposed method does just as good as the classical indirect T1-based pipeline.

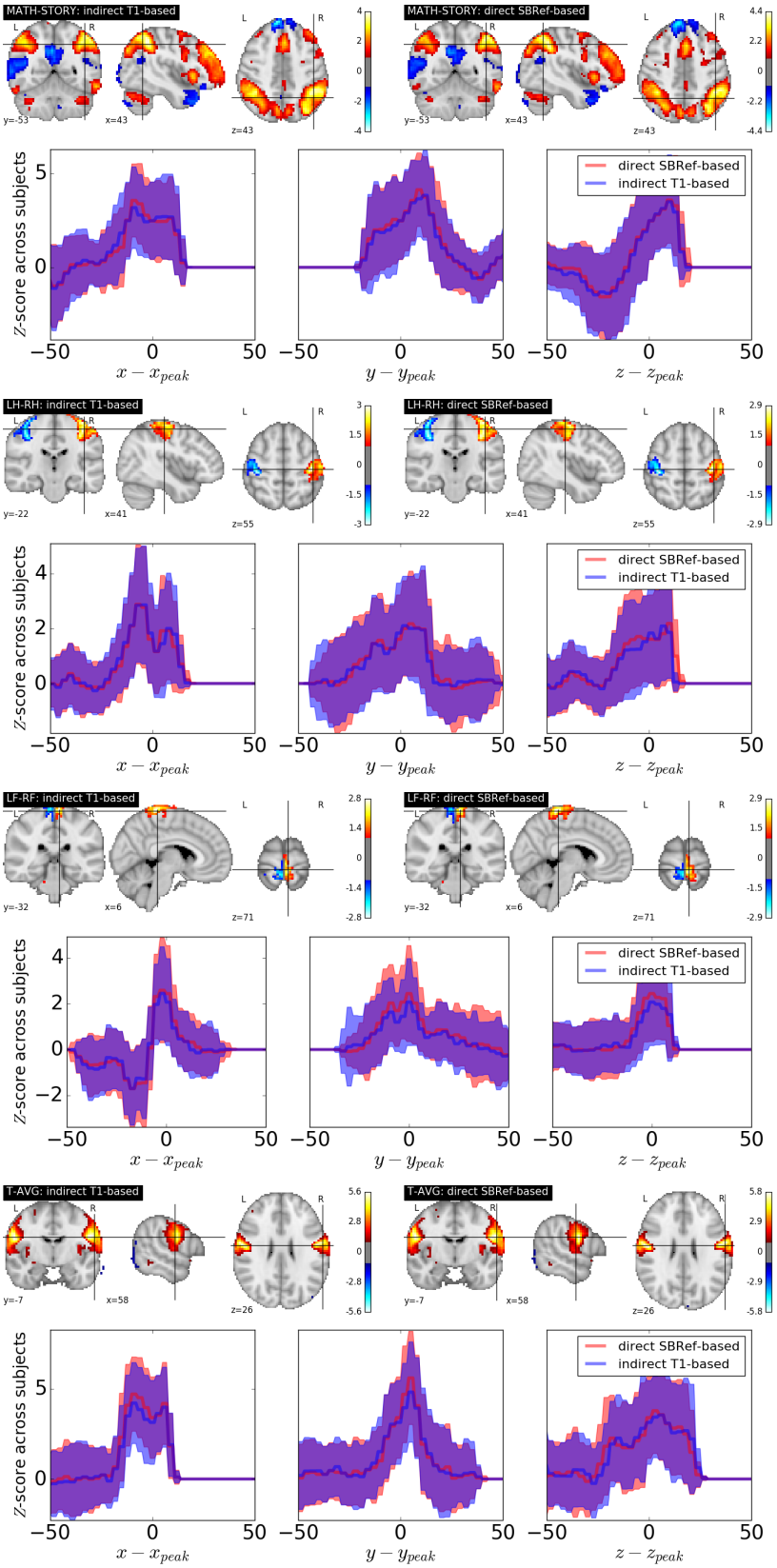


Figure 8.6: **Qualitative comparison** of pipelines via GLM results. Across-subject mean activation maps of Z-scores for different contrasts. Here we see that our proposed direct EPI-based registration scheme leads to slightly higher across-subject mean activation peaks. For each contrast, a cut has been made around the location of the activation peak, to display curves of the activation profile and across-subject variability thereof, in a neighborhood of this peak location.

This is remarkable, as the former pipeline does not use any anatomical data. However, as noted in [Thirion et al., 2007, Thyreau et al., 2012], the inter-subject variability in GLM results is not due to misregistration, but intrinsic subject differences with a more physiological nature: the response of subjects to the same stimulus / task is modulated differently, and is more dependent on effect size fluctuations than position.

This is confirmed in the curves in Figure 8.6, where we can see that the spatial across-subject activation profiles are very similar between the compared registration methods, except for the already noted slight improvement of the peak mean activation pattern obtained by our proposed method.

Finally, Figure 8.7 comparing the functional brain networks obtained by running ICA on the images registered with each of the pipelines, shows essentially the same network patterns. The absence of a difference between these maps can be explained by the fact that resting state networks are less focal than task-based activation-patterns, and so the former are less sensitive to the quality of the underlying registration procedure.

8.6 Discussion and concluding remarks

Classical inter-subject registration pipelines use the T1-weighted (anatomical) image of a subject to estimate the subject-to-template warp. An obvious issue is that high-quality T1-weighted images are not always available, but more generally, it is not always possible to completely align the EPI images of a subject to their T1-weighted image via coregistration. Added to this is the possibility that such an intermediate registration step is a potential source of interpolation artifacts, not to mention the added computational cost (which may exceed the rest of the computation time by many orders of magnitude, for example, in the case of surface-based methods). As shown by our experiments on the HCP dataset [van Essen et al., 2012] (Figure 8.1), this is for example the case in the presence of distortions [Wan et al., 1997a, Mangin et al., 2002, Zeng and Constable, 2002, Andersson et al., 2003] that persist even after correction. Further, as noted in [Yamada et al., 2014], distortions cannot be separated with registration based techniques, which are compensatory operations. Consequently, the efficient distortion correction method in EPI data remains an open question. Our work proposes a direct EPI-based inter-subject registration pipeline that to some extent evades these bottlenecks.

We have proposed a computationally cheap EPI-based pipeline for direct inter-subject nonlinear registration of functional data. Our method has been empirically validated on the HCP dataset [van Essen et al., 2012], where we have shown that we obtain registered subject images with less inter-subject variability. Such direct EPI-based methods should replace the well-accepted classical T1-based strategy. Results on the HCP dataset [van Essen et al., 2012] show that the proposed pipeline outperforms the classical T1-based indirect registration strategy, according to a variety of different quality metrics: Normalized Mutual Information –NMI (Figure 8.3), residual inter-subject variance (Figure 8.4), and quality of estimated group template (Figure 8.5), without compromising the quality of post-registration statistical analyses results (GLM, ICA, etc.). These results replicate the findings of

[Grabner et al., 2014] on a larger dataset (110 subjects, compared to 10 in the reference paper) and in a 3T setting.

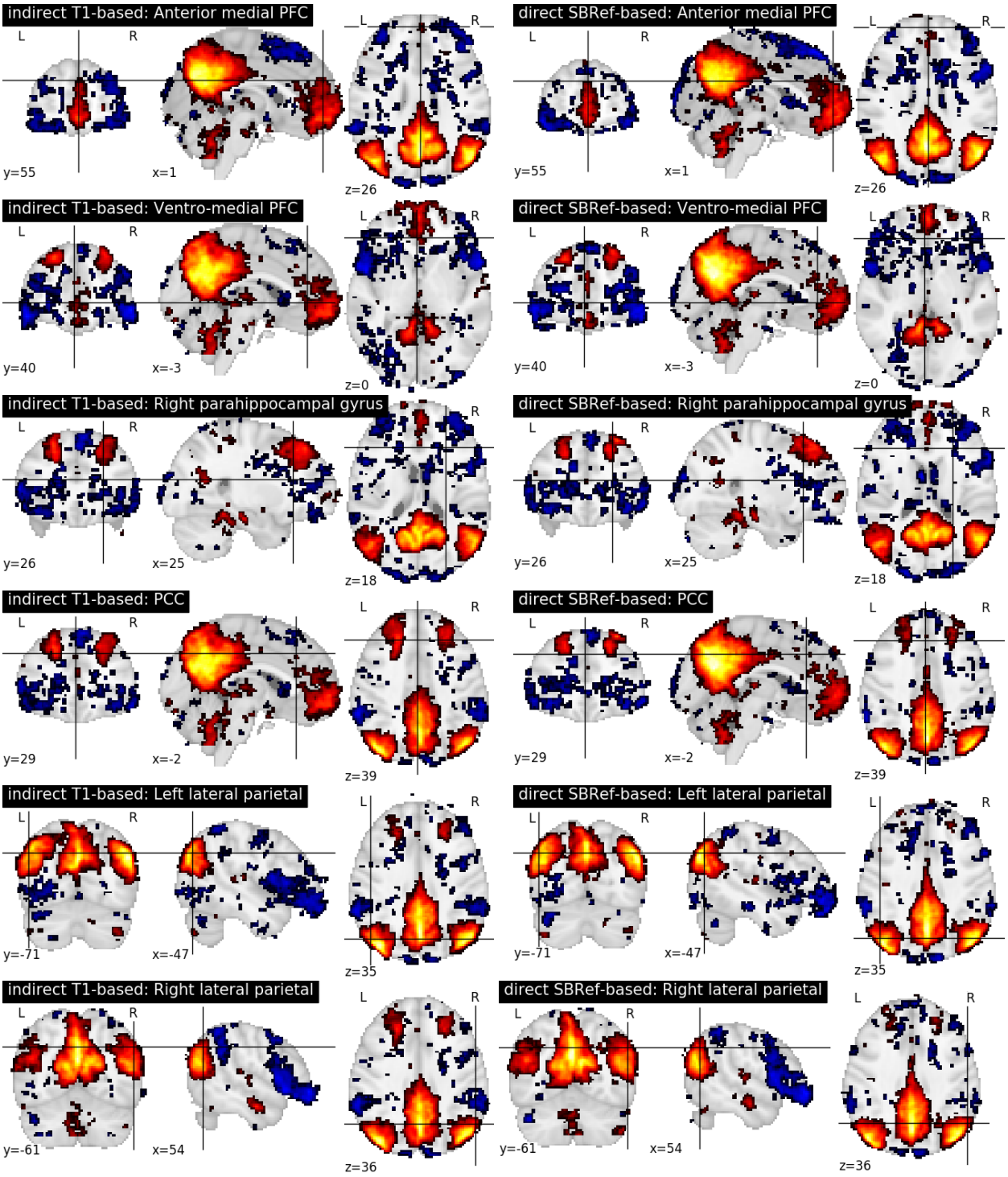


Figure 8.7: Comparing functional brain networks from subject fMRI images registered with both pipelines, namely the classical indirect T1-based method, and our proposed direct EPI-based method). Shown here are group-level unthresholded sub-component maps of the Default Mode Network (DMN) [Raichle et al., 2001], using MNI coordinates reported in Table 1 of [Watanabe et al., 2013].

Remarkably, we observed that according to low-level QA metrics like

NMI (Figure 8.3), residual inter-subject spatial variability (Figure 8.4) and the quality of across-subject mean registered EPI image (Figure 8.5), our proposes method outperforms the classical indirect T1-based registration. In terms of more high-level metrics like group-level GLM statistics, these gains though still present, are as not as pronounced (refer to Figure 8.6). Indeed, as noted in [Thirion et al., 2007, Thyreau et al., 2012, Xu et al., 2009], the inter-subject variability in GLM results is not due to misregistration, but intrinsic subject differences with a more physiological nature: the size of effects and the anatomical localization are subject-specific. In chapter 11, we show that resting-state fMRI data can be used to predict the activation maps of a subject to a task, with an R^2-score which can be up to **0.5** for some subjects and task. This is an enhancement on previous work by [Tavor et al., 2016], and shows differences in task-based brain activations are largely physiological –in contrast to being driven by subjects' brain morphological differences– and can be predicted from resting state fMRI data.

In a separate work [Dohmatob et al., 2016], also presented in chapter 9 in detail, we have considered the possibility of explicitly modeling this physiology differences by estimating latent factors of variability across-subjects in a data-driven way using dictionary-learning. The motivating idea behind such a model, is that activation across-subjects would be governed by the same generative model (the latent model), and modulated on the subject-level by subject-specific physiology.

Bibliography

K. Amunts, M. J. Hawrylycz, D. C. Van Essen, J. D. Van Horn, N. Harel, J. B. Poline, F. De Martino, J. G. Bjaalie, G. Dehaene-Lambertz, S. Dehaene, P. Valdes-Sosa, B. Thirion, K. Zilles, S. L. Hill, M. B. Abrams, P. A. Tass, W. Vanduffel, A. C. Evans, and S. B. Eickhoff. Interoperable atlases of the human brain. *Neuroimage*, 99:525–532, Oct 2014.

J. L. Andersson, S. Skare, and J. Ashburner. How to correct susceptibility distortions in spin-echo echo-planar images: application to diffusion tensor imaging. *Neuroimage*, 20(2):870–888, Oct 2003.

B. B. Avants, C. L. Epstein, M. Grossman, and J. C. Gee. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Med Image Anal*, 12(1): 26–41, Feb 2008.

B. B. Avants, N. J. Tustison, G. Song, P. A. Cook, A. Klein, and J. C. Gee. A reproducible evaluation of ANTs similarity metric performance in brain image registration. *Neuroimage*, 54(3):2033–2044, Feb 2011.

Brian B Avants, Nick Tustison, and Gang Song. Advanced normalization tools (ANTS). *Insight J*, pages 1–35, 2009.

Hsuan Chang and J Michael Fitzpatrick. A technique for accurate magnetic resonance imaging in the presence of field inhomogeneities. *IEEE transactions on medical imaging*, 11(3):319–329, 1992.

D. L. Collins, P. Neelin, T. M. Peters, and A. C. Evans. Automatic 3D intersubject registration of MR volumetric data in standardized Talairach space. *J Comput Assist Tomogr*, 18(2):192–205, 1994.

Elvis Dohmatob, Arthur Mensch, Gaël Varoquaux, and Thirion Bertrand. Learning brain regions via large-scale online structured sparse dictionary-learning. In *NIPS*, 2016.

L. Freire, A. Roche, and J. F. Mangin. What is the best similarity measure for motion correction in fMRI time series? *IEEE Trans Med Imaging*, 21 (5):470–484, May 2002.

Karl J Friston, Andrew P Holmes, Keith J Worsley, J-P Poline, Chris D Frith, and Richard SJ Frackowiak. Statistical parametric maps in functional imaging: a general linear approach. *Human brain mapping*, 2(4):189–210, 1994.

Karl. J. Friston, J. Ashburner, C. D. Frith, J.-B. Poline, J. D. Heather, and R. S. J. Frackowiak. Spatial registration and normalization of images. *Human Brain Mapping*, 3(3):165–189, 1995. ISSN 1097-0193. doi: 10.1002/hbm.460030303. URL `http://dx.doi.org/10.1002/hbm.460030303`.

M. F. Glasser, S. N. Sotiropoulos, J. A. Wilson, T. S. Coalson, B. Fischl, J. L. Andersson, J. Xu, S. Jbabdi, M. Webster, J. R. Polimeni, D. C. Van Essen, and M. Jenkinson. The minimal preprocessing pipelines for the Human Connectome Project. *Neuroimage*, 80:105–124, Oct 2013.

K. J. Gorgolewski, G. Varoquaux, G. Rivera, Y. Schwarz, S. S. Ghosh, C. Maumet, V. V. Sochat, T. E. Nichols, R. A. Poldrack, J. B. Poline, T. Yarkoni, and D. S. Margulies. NeuroVault.org: a web-based repository for collecting and sharing unthresholded statistical maps of the human brain. *Front Neuroinform*, 9:8, 2015.

Krzysztof Gorgolewski, Christopher D Burns, Cindee Madison, Dav Clark, Yaroslav O Halchenko, Michael L Waskom, and Satrajit S Ghosh. Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python. *Front Neuroinform*, 5, 08 2011. ISSN 1662-5196.

G. Grabner, A. L. Janke, M. M. Budge, D. Smith, J. Pruessner, and D. L. Collins. Symmetric atlasing and model based segmentation: an application to the hippocampus in older adults. *Med Image Comput Comput Assist Interv*, 9(Pt 2):58–66, 2006.

G. Grabner, B. A. Poser, K. Fujimoto, J. R. Polimeni, L. L. Wald, S. Trattnig, I. Toni, and M. Barth. A study-specific fMRI normalization approach that operates directly on high resolution functional EPI data at 7 Tesla. *Neuroimage*, 100:710–714, Oct 2014.

Douglas N Greve and Bruce Fischl. Accurate and robust brain image alignment using boundary-based registration. *Neuroimage*, 48(1):63–72, 2009.

Peter Jezzard and Robert S Balaban. Correction for geometric distortion in echo planar images from B0 field variations. *Magnetic resonance in medicine*, 34(1):65–73, 1995.

A. Klein, J. Andersson, B. A. Ardekani, J. Ashburner, B. Avants, M. C. Chiang, G. E. Christensen, D. L. Collins, J. Gee, P. Hellier, J. H. Song, M. Jenkinson, C. Lepage, D. Rueckert, P. Thompson, T. Vercauteren, R. P. Woods, J. J. Mann, and R. V. Parsey. Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. *Neuroimage*, 46(3): 786–802, Jul 2009.

Frederik Maes, Andre Collignon, Dirk Vandermeulen, Guy Marchal, and Paul Suetens. Multimodality image registration by maximization of mutual information. *IEEE transactions on medical imaging*, 16(2):187–198, 1997.

J. F. Mangin, C. Poupon, C. Clark, D. Le Bihan, and I. Bloch. Distortion correction and robust tensor estimation for MR diffusion imaging. *Med Image Anal*, 6(3):191–198, Sep 2002.

D. Mattes, D. R. Haynor, H. Vesselle, T. K. Lewellen, and W. Eubank. PET-CT image registration in the chest using free-form deformations. *IEEE Trans Med Imaging*, 22(1):120–128, Jan 2003.

J. C. Mazziotta, A. W. Toga, A. Evans, P. Fox, and J. Lancaster. A probabilistic atlas of the human brain: theory and rationale for its development. The International Consortium for Brain Mapping (ICBM). *Neuroimage*, 2(2): 89–101, Jun 1995.

Karla L. Miller. Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nat Neurosci*, 2016.

J. P. Pluim, J. B. Maintz, and M. A. Viergever. Mutual-information-based registration of medical images: a survey. *IEEE Trans Med Imaging*, 22(8): 986–1004, Aug 2003.

Russell Poldrack, Chris I Baker, Joke Durnez, Krzysztof Gorgolewski, Paul M Matthews, Marcus Munafo, Thomas Nichols, Jean-Baptiste Poline, Edward Vul, and Tal Yarkoni. Scanning the Horizon: Future challenges for neuroimaging research. *bioRxiv*, 2016.

M.E. Raichle, A.M. MacLeod, A.Z. Snyder, W.J. Powers, D.A. Gusnard, and G.L. Shulman. A default mode of brain function. *Proceedings of the National Academy of Sciences*, 98, 2001.

Stephen M Smith, Mark Jenkinson, Mark W Woolrich, Christian F Beckmann, Timothy EJ Behrens, Heidi Johansen-Berg, Peter R Bannister, Marilena De Luca, Ivana Drobnjak, David E Flitney, et al. Advances in functional and structural MR image analysis and implementation as FSL. *Neuroimage*, 23, 2004.

I Tavor, O Parker Jones, RB Mars, SM Smith, TE Behrens, and S Jbabdi. Task-free MRI predicts individual differences in brain activity during task performance. *Science*, 352(6282):216–220, 2016.

Bertrand Thirion, Philippe Pinel, Sébastien Mériaux, Alexis Roche, Stanislas Dehaene, and Jean-Baptiste Poline. Analysis of a large fMRI cohort: Statistical and methodological issues for group analyses. *Neuroimage*, 35 (1):105–120, 2007.

B. Thyreau, Y. Schwartz, B. Thirion, V. Frouin, E. Loth, S. Vollstadt-Klein, T. Paus, E. Artiges, P. J. Conrod, G. Schumann, R. Whelan, and J. B. Poline. Very large fMRI study using the IMAGEN database: sensitivity-specificity and population effect modeling in relation to the underlying anatomy. *Neuroimage*, 61(1):295–303, May 2012.

D.C. van Essen et al. The human connectome project: A data acquisition perspective. *NeuroImage*, 62(4), 2012.

T. D. Wager, M. Lindquist, and L. Kaplan. Meta-analysis of functional neuroimaging data: current and future directions. *Soc Cogn Affect Neurosci*, 2(2):150–158, Jun 2007.

X. Wan, G. T. Gullberg, D. L. Parker, and G. L. Zeng. Reduction of geometric and intensity distortions in echo-planar imaging using a multireference scan. *Magn Reson Med*, 37(6):932–942, Jun 1997a.

Xin Wan, Grant T Gullberg, Dennis L Parker, and Gengsheng L Zeng. Reduction of geometric and intensity distortions in echo-planar imaging using a multireference scan. *Magnetic Resonance in Medicine*, 37(6):932–942, 1997b.

Sijia Wang, Daniel J. Peterson, J. C. Gatenby, Wenbin Li, Thomas J. Grabowski, and Tara M. Madhyastha. Evaluation of Field Map and Nonlinear Registration Methods for Correction of Susceptibility Artifacts in Diffusion MRI. *Frontiers in Neuroinformatics*, 11:17, 2017.

Takamitsu Watanabe, Satoshi Hirose, Hiroyuki Wada, Yoshio Imai, Toru Machida, Ichiro Shirouzu, Seiki Konishi, Yasushi Miyashita, and Naoki Masuda. A pairwise maximum entropy model accurately describes resting-state human brain networks. *Nat Commun*, 4:1370, Jan 2013.

Lei Xu, Timothy D. Johnson, Thomas E. Nichols, and Derek E. Nee. Modeling inter-subject variability in fMRI activation location: A Bayesian hierarchical spatial model. *Biometrics*, 65(4):1041–1051, Dec 2009.

H. Yamada, O. Abe, T. Shizukuishi, J. Kikuta, T. Shinozaki, K. Dezawa, A. Nagano, M. Matsuda, H. Haradome, and Y. Imamura. Efficacy of distortion correction on diffusion imaging: comparison of FSL eddy and eddy_correct using 30 and 60 directions diffusion encoding. *PLoS ONE*, 9 (11):e112411, 2014.

H. Zeng and R. T. Constable. Image distortion correction in EPI: comparison of field mapping with point spread function mapping. *Magn Reson Med*, 48(1):137–146, Jul 2002.

Learning patterns of inter-subject functional variability from data

Contents

In neuro-imaging, inter-subject variability is often handled as a statistical residual and discarded. Yet there is evidence that it displays structure and contains important information. Univariate models are ineffective both computationally and statistically due to the large number of voxels compared to the number of subjects. Likewise, statistical analysis of weak effects on medical images often relies on defining regions of interests (ROIs). For instance, pharmacology with Positron Emission Tomography (PET) often studies metabolic processes in specific organ sub-parts that are defined from anatomy. Population-level tests of tissue properties, such as diffusion, or simply their density, are performed on ROIs adapted to the spatial impact of the pathology of interest. In functional brain imaging, e.g functional magnetic resonance imaging (fMRI), ROIs must be adapted to the cognitive process under study, and are often defined by the very activation elicited by

a closely related process [Saxe et al., 2006]. ROIs can boost statistical power by reducing multiple comparisons that plague image-based statistical testing. If they are defined to match spatially the differences to detect, they can also improve the signal-to-noise ratio by averaging related signals. However, the crux of the problem is how to define these ROIs in a principled way. Indeed, standard approaches to region definition imply a segmentation step. Segmenting structures on first-level statistical maps, as in fMRI, typically yields meaningful units, but is limited by the noise inherent to these maps. Relying on a different imaging modality hits cross-modality correspondence problems.

9.1 Introduction and sketch of our contributions

In this chapter, we propose to use the *variability* of the statistical maps across the population to define regions. This idea is reminiscent of clustering approaches, that have been employed to define spatial units for quantitative analysis of information as diverse as brain fiber tracking, brain activity, brain structure, or even imaging-genetics. See [Varol and Davatzikos, 2014, Hibar et al., 2013] and references therein. The key idea is to group together features –voxels of an image, vertices on a mesh, fiber tracts– based on the quantity of interest, to create regions –or fiber bundles– for statistical analysis. However, unlike clustering that models each observation as an instance of a cluster, we use a model closer to the signal, where each observation is a linear mixture of several signals. The model is closer to mode finding, as in a principal component analysis (PCA), or an independent component analysis (ICA), often used in brain imaging to extract functional units [Beckmann and Smith, 2004]. Yet, an important constraint is that the modes should be sparse and spatially-localized. For this purpose, the problem can be reformulated as a linear decomposition problem like ICA/PCA, with appropriate spatial and sparse penalties [Varoquaux et al., 2011, Abraham et al., 2013].

We propose a multivariate online dictionary-learning method for obtaining decompositions with structured and sparse components (aka atoms). Sparsity is to be understood in the usual sense: the atoms contain mostly zeros. This is imposed via an ℓ_1 penalty on the atoms. By "structured", we mean that the atoms are piece-wise smooth and compact, thus making up blobs, as opposed to scattered patterns of activation. We impose this type of structure via a Laplacian penalty[1] on the dictionary atoms. Combining the two penalties, we therefore obtain decompositions that are closer to known functional organization of the brain. This non-trivially extends the online dictionary-learning / dictionary-learning work [Mairal et al., 2010], with only a factor of 2 or 3 on the running time. By means of experiments on a large public dataset, we show the improvements brought by the spatial regularization with respect to traditional ℓ_1-regularized dictionary learning. We also provide a concise study of the impact of hyper-parameter selection on this problem and describe the optimality regime, based on relevant criteria (reproducibility, captured variability, explanatory power in prediction problems).

[1] This is a slight abuse of language as we really mean the Sobolev semi-norm $\mathbf{v} \mapsto \mathbf{v}^T \Delta \mathbf{v} = (\nabla \mathbf{v})^T \nabla \mathbf{v} \geq 0$ and not the Laplacian linear operator $\Delta := \nabla^T \nabla$.

Applied to functional brain imaging, it separates successfully activation maps into localized units of brain activity. Our contribution is to frame spatial penalties as a particular case of more general Laplacian regularization and introduce an efficient online algorithm for dictionary-learning in these settings. Applied to functional brain imaging, it separates successfully activation maps into localized units of brain activity. Here we do things that are closer to probabilistic segmentations

9.2 Smooth Sparse Online Dictionary-Learning (Smooth-SODL)

Consider a stack $\mathbf{X} \in \mathbb{R}^{n \times p}$ of n subject-level brain images $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ each of shape $n_1 \times n_2 \times n_3$, seen as p-dimensional row vectors –with $p = n_1 \times n_2 \times n_3$, the number of voxels. These could be images of fMRI activity patterns like statistical parametric maps of brain activation, raw pre-registered (into a common coordinate space) fMRI time-series, PET images, etc. We would like to decompose these images as a product of $k \leq \min(n, p)$ component maps (aka latent factors or dictionary atoms) $\mathbf{d}^1, \ldots, \mathbf{d}^k \in \mathbb{R}^{p \times 1}$ and modulation coefficients $\mathbf{c}_1, \ldots, \mathbf{c}_n \in \mathbb{R}^{k \times 1}$ called *codes* (one k-dimensional code per sample point), i.e

$$\mathbf{x}_i \approx \mathbf{D}\mathbf{c}_i, \text{ for } i = 1, 2, \ldots, n \tag{9.1}$$

where $\mathbf{D} := [\mathbf{d}^1 | \ldots | \mathbf{d}^k] \in \mathbb{R}^{p \times k}$, an unknown dictionary to be estimated.

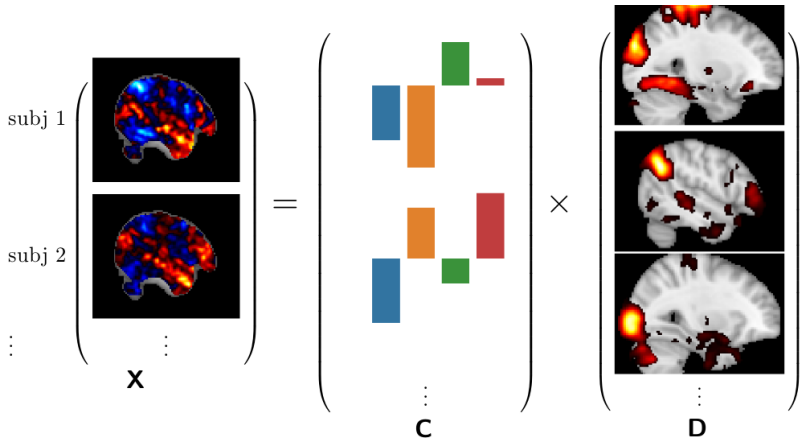


Figure 9.1: **Dictionary-learning** with smoothness and sparsity constraints on the atoms. On the right, each time point is a masked 3D brain image and corresponds to a sample, and each voxel corresponds to a feature, giving an p-by-n matrix $\mathbf{X}$, where n is the number of samples and p is the number of features. On the right of the equation, the sought-for p-by-k dictionary $\mathbf{D}$ is a low-dimensional representation these images, by means of a (non-orthonormal) basis of $k \ll \min(n, p)$ smooth and sparse 3D brain images called *atoms*. In this representation, each sample point (i.e 3D brain image) $\mathbf{X}_i \in \mathbb{R}^p$ is mapped onto k-dimensional vector $\mathbf{c}_i$, called the *code* of $\mathbf{X}_i$.

Typically, $p \sim 10^5 - 10^6$ (in full-brain high-resolution fMRI) and $n \sim 10^2 - 10^5$ (for example, in considering all the 500 subjects and all the about 15 –20 functional tasks of the Human Connectome Project dataset [van Essen et al., 2012]). Our work handles the extreme case where both n and p are large (massive-data setting). It is reasonable then to only consider undercomplete dictionaries: $k \leq \min(n, p)$. Typically, we use $k \sim 50$ or 100 components. It should be noted that online optimization is not only crucial in the case where n/p is big; it is relevant whenever n is large, leading to prohibitive memory issues irrespective of how big or small p is.

As explained in section 9.1, we want the component maps (aka dictionary atoms) $\mathbf{d}^j$ to be sparse and spatially smooth (illustrated in Fig. 9.1). A principled way to achieve such a goal is to impose a boundedness constraint on

ℓ_1-like norms of these maps to achieve sparsity and simultaneously impose smoothness by penalizing their Laplacian. Thus, we propose the following penalized dictionary-learning model

$$\min_{\mathbf{D}\in\mathbb{R}^{p\times k}} \left(\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \min_{\mathbf{c}_i\in\mathbb{R}^k} \frac{1}{2}\|\mathbf{x}_i - \mathbf{D}\mathbf{c}_i\|_2^2 + \frac{1}{2}\alpha\|\mathbf{c}_i\|_2^2 \right) + \gamma \sum_{j=1}^{k} \mathrm{Lap}(\mathbf{d}^j). \quad (9.2)$$
$$\text{subject to } \mathbf{d}^1, \ldots, \mathbf{d}^k \in C$$

The ingredients in the model can be broken down as follows:

- Each of the terms $\max_{\mathbf{c}_i\in\mathbb{R}^k} \frac{1}{2}\|\mathbf{x}_i - \mathbf{D}\mathbf{c}_i\|_2^2$ measures how well the current dictionary $\mathbf{D}$ explains data $\mathbf{x}_i$ from subject i. The Ridge penalty term $\phi(\mathbf{c}_i) \equiv \frac{1}{2}\alpha\|\mathbf{c}_i\|_2^2$ amounts to placing an isotropic Gaussian prior on the codes, namely $p(\mathbf{c}_i) \propto \exp(-\frac{1}{2}\alpha\|\mathbf{c}_i\|_2^2)$. In the context of a specific neuro-imaging problem, if there are good grounds to assume that each sample / subject should be sparsely encoded across only a few atoms of the dictionary, then we can use the ℓ_1 penalty $\phi(\mathbf{c}_i) := \alpha\|\mathbf{c}_i\|_1$ as in [Mairal et al., 2010], corresponding to a Laplace prior on the codes, namely $p(\mathbf{c}_i) \propto \exp(-\alpha\|\mathbf{c}_i\|_1)$. We note that in contrast to the ℓ_1 penalty, the Ridge leads to stable codes. The parameter $\alpha > 0$ controls the amount of penalization on the codes.

- The constraint set C is a sparsity-inducing compact simple[2] convex subset of $\mathbb{R}^p$ like an ℓ_1-ball $\mathbb{B}_{p,\ell_1}(\tau)$ or a simplex $\mathcal{S}_p(\tau)$, defined respectively as
$$\mathbb{B}_{p,\ell_1}(\tau) := \left\{ \mathbf{d} \in \mathbb{R}^p \text{ s.t } |\mathbf{d}_1| + |\mathbf{d}_2| + \ldots + |\mathbf{d}_p| \leq \tau \right\},$$
and $\mathcal{S}_p(\tau) := \mathbb{B}_{p,\ell_1}(\tau) \cap \mathbb{R}^p_+$. Other choices (e.g ElasticNet ball) are of course possible. The radius parameter $\tau > 0$ controls the amount of sparsity: smaller values lead to sparser atoms. The Laplacian regularization Lap (see Table 1 for definition) is meant to impose blobs. $\gamma \geq 0$ controls how much regularization we impose on the atoms, compared to the reconstruction error.

[2] Mainly in the sense that the Euclidean projection onto C should be easy to compute.

The above formulation, which we dub *Smooth Sparse Online Dictionary-Learning* (Smooth-SODL) is inspired by, and generalizes the standard dictionary-learning framework of [Mairal et al., 2010] –henceforth referred to as *Sparse Online Dictionary-Learning* (SODL); setting $\gamma = 0$, we recover SODL [Mairal et al., 2010].

9.3 Algorithms

The objective function in problem (9.2) is separately convex and block-separable w.r.t each of $\mathbf{C}$ and $\mathbf{D}$ but is not jointly convex in $(\mathbf{C}, \mathbf{D})$. Also, it is continuously differentiable on the constraint set, which is compact and convex. Thus by classical results (e.g Bertsekas [Bertsekas, 1999]), the problem can be solved via Block-Coordinate Descent (BCD) [Mairal et al., 2010]. Reasoning along the lines of [Jenatton et al., 2010], we derive that the BCD iterates are as given in Alg. 3. A crucial advantage of using a BCD scheme is that it is parameter free: there is not step size to tune. The resulting algorithm Alg. 3, is adapted from [Mairal et al., 2010]. It relies on Alg. 4

for performing the structured dictionary updates, the details of which are discussed below.

Algorithm 3: Online algorithm for the dictionary-learning problem (9.2)

Require: Regularization parameters $\alpha, \gamma > 0$; initial dictionary $\mathbf{D} \in \mathbb{R}^{p\times k}$, number of passes / iterations T on the data.

1: $\mathbf{A}_0 \leftarrow 0 \in \mathbb{R}^{k\times k}$, $\mathbf{B}_0 \leftarrow 0 \in \mathbb{R}^{p\times k}$ (historical "sufficient statistics")

2: **for** $t = 1$ to T **do**

3: Empirically draw a sample point $\mathbf{x}_t$ at random.

4: Code update: Ridge-regression (via SVD of current dictionary $\mathbf{D}$)

$$\mathbf{c}_t \leftarrow \arg\min_{\mathbf{c}\in\mathbb{R}^k} \frac{1}{2}\|\mathbf{x}_t - \mathbf{D}\mathbf{c}\|_2^2 + \frac{1}{2}\alpha\|\mathbf{c}\|_2^2. \tag{9.3}$$

5: Rank-1 updates: $\mathbf{A}_t \leftarrow \mathbf{A}_{t-1} + \mathbf{c}_t\mathbf{c}_t^T$, $\mathbf{B}_t \leftarrow \mathbf{B}_{t-1} + \mathbf{x}_t\mathbf{c}_t^T$

6: BCD dictionary update: Compute update for dictionary $\mathbf{D}$ using Alg. 4.

7: **end for**

Update of the codes: Ridge-coding. The Ridge sub-problem for updating the codes

$$\mathbf{c}_t = (\mathbf{D}^T\mathbf{D} + \alpha\mathbf{I})^{-1}\mathbf{D}^T\mathbf{x}_t \tag{9.4}$$

is computed via an SVD of the current dictionary $\mathbf{D}$. For $\alpha \approx 0$, $\mathbf{c}_t$ reduces to the orthogonal projection of $\mathbf{x}_t$ onto the image of the current dictionary $\mathbf{D}$. As in [Mairal et al., 2010], we speed up the overall algorithm by sampling mini-batches of η samples $\mathbf{x}_t, \ldots, \mathbf{x}_\eta$ and compute the corresponding codes $\mathbf{c}_1$, $\mathbf{c}_2$, ..., $\mathbf{c}_\eta$ at once. We typically use we use mini-batches of size $\eta \sim 20$ images.

BCD dictionary update for the dictionary atoms. Let us define time-varying matrices $\mathbf{A}_t := \sum_{i=1}^t \mathbf{c}_i\mathbf{c}_i^T \in \mathbb{R}^{k\times k}$ and $\mathbf{B}_t := \sum_{i=1}^t \mathbf{x}_i\mathbf{c}_i^T \in \mathbb{R}^{p\times k}$, where $t = 1, 2, \ldots$ denotes time. We fix the matrix of codes $\mathbf{C}$, and for each j, consider the update of the jth dictionary atom, with all the other atoms $\mathbf{d}^{k\neq j}$ kept fixed. The update for the atom $\mathbf{d}^j$ can then be written as

$$\begin{aligned}
\mathbf{d}^j &= \arg\min_{\mathbf{d}^j\in C} \frac{1}{t}\sum_{i=1}^t \left(\frac{1}{2}\|\mathbf{x}_i - \mathbf{D}\mathbf{c}_i\|_2^2\right) + \gamma\mathrm{Lap}(\mathbf{d}^j) \\
&= \arg\min_{\mathbf{d}^j\in C} \left(\sum_{i=1}^t \frac{1}{2}\|\mathbf{x}_i - \mathbf{D}\mathbf{c}_i\|_2^2\right) + \gamma t\mathrm{Lap}(\mathbf{d}^j) \\
&= \arg\min_{\mathbf{d}^j\in C} F_{\gamma(\mathbf{a}_{j,j}/t)^{-1}}(\mathbf{d}^j, \underbrace{\mathbf{d}^j_{\text{old}} + \mathbf{a}_{j,j}^{-1}(\mathbf{b}^j - \mathbf{D}\mathbf{a}^j)}_{\text{see chapter 10 below}}),
\end{aligned} \tag{9.5}$$

where

$$F_{\tilde{\gamma}}(\mathbf{d}^j, \mathbf{a}) \equiv \frac{1}{2}\|\mathbf{d}^j - \mathbf{a}\|_2^2 + \tilde{\gamma}\mathrm{Lap}(\mathbf{d}^j) = \frac{1}{2}\|\mathbf{d}^j - \mathbf{a}\|_2^2 + \frac{1}{2}\|\nabla\mathbf{d}^j\|^2. \tag{9.6}$$

Problem (9.5) is thus a compactly-constrained minimization of the 1-strongly-convex quadratic functions $F_{\tilde{\gamma}}(., \mathbf{a}) : \mathbb{R}^p \to \mathbb{R}$ defined above. This problem can further be identified with a denoising instance (i.e in which the design

Algorithm 4: BCD dictionary update with Laplacian prior

Require: $\mathbf{D} = [\mathbf{d}^1|\ldots|\mathbf{d}^k] \in \mathbb{R}^{p\times k}$ (input dictionary),

$\mathbf{A}_t = [\mathbf{a}^1|\ldots|\mathbf{a}^k] \in \mathbb{R}^{k\times k}$, $\mathbf{B}_t = [\mathbf{b}^1|\ldots|\mathbf{b}^k] \in \mathbb{R}^{p\times k}$ (history)

while stopping criteria not met, **do**

  for $j = 1$ to r **do**

    Fix the code $\mathbf{C}$ and all atoms $k \neq j$ of the dictionary $\mathbf{D}$ and then update $\mathbf{d}^j$ as follows

$$\mathbf{d}^j \leftarrow \underset{\mathbf{d}^j \in C}{\arg\min}\, F_{\gamma(\mathbf{a}_{j,j}/t)^{-1}}(\mathbf{d}^j, \mathbf{d}^j_{\text{old}} + \mathbf{a}^{-1}_{j,j}(\mathbf{b}^j - \mathbf{D}\mathbf{a}^j)) \tag{9.7}$$

    (See below for details on the derivation and the resolution of this problem)

  end for

end while

matrix or deconvolution operator is the identity operator) of the GraphNet model [Grosenick et al., 2013, Hebiri and van de Geer, 2011]. Fast first-order methods like FISTA [Beck and Teboulle, 2009] with optimal rates $O(L/\sqrt{\epsilon})$ are available[3] for solving such problems to arbitrary precision $\epsilon > 0$. One computes the Lipschitz constant to be $L_{F_{\tilde{\gamma}}(.,\mathbf{a})} \equiv 1 + \tilde{\gamma}L_{\text{Lap}} = 1 + 4D\tilde{\gamma}$, where as before, D is the number of spatial dimensions with $D = 3$ for volumetric images. One should also mention that under certain circumstances, it is possible to perform the dictionary updates in the Fourier domain, via FFT. This alternative approach is developed in the Appendix of [Dohmatob et al., 2016].

[3] For example, see [Dohmatob et al., 2014, Varoquaux et al., 2015], implemented as part of the *Nilearn* open-source Python library [Abraham et al., 2014].

Finally, one notes that, since constraints in problem (9.2) are separable in the dictionary atoms $\mathbf{d}^j$, the BCD dictionary-update algorithm Alg. 4 is guaranteed to converge to a global optimum, at each iteration [Bertsekas, 1999, Mairal et al., 2010].

How difficult is the dictionary update for our proposed model ? A favorable property of the vanilla dictionary-learning [Mairal et al., 2010] is that the BCD dictionary updates amount to Euclidean projections onto the constraint set C, which can be easily computed for a variety of choices (simplexes, closed convex balls, etc.). One may then ask: do we retain a comparable algorithmic simplicity even with the additional Laplacian terms $y\text{Lap}(\mathbf{d}^j)$? The short answer is yes: empirically, we found that 1 or 2 iterations of FISTA [Beck and Teboulle, 2009] are sufficient reach an accuracy of 10^{-6} in problem (9.5), which is sufficient to obtain a good decomposition in the overall algorithm. However, choosing γ "too large" will provably cause the dictionary updates to eventually take forever to run. Indeed, the Lipschitz constant in problem (9.5) is $L_t = 1 + 4D\gamma(a_{j,j}/t)^{-1}$, which will blow-up (leading to arbitrarily small step-sizes) unless γ is chosen so that

$$\gamma = \gamma_t = O\left(\max_{1\leq j\leq k} a_{j,j}\right) = O\left(\max_{1\leq j\leq k}\sum_{i=1}^{t}\|\mathbf{C}^j\|_2^2/t\right) = O(\|A_t\|_{\infty,\infty}/t). \tag{9.8}$$

Finally, the Euclidean projections onto the ℓ_1 ball C can be computed exactly in linear-time $O(p)$ (see for example [Condat, 2014, Duchi et al., 2008]). The dictionary atoms j are repeatedly cycled and problem (9.5) solved. All

in all, in practice we observe that a single iteration is sufficient for the dictionary update sub-routine in Alg. 4 to converge to a qualitatively good dictionary.

9.4 Implementation and practical considerations

Software implementation. All aspects of the code where implemented in the Python programming language. For the implementation of the proposed Alg. 3, we implemented a modified version of *scikit-learn* Python library's *dict_learning* module, to handle more general constraint sets and more general penalties both for the codes $\mathbf{c}_i$ and for the dictionary atoms $\mathbf{d}^j$. The projection onto the ℓ_1-ball C was coded in Cython, a toolkit for writing Python code to run at the speed of the C language.

Convergence of the overall algorithm. The Convergence of our algorithm (to a local optimum) is guaranteed since all hypotheses of [Mairal et al., 2010] are satisfied. For example, assumption **(A)** is satisfied because fMRI data are naturally compactly supported. Assumption **(C)** is satisfied since the ridge-regression problem (9.3) has a unique solution. More details are provided in the Appendix of [Dohmatob et al., 2016].

9.4.1 Practical considerations

Hyper-parameter tuning. Parameter-selection in dictionary-learning is known to be a difficult unsolved problem [Mairal et al., 2010, Jenatton et al., 2010], and our proposed model (9.2) is not an exception to this rule. We did an extensive study of the quality of estimated dictionary varies with the model hyper-parameters (α, γ, τ). The data experimental setup is described in Section 11.5. The results are presented in Fig. 9.2. We make the following observations: Taking the sparsity parameter τ in (9.2) too large leads to dense atoms that perfectly explain the data but are not very intepretable. Taking it too small leads to overly sparse maps that barely explain the data. This normalized sparsity metric (small is better, *ceteris paribus*) is defined as the mean ratio $\|\mathbf{d}^j\|_1 / \|\mathbf{d}^j\|_2$ over the dictionary atoms.

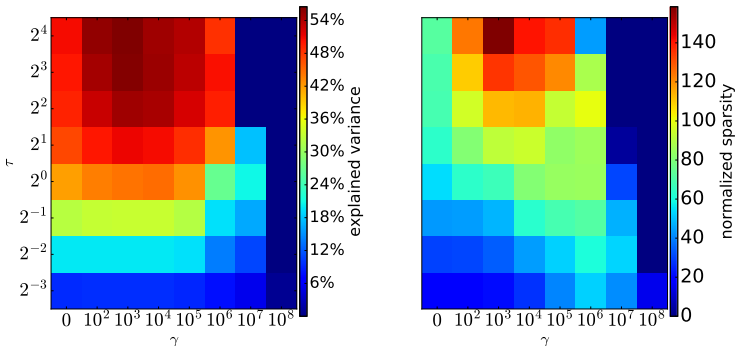


Figure 9.2: **Influence of model parameters.** In the experiments, α was chosen according to (9.9). **Left:** Percentage explained variance of the decomposition, measured on left-out data split. **Right:** Average normalized sparsity of the dictionary atoms.

Concerning the α parameter, inspired by [Ying and Zhou, 2006], we have found the following time-varying data-adaptive choice for the α parameter to work very well in practice:

$$\alpha = \alpha_t \sim t^{-1/2}. \tag{9.9}$$

Likewise, care must be taken in selecting the Laplacian regularization parameter γ. Indeed taking it too small amounts to doing vanilla dictionary-learning model [Mairal et al., 2010]. Taking it too large can lead to degenerate maps, as the spatial regularization then dominates the reconstruction error (data fidelity) term. We find that there is a safe range of the parameter pair (γ, τ) in which a good compromise between the sparsity of the dictionary (thus its intepretability) and its explanation power of the data can be reached. See Fig. 9.2. K-fold cross-validation with explained variance metric was retained as a good strategy for setting the Laplacian regularization γ parameter and the sparsity parameter τ.

Initialization of the dictionary. Problem (9.2) is non-convex jointly in $(\mathbf{C}, \mathbf{D})$, and so initialization might be a might be a crucial issue. However, in our experiments, we have observed that even randomly initialized dictionaries eventually produce sensible results that do not jitter much across different runs of the same experiment.

9.4.2 Interlude: Working in the Fourier domain (when possible)

To close this section, let us point out a few special instances cases of problem (9.6), for peculiar choices of the constraint set Q. First note that the objective in problem (7) can be conveniently rewritten as

$$\begin{aligned} F_{\gamma \mathbf{A}_t[j,j]^{-1}}(\mathbf{v}, \mathbf{v}^j + \mathbf{A}_t[j,j]^{-1}(\mathbf{v}\mathbf{A}^j - \mathbf{B}_t^j)) &= \frac{1}{2}(\mathbf{v} - \tilde{\mathbf{v}}^j)^T(\mathbf{I} - \gamma \mathbf{A}_t[j,j]^{-1}\Delta)(\mathbf{v} - \tilde{\mathbf{v}}^j) \\ &= \frac{1}{2}(\hat{\mathbf{v}} - \hat{\tilde{\mathbf{v}}}^j)^T(\mathbf{I} - \gamma \mathbf{A}_t[j,j]^{-1}\Delta)(\hat{\mathbf{v}} - \hat{\tilde{\mathbf{v}}}^j), \end{aligned} \tag{9.10}$$

with

$$\tilde{\mathbf{v}}^j := (\mathbf{A}_t[j,j]\mathbf{I} - \gamma\Delta)^{-1}\left(\mathbf{v}^j + \mathbf{A}_t[j,j]^{-1}(\mathbf{v}\mathbf{A}^j - \mathbf{B}_t^j)\right). \tag{9.11}$$

We note that the matrix-inversion $(\mathbf{I} - \tilde{\gamma}\Delta)^{-1}$ that appears in the formula above is a Laplacian filter, and can be efficiently applied in closed-form (i.e non-iteratively) in the Fourier / frequency domain. Indeed, under periodic boundary conditions, the discrete Laplacian Δ is Block-Circulant with Circulant Blocks (BCCB) and so is diagonalizable in the Fourier domain. Precisely,

$$\Delta = \mathcal{F}^* \Lambda \mathcal{F} \tag{9.12}$$

where the complex orthonormal operator $\mathcal{F}$ represents the fast Discrete Fourier Transform (DFT), and Λ is diagonal matrix made p eigenvalues (including multiplicities) of the Laplace operator Δ, given by

$$\begin{aligned} \Lambda(\omega) := -\sum_{d=1}^{3}\left(2\sin\left(\frac{\omega_d \pi}{2n_d}\right)\right)^2 = -2\sum_{d=1}^{3}\left(1 - \cos\left(\frac{\omega_d \pi}{n_d}\right)\right) \leq 0, \\ \text{for } \omega = (\omega_1, \omega_2, \omega_3) \in [\![0, n_1 - 1]\!] \times [\![0, n_2 - 1]\!] \times [\![0, n_3 - 1]\!]. \end{aligned}$$

We note that the spectral norm of Laplace operator in D dimensions (here $D = 3$) is $\|\Delta\|_2 = \tilde{\gamma}_{\max}(-\Delta) = 2 \times D \times (1 + 1) = 4D$.

Now, one can then harvest the closed-form solution

$$(\mathbf{I} - \tilde{\gamma}\Delta)^{-1}\mathbf{a} = (\mathcal{F}^{-1}(\mathbf{I} - \tilde{\gamma}\Lambda)^{-1}\mathcal{F})(\mathbf{a}) = \mathcal{F}^{-1}(\mathbf{s}), \tag{9.13}$$

where $\mathbf{s} \in \mathbb{R}^p$ is defined by $\mathbf{s}(\omega) := \dfrac{\hat{\mathbf{a}}(\omega)}{1 - \tilde{\gamma}\hat{\Delta}(\omega)}$, with $\hat{\mathbf{a}} := \mathcal{F}(\mathbf{a})$. These DFT computations have complexity $O(p \log p)$.

For applying the DFTs above, one can use the FFTW[4] –or *Fastest Fourier Transform in the West*– library for computing the forward and inverse Fourier transforms needed to apply the Laplacian filter.

[4] FFTW is generally taught to be one of the fastest implementations of the FFT, yielding up to 3× speedup against competing libraries like LAPACK.

Pure ℓ_2 constraint. Here, the constraint set C is an L2 ball (with radius = 1, w.l.o.g) in $\mathbb{R}^2$. By the Rayleigh energy theorem (aka Parseval's identity for the DFT), one has

$$\|\hat{\mathbf{v}}\|^2 = p\|\mathbf{v}\|_2^2, \ \forall \mathbf{v} \in \mathbb{R}^p$$

and so problem (7) can be written as

$$\begin{aligned}
\mathbf{v}^j &\leftarrow \underset{\mathbf{v}\in\mathbb{R}^p,\ \|\mathbf{v}\|_2^2 \le 1}{\arg\min} \frac{1}{2}(\hat{\mathbf{v}} - \hat{\mathbf{v}}^j)^*(\mathbf{I} - \gamma \mathbf{A}_t[j,j]^{-1}\Lambda)(\hat{\mathbf{v}} - \hat{\mathbf{v}}^j) \\
&= \mathcal{F}^*\left(\underset{\hat{\mathbf{v}}\in\mathbb{C}^p,\ \|\hat{\mathbf{v}}\|_2^2 \le p}{\arg\min} \frac{1}{2}(\hat{\mathbf{v}} - \hat{\mathbf{v}}^j)^*(\mathbf{I} - \gamma \mathbf{A}_t[j,j]^{-1}\Lambda)(\hat{\mathbf{v}} - \hat{\mathbf{v}}^j) \right) \\
&= \mathcal{F}^*\left(P_{\mathcal{E}}(\hat{\mathbf{v}}^j) \right)
\end{aligned} \tag{9.14}$$

where

$$\mathcal{E} := \left\{ (\mathbf{I} - \gamma \mathbf{A}_t[j,j]^{-1}\Lambda)^{\frac{1}{2}} \hat{\mathbf{v}} \text{ s.t } \hat{\mathbf{v}} \in \mathbb{C}^p, \ \|\hat{\mathbf{v}}\|_2^2 \le p \right\}, \tag{9.15}$$

a hyper-ellipsoid in standard position (i.e $\mathbf{0}$-centered and axes-aligned). Using elementary geometric arguments, one can show that the projection $P_{\mathcal{E}}(\hat{\mathbf{v}}^j)$ can be computed efficiently using a kind of root-finding algorithm [Dai, 2006], and converges exponentially fast.

Non-negative Lasso. In case the constraint set C for the dictionary atoms is a simplex $\mathcal{S}_p(\tau)$, the simplex (see section 9.2), then the BCD update for the jth atom becomes

$$\begin{aligned}
\mathbf{v}^j &\leftarrow \underset{\mathbf{v}\in\mathbb{R}^p,\ \mathbf{v}\ge 0,\ \mathbf{1}^T\mathbf{v}\le 1}{\arg\min} \frac{1}{2}(\hat{\mathbf{v}} - \hat{\mathbf{v}}^j)^*(\mathbf{I} - \gamma \mathbf{A}_t[j,j]^{-1}\Lambda)(\hat{\mathbf{v}} - \hat{\mathbf{v}}^j) \\
&= \mathcal{F}^*\left(\underset{\hat{\mathbf{v}}\in\mathbb{C}^p,\ -\mathcal{F}^*\hat{\mathbf{v}}\le 0,\ \hat{\mathbf{1}}^T\hat{\mathbf{v}}\le 1}{\arg\min} \frac{1}{2}(\hat{\mathbf{v}} - \hat{\mathbf{v}}^j)^*(\mathbf{I} - \gamma \mathbf{A}_t[j,j]^{-1}\Lambda)(\hat{\mathbf{v}} - \hat{\mathbf{v}}^j) \right),
\end{aligned} \tag{9.16}$$

which is a diagonal quadratic program with linear constraints, and can be effectively solved via the well-known simplex method, for example.

9.5 Related works

While there exist algorithms for online sparse dictionary-learning that are very efficient in large-scale settings (for example [Mairal et al., 2010], or more recently [Mensch et al., 2016]) imposing spatial structure introduces couplings in the corresponding optimization problem [Dohmatob et al., 2014]. So far, spatially-structured decompositions have been solved by very slow alternated optimization [Varoquaux et al., 2011, Abraham et al., 2013]. Notably, structured priors such as TV-ℓ_1 [Baldassarre et al., 2012] minimization, were used by [Abraham et al., 2013] to extract data-driven state-of-the-art atlases of brain function. However, alternated minimization is

very slow, and large-scale medical imaging has shifted to online solvers for dictionary-learning like [Mairal et al., 2010] and [Mensch et al., 2016]. These do not readily integrate structured penalties. As a result, the use of structured decompositions has been limited so far, mostly due to the computational cost of the ensuing algorithms. Our approach instead uses a Laplacian penalty to impose spatial structure at a very minor cost and adapts the online-learning dictionary-learning framework [Mairal et al., 2010], resulting in a fast and scalable structured decomposition. Second, the approach in [Abraham et al., 2013] though very novel, is heuristic, as it does not come with theoretical guarantees. In contrast, our method enjoys the same convergence guarantees and comparable numerical complexity as the basic unstructured online dictionary-learning [Mairal et al., 2010].

Finally, one should also mention [Varoquaux et al., 2013] that introduced an online group-level functional brain mapping strategy for differentiating regions reflecting the variety of brain network configurations observed a the population, by learning a sparse representation of these in the spirit of [Mairal et al., 2010].

9.6 Experiments

Setup. Our experiments were done on task fMRI data from 500 subjects from the HCP –Human Connectome Project– dataset [van Essen et al., 2012]. These task fMRI data were acquired in an attempt to assess major domains that are thought to sample the diversity of neural systems of interest in functional connectomics. We studied the activation maps related to a task that involves language (story understanding) and mathematics (mental computation). This particular task is expected to outline number, attentional and language networks, but the variability modes observed in the population cover even wider systems. For the experiments, mass-univariate General Linear Models (GLMs) [Friston et al., 1995] for $n = 500$ subjects were estimated for the *Math vs Story* contrast (language protocol), and the corresponding full-brain Z-score maps each containing $p = 2.6 \times 10^5$ voxels, were used as the input data $\mathbf{X} \in \mathbb{R}^{n \times p}$, and we sought a decomposition into a dictionary of $k = 40$ atoms (components). The input data $\mathbf{X}$ were shuffled and then split into two groups of the same size.

Models compared and metrics. We compared our proposed Smooth-SODL model (9.2) against both the Canonical ICA –CanICA [Varoquaux et al., 2010], a single-batch multi-subject PCA/ICA-based method, and the standard SODL (sparse online dictionary-learning) [Mairal et al., 2010]. While the CanICA model accounts for subject-to-subject differences, one of its major limitations is that it does not model spatial variability across subjects. Thus we estimated the CanICA components on smoothed data: isotropic FWHM of 6mm, a necessary preprocessing step for such methods. In contrast, we did no pre-smoothing for the SODL of Smooth-SODL models. The different models were compared across a variety of qualitative and quantitative metrics: visual quality of the dictionaries obtained, explained variance, stability of the dictionary atoms, their reproducibility, performance of the dictionaries in predicting behavioral scores (IQ, picture

vocabulary, reading proficiency, etc.) shipped with the HCP data [van Essen et al., 2012]. For both SODL [Mairal et al., 2010] and our proposed Smooth-SODL model, the constraint set for the dictionary atoms was taken to be a simplex $C := \mathcal{S}_p(\tau)$ (see section 9.2 for definition). The results of these experiments are presented in Fig. 9.3 and 9.5.

9.7 Results

Qualitative assessment of dictionaries. As can be seen in Fig. 9.3, all methods recover dictionary atoms that represent known functional brain organization; notably the dictionaries all contain the well-known executive control and attention networks, at least in part. Vanilla dictionary-learning leverages the denoising properties of the ℓ_1 sparsity constraint, but the voxel clusters are not very structured. For, example most blobs are surrounded with a thick ring of very small nonzero values. In contrast, our proposed regularization model leverages both sparse and structured dictionary atoms, that are more spatially structured and less noisy.

In contrast to both SODL and Smooth-SODL, CanICA [Varoquaux et al., 2010] is an ICA-based method which enforces no notion of sparsity whatsoever. The result are therefore dense and noisy dictionary atoms that explain the data very well (Fig. 9.4 but which are completely unintepretable. In a futile attempt to remedy the situation, in practice such PCA/ICA-based methods (including FSL's MELODIC tool [Smith et al., 2004]) are hard-thresholded in order to see information. For CanICA, the hard-thresholded version has been named tCanICA in Fig. 9.3. That notwithstanding, notice how the major structures (parietal lobes, sulci, etc.) in each atom are reproducible across the different models.

Stability-fidelity trade-offs. PCA/ICA-based methods like CanICA [Varoquaux et al., 2010] and MELODIC [Smith et al., 2004] are the optimal linear decomposition method to maximize explained variance on a dataset. On the training set, CanICA [Varoquaux et al., 2010] out-performs all others algorithms with about 66% (resp. 50% for SODL [Mairal et al., 2010] and 58% for Smooth-SODL) of explained variance on the training set, and 60% (resp. 49% for SODL and **55%** for Smooth-SODL) on left-out (test) data. See Fig. 9.4. However, as noted in the above paragraph, such methods lead to dictionaries that are hardly intepretable and thus the user must recourse to some kind of post-processing hard-thresholding step, which destroys the estimated model. More so, assessing the stability of the dictionaries, measured by mean correlation between corresponding atoms, across different splits of the data, CanICA [Varoquaux et al., 2010] scores a meager 0.1, whilst the hard-thresholded version tCanICA obtains 0.2, compared to **0.4** for Smooth-SODL and 0.1 for SODL. As is to be expected, notice how the RAW model over-fits. The voxel space of worth $p = 261596$ voxels is reduced to $k = 40$ components, and then each subject Z-map $\mathbf{x}_t$ of worth $p = 261596$ voxels is reduced to $k = 40$ coefficients via a simple ridge regression (9.3).

Prediction of behavioral variables. If Smooth-SODL captures the patterns of inter-subject variability, then it should be possible to predict cogni-

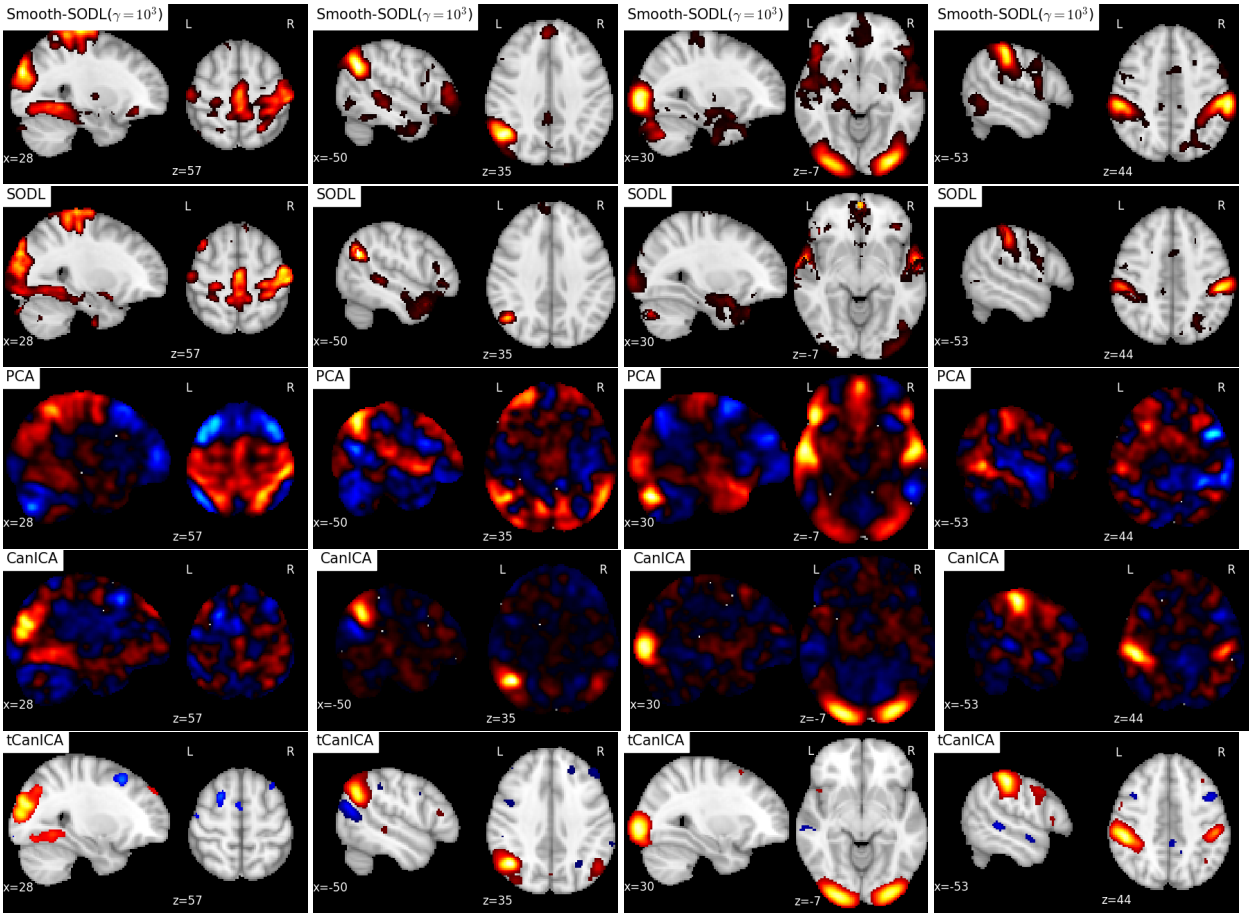


Figure 9.3: **Qualitative comparison of the estimated dictionaries.** Each column represents an atom of the estimated dictionary, where atoms from the different models (the rows of the plots) have been matched via a Hungarian algorithm. Here, we only show a limited number of the most "intepretable" atoms. Notice how the major structures in each atom are reproducible across the different models. Maps corresponding to hard-thresholded CanICA [Varoquaux et al., 2010] components have also been included, and have been called tCanICA. In contrast, the maps from the SODL [Mairal et al., 2010] and our proposed Smooth-SODL (9.2) have not been thresholded.

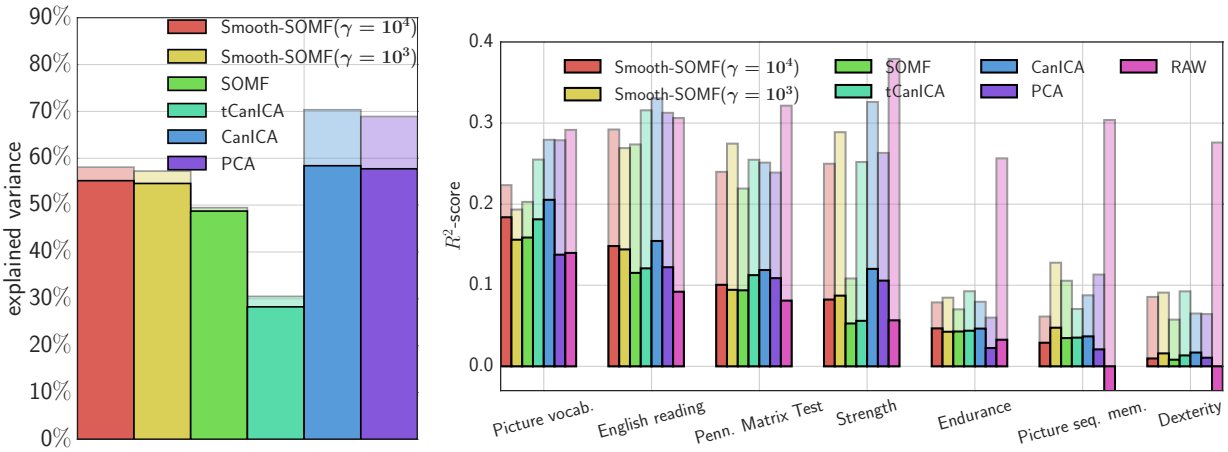
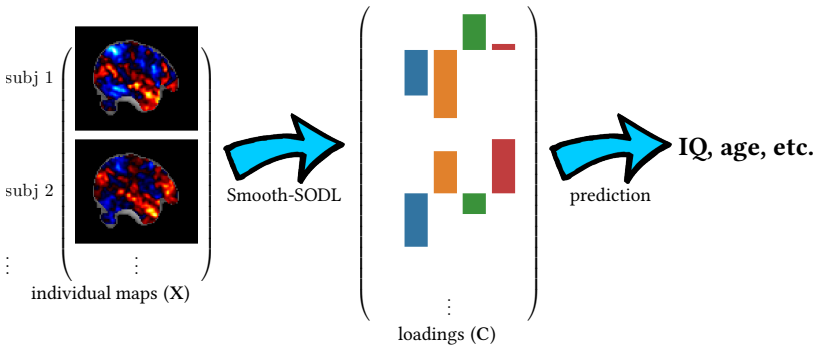


Figure 9.4: **Left:** Mean explained variance of the different models on both training data and test (left-out) data. **N.B.:** Bold bars represent performance on **test** set while faint bars in the background represent performance on **train** set. **Right:** Predicting behavioral variables of the HCP [van Essen et al., 2012] dataset using subject-level Z-maps.

tive scores $\mathbf{y}$ like picture vocabulary, reading proficiency, math aptitude, etc. (the behavioral variables are explained in the HCP wiki [hcp]) by projecting new subjects' data into this learned low-dimensional space (via solving the ridge problem (9.3) for each sample $\mathbf{x}_t$), without loss of performance compared with using the raw Z-values values $\mathbf{X}$.



Let RAW refer to the direct prediction of targets $\mathbf{y}$ from $\mathbf{X}$, using the top 2000 most voxels most correlated with the target variable. Results of the comparison are shown in Fig. 9.4. Only variables predicted with a positive mean (across the different methods and across subjects) R-score are reported. We see that the RAW model, as expected over-fits drastically, scoring an R^2 of 0.3 on training data and only 0.14 on test data. Overall, for this metric CanICA performs best than all the other models in predicting the different behavioral variables on test data. However, our proposed Smooth-SODL model outperforms both SODL [Mairal et al., 2010] and tCanICA, the thresholded version of CanICA.

Running time. On the computational side, the vanilla dictionary-learning SODL algorithm [Mairal et al., 2010] with a batch size of $\eta = 20$ took about 110s ($\approx$ 1.7 minutes) to run, whilst with the same batch size, our proposed Smooth-SODL model (9.2) implemented in Alg. 3 took 340s ($\approx$ 5.6 minutes), which is slightly less than **3 times** slower than SODL. Finally, CanICA [Varoquaux et al., 2010] for this experiment took 530s ($\approx$ 8.8 minutes) to run, which is about **5 times** slower than the SODL model and **1.6 times** slower than our proposed Smooth-SODL (9.2) model. All experiments were run on a single CPU of a modern laptop.

Is spatial regularization really needed ? Ideally, one does not need spatial regularization if data are abundant (like in the HCP). So we computed learning curves of mean explained variance (EV) on test data, as a function of the amount training data seen by both Smooth-SODL and SODL [Mairal et al., 2010] (Fig. 9.5). In the beginning of the curve, our proposed spatially regularized Smooth-SODL model starts off with more than 31% explained variance (computed on 241 subjects), after having pooled only 17 subjects. In contrast, the vanilla SODL model [Mairal et al., 2010] scores a meager 2% explained variance; this corresponds to a 14-fold gain of Smooth-SODL over SODL. As more and more that are pooled, both models explain more variance, and the gap between Smooth-SODL and SODL reduces, and both models perform comparably asymptotically.

# subjects pooled	vanilla SODL	proposed model	gain factor
17	2%	31%	**x13.8**
92	37%	50%	**x1.35**
167	47%	54%	**x1.15**
241	49%	55%	**x1.11**

Figure 9.5: **Learning-curve** for boost in explained variance of our proposed Smooth-SODL model (9.2) over the reference sparse online dictionary-learning (SODL) model [Mairal et al., 2010]. Note the reduction in the gain in EV as more data are pooled.

9.8 Concluding remarks

To extract structured functionally discriminating patterns from massive brain data (i.e data-driven atlases), we have extended the online dictionary-learning framework first developed in [Mairal et al., 2010], to learn structured regions representative of brain organization. To this end, we have successfully augmented [Mairal et al., 2010] with a Laplacian prior on the component maps, while conserving the low numerical complexity of the latter. Through experiments, we have shown that the resulting model –Smooth-SODL model (9.2)– extracts structured and denoised dictionaries that are more intepretable and better capture inter-subject variability in small medium, and large-scale regimes alike, compared to state-of-the-art models. We believe such online multivariate online methods shall become the de facto way do dimensionality reduction and ROI extraction in future.

Implementation. The authors' implementation of the proposed SSODL (9.2) model will soon be made available as part of the *Nilearn* package [Abraham et al., 2014].

9.8.1 Possible extensions

More general structure-imposing penalties. One can envisage to replace the Laplacian regularization with a general structure-inducing penalty for which the proximal operator is easy to compute. Such a framework is developed in chapter 10, and produces an entire family of models, with potentially different properties.

Replacing the dictionary with a general neural net. One notes that the proposed model (9.2) can be seen as an auto-encoding model of brain data with linear generator

$$G_{\mathbf{D}} = \langle \mathbf{D}, . \rangle : \mathbb{R}^k \to \mathbb{R}^p, \ \mathbf{c} \mapsto \mathbf{x} := \mathbf{D}\mathbf{c}, \tag{9.17}$$

parametrized by the shared learned dictionary $\mathbf{D}$, and an *implicit* encoder

$$E_{\mathbf{D}} : \mathbb{R}^p \to \mathbb{R}^k, \ \mathbf{x} \mapsto \arg\min_{\mathbf{c} \in \mathbb{R}^k} -\text{loglik}(\mathbf{x}|\mathbf{c}, \mathbf{D}) = \arg\min_{\mathbf{c} \in \mathbb{R}^k} \ell(G_{\mathbf{D}}(\mathbf{c}), \mathbf{x}) + \alpha\phi(\mathbf{c}), \tag{9.18}$$

which is by construction, the optimal encoder [5] for the the generator $G_{\mathbf{D}}$. One could obtain much greater modeling flexibility by replacing the generator (9.17) by a neural network (see Fig. 9.6)

$$G_{\boldsymbol{\theta}} : \mathbb{R}^k \to \mathbb{R}^p, \ \mathbf{c} \mapsto \mathbf{x} := G_{\boldsymbol{\theta}}(\mathbf{c}),$$

with parameters $\boldsymbol{\theta} \in \Theta$. The model could be successfully trained via stochastic gradient descent (SGD) or its various enhanced variants. Such models,

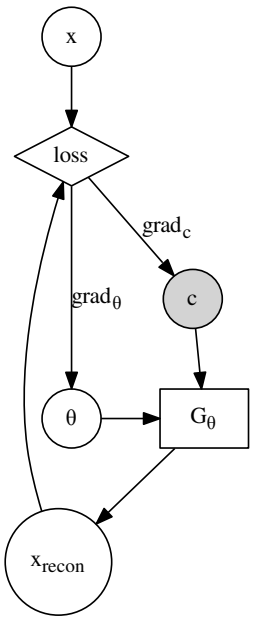


Figure 9.6: **General neural net** schema for structured online dictionary learning. The model is trained via stochastic gradient descent –SGD. For an incoming image $\mathbf{x} \in \mathbb{R}^p$ (or mini-batch of images), a code $\mathbf{z} \in \mathbb{R}^k$ is sampled and fed into th generator $G_{\boldsymbol{\theta}}$ to produce a reconstructed image $\mathbf{x}_{\text{recon}} \in \mathbb{R}^p$, which is compared with the original via a loss function. The model (9.2) corresponds to a linear generator $G_{\boldsymbol{\theta}} = G_{\mathbf{D}} = \langle \mathbf{D}, . \rangle$, and can be solved via Alg. 4.

[5] Under the prior for the codes $p(\mathbf{c}) \propto \exp(-\alpha\phi(\mathbf{c}))$.

which can be referred to as "encoder-less auto-encoders", have recently been proposed in the compressed sensing [Bora et al., 2017] and computer vision literatures [Bojanowski et al., 2017].

Bibliography

HCP wiki. `https://wiki.humanconnectome.org/display/PublicData/HCP+Data+Dictionary+Public-+500+Subject+Release`. Accessed: 2010-09-30.

Alexandre Abraham, Elvis Dohmatob, Bertrand Thirion, Dimitris Samaras, and Gael Varoquaux. Extracting brain regions from rest fMRI with Total-Variation constrained dictionary learning. In *MICCAI*. 2013.

Alexandre Abraham, Fabian Pedregosa, Michael Eickenberg, Philippe Gervais, Andreas Mueller, Jean Kossaifi, Alexandre Gramfort, Bertrand Thirion, and Gael Varoquaux. Machine learning for neuroimaging with scikit-learn. *Frontiers in Neuroinformatics*, 2014.

Luca Baldassarre, Janaina Mourao-Miranda, and Massimiliano Pontil. Structured sparsity models for brain decoding from fMRI data. In *PRNI*, 2012.

A. Beck and M. Teboulle. A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems. *SIAM Journal on Imaging Sciences*, 2, 2009.

C. F. Beckmann and S. M. Smith. Probabilistic independent component analysis for functional magnetic resonance imaging. *Trans Med Im*, 23, 2004.

Dimitri P Bertsekas. *Nonlinear programming*. Athena Scientific, 1999.

Piotr Bojanowski, Armand Joulin, David Lopez-Paz, and Arthur Szlam. Optimizing the Latent Space of Generative Networks. *arXiv preprint arXiv:1707.05776*, 2017.

Ashish Bora, Ajil Jalal, Eric Price, and Alexandros G. Dimakis. Compressed Sensing using Generative Models. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pages 537–546, 2017.

Laurent Condat. Fast projection onto the simplex and the ℓ_1 ball. *Mathematical Programming*, 2014.

Yu-Hong Dai. Fast Algorithms for Projection on an Ellipsoid. *SIAM Journal on Optimization*, 16(4), 2006.

Elvis Dohmatob, Alexandre Gramfort, Bertrand Thirion, and Gaël Varoquaux. Benchmarking solvers for TV-l1 least-squares and logistic regression in brain imaging. In *PRNI*. IEEE, 2014.

Elvis Dohmatob, Arthur Mensch, Gaël Varoquaux, and Thirion Bertrand. Learning brain regions via large-scale online structured sparse dictionary-learning. In *NIPS*, 2016.

John Duchi, Shai Shalev-Shwartz, Yoram Singer, and Tushar Chandra. Efficient projections onto the l 1-ball for learning in high dimensions. In *Proceedings of the 25th international conference on Machine learning*. ACM, 2008.

K. J. Friston, A. P. Holmes, K. J. Worsley, J.-B. Poline, C. Frith, and R. S. J. Frackowiak. Statistical Parametric Maps in Functional Imaging: A General Linear Approach. *Hum Brain Mapp*, 1995.

Logan Grosenick, Brad Klingenberg, Kiefer Katovich, Brian Knutson, and Jonathan E Taylor. Interpretable whole-brain prediction analysis with GraphNet. *NeuroImage*, 72, 2013.

M. Hebiri and S. van de Geer. The Smooth-Lasso and other $\ell_1 + \ell_2$-penalized methods. *Electron. J. Stat.*, 5, 2011.

Derrek P Hibar, Sarah E Medland, Jason L Stein, Sungeun Kim, Li Shen, Andrew J Saykin, Greig I De Zubicaray, Katie L McMahon, Grant W Montgomery, Nicholas G Martin, et al. Genetic clustering on the hippocampal surface for genome-wide association studies. In *MICCAI*. 2013.

Rodolphe Jenatton, Guillaume Obozinski, and Francis Bach. Structured sparse principal component analysis. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2010.

J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research*, 11, 2010.

Arthur Mensch, Julien Mairal, Bertrand Thirion, and Gaël Varoquaux. Dictionary Learning for Massive Matrix Factorization. *arXiv:1605.00937*, 2016.

R. Saxe, M. Brett, and N. Kanwisher. Divide and conquer: a defense of functional localizers. *Neuroimage*, 30(4), May 2006.

Stephen M Smith, Mark Jenkinson, Mark W Woolrich, Christian F Beckmann, Timothy EJ Behrens, Heidi Johansen-Berg, Peter R Bannister, Marilena De Luca, Ivana Drobnjak, David E Flitney, et al. Advances in functional and structural MR image analysis and implementation as FSL. *Neuroimage*, 23, 2004.

D.C. van Essen et al. The human connectome project: A data acquisition perspective. *NeuroImage*, 62(4), 2012.

Erdem Varol and Christos Davatzikos. Supervised block sparse dictionary learning for simultaneous clustering and classification in computational anatomy. *Med Image Comput Comput Assist Interv*, 17(Pt 2), 2014.

Gaël Varoquaux, Sepideh Sadaghiani, Philippe Pinel, Andreas Kleinschmidt, Jean-Baptiste Poline, and Bertrand Thirion. A group model for stable multi-subject ICA on fMRI datasets. *Neuroimage*, 2010.

Gaël Varoquaux, A. Gramfort, F. Pedregosa, V. Michel, and B. Thirion. Multi-subject dictionary learning to segment an atlas of brain spontaneous activity. In *Inf Proc Med Imag*, 2011.

Gaël Varoquaux, Yannick Schwartz, Philippe Pinel, and Bertrand Thirion. Cohort-level brain mapping: learning cognitive atoms to single out specialized regions. In *IPMI*, 2013.

Gaël Varoquaux, Michael Eickenberg, Elvis Dohmatob, and Bertand Thirion. FAASTA: A fast solver for total-variation regularization of ill-conditioned problems with application to brain imaging. *arXiv:1512.06999*, 2015.

Yiming Ying and D-X Zhou. Online regularized classification algorithms. *IEEE Transactions on Information Theory*, 52(11), 2006.

Proximal updates for online dictionary-learning

Contents

In this chapter, we consider possible generalization of model (9.2) proposed in chapter 9, by allowing for more general block-separable penalties on the dictionary.

10.1 The power of the prox

Recall from equation (9.2) of chapter 9 that after n passes over the data, the unpenalized objective (i.e loss) function whose minimization gives the dictionary updates is $E(\mathbf{D}) := \frac{1}{2n}\|\mathbf{X} - \mathbf{DC}\|_{\mathrm{F}}^2$, where $\mathbf{C} \in \mathbb{R}^{n\times k}$ is the fixed matrix of codes computed up to this point, and $\mathbf{D} \in \mathbb{R}^{p\times k}$ is the dictionary variable. Adding a **block-separable** penalty term $\gamma \sum_{j=1}^{k} g_j(\mathbf{d}^j)$, the energy becomes

$$E(\mathbf{D}) = \frac{1}{2n}\|\mathbf{X} - \mathbf{DC}\|_{\mathrm{F}}^2 + \gamma \sum_j g_j(\mathbf{d}^j). \tag{10.1}$$

In particular, taking $g_j(\mathbf{d}^j) := \frac{1}{2}\|\nabla \mathbf{d}^j\|_{\mathrm{F}}^2 + i_{\mathbb{B}_{p,1}}(\mathbf{d}^j)$ corresponds to the model proposed in chapter 9. Now, one easily computes

$$\nabla_{\mathbf{D}}\left(\frac{1}{2n}\|\mathbf{X} - \mathbf{DC}\|_{\mathrm{F}}^2\right) = \frac{1}{n}(\mathbf{DC} - \mathbf{X})\mathbf{C}^T = \mathbf{DA} - \mathbf{B},$$

where $\mathbf{A} := \frac{1}{n}\mathbf{C}\mathbf{C}^T = \frac{1}{n}\sum_{i=1}^{n} \mathbf{c}_i\mathbf{c}_i^T \in \mathbb{R}^{k\times k}$ and $\mathbf{B} := \frac{1}{n}\mathbf{X}\mathbf{C}^T = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i\mathbf{c}_i^T \in \mathbb{R}^{p\times k}$. Now in BCD, the jth atom is updated whilst all the others are held constant. Selecting the jth column of (10.1), we get $\nabla_{\mathbf{d}^j}\left(\frac{1}{2n}\|\mathbf{X} - \mathbf{DC}\|_{\mathrm{F}}^2\right) =$

$\mathbf{D}\mathbf{a}^j - \mathbf{b}^j$. Putting things together, we get

$$\mathbf{p} = \underset{\mathbf{d}^j \in \mathbb{R}^p,\ \mathbf{d}^l \text{ fixed } \forall l \neq j}{\arg\min} E(\mathbf{D}) \iff \mathbf{0} \in \partial_{\mathbf{d}^j} E(\mathbf{D}) = \mathbf{D}\mathbf{a}^j - \mathbf{b}^j\big|_{\mathbf{d}^j = \mathbf{p}} + \gamma \partial g_j(\mathbf{p})$$

$$\iff \left(\mathbf{b}^j - \sum_{l \neq j} a_{j,l}\mathbf{d}^l\right) - a_{j,j}\mathbf{p} \in \gamma \partial g_j(\mathbf{p}) \overset{\text{Lemma2}}{\iff} \mathbf{p} = \text{prox}_{\gamma a_{j,j}^{-1} g_j}(\mathbf{z}^{-j})$$

where

$$\mathbf{z}^{-j} := a_{j,j}^{-1}\left(\mathbf{b}^j - \textstyle\sum_{l \neq j} a_{j,l}\mathbf{d}^l\right),$$

and the last equivalence results from the following elementary lemma which reveals that the prox of a function at a point can be seen as an *implicit* gradient step. Viz

Lemma 2. *For a function $f : \mathcal{H} \to (-\infty, +\infty]$ (convex or not), recall the definition of its subdiffferential at a point $\mathbf{p} \in \mathcal{H}$, namely $\partial f(\mathbf{p}) := \{\mathbf{v} \in \mathcal{H} | f(\mathbf{z}) \geq f(\mathbf{p}) + \mathbf{v}^T(\mathbf{z} - \mathbf{p})\ \forall \mathbf{z} \in \mathcal{H}\}$. We have the following characterization of the prox*

$$\mathbf{p} \in \text{prox}_f(\mathbf{d}) \iff \mathbf{d} - \mathbf{p} \in \partial f(\mathbf{p}). \tag{10.2}$$

Proof.

$$\begin{aligned}
\mathbf{p} \in \text{prox}_f(\mathbf{d}) &\iff \frac{1}{2}\|\mathbf{p} - \mathbf{d}\|_2^2 + f(\mathbf{p}) \leq \frac{1}{2}\|\mathbf{z} - \mathbf{d}\|_2^2 + f(\mathbf{z})\ \forall \mathbf{z} \\
&\iff f(\mathbf{p}) + \frac{1}{2}\|\mathbf{p}\|_2^2 - \mathbf{d}^T\mathbf{p} \leq f(\mathbf{z}) + \frac{1}{2}\|\mathbf{z}\|_2^2 - \mathbf{d}^T\mathbf{z}\ \forall \mathbf{z} \\
&\iff f(\mathbf{p}) + \frac{1}{2}\|\mathbf{p}\|_2^2 + \mathbf{d}^T(\mathbf{z} - \mathbf{p}) \leq f(\mathbf{z}) + \frac{1}{2}\|\mathbf{z}\|_2^2\ \forall \mathbf{z} \\
&\iff \mathbf{d} \in \partial(f + \frac{1}{2}\|.\|_2^2)(\mathbf{p}) = \mathbf{p} + \partial f(\mathbf{p}) \\
&\iff \mathbf{d} - \mathbf{p} \in \partial f(\mathbf{p}).
\end{aligned}$$

□

Putting everything together, we have the close-form BCD update formula for regularized online dictionary learning[1]

[1] We assumed W.L.O.G that $a_{j,j} \neq 0$ (i.e $a_{j,j} > 0$), meaning that the jth atom is active in the representation of at least one sample. Otherwise, we can update the jth atom with a random vector, or even skip it altogether.

BCD updates DL with general separable penalties. The BCD updates for a penalized DL model (10.1) is

$$\mathbf{d}^j \leftarrow \text{prox}_{\gamma a_{j,j}^{-1} g_j}(\mathbf{z}^{-j}), \tag{10.3}$$

where $\mathbf{z}^{-j} := a_{j,j}^{-1}\mathbf{r}^j$ and $\mathbf{r}^j := \left(\mathbf{b}^j - \sum_{l \neq j} a_{j,l}\mathbf{d}^l\right)$.

For the purposes of practical implementation, one notes that $\mathbf{r}^j$ is precisely the jth column of the p-by-k matrix

$$\mathbf{R} := \mathbf{B} - \mathbf{D}\mathbf{A} + \mathbf{d}^j \circ \mathbf{a}^j. \tag{10.4}$$

Thus at the begining of the BCD updates, we precompute the difference $\mathbf{D}\mathbf{A} - \mathbf{B}$, and substract (resp. add) the rank-1 term $\mathbf{d}^j \circ \mathbf{a}^j$ before and after updating the j atom $\mathbf{d}$ according to (10.3). Such rank-1 updates are natively optimized in all linear algebra scientific computation packages.

10.2 Applications

Now that we have the hammer, where are the nails...

10.2.1 Special cases

Constraint sets

If we take $g_j := i_{C_j}$, the indicator function of a closed convex subset of $\mathbb{R}^p$, so that each atom $\mathbf{d}^j$ is constrained to satisfy a set of constraints prescribed by C_j, then the above updates reduce to projecting $\mathbf{z}^{-j}$ onto C_j, namely

$$\mathbf{d}^j \leftarrow \text{proj}_{C_j}(\mathbf{z}^{-j}). \tag{10.5}$$

This is interesting as long as the C_j's are sufficiently "simple" to allow us compute the above projection easily (preferably in closed form).

Classical choice. Taking C_j to be $\mathbb{B}_2$, the unit ball for the euclidean norm on $\mathbb{R}^p$, we recover the updates proposed in [Mairal et al., 2009, 2010], namely

$$\mathbf{d}^j \leftarrow \frac{\mathbf{z}^{-j}}{\max(1, \|\mathbf{z}^{-j}\|_2)}. \tag{10.6}$$

One notes that these constraint has no structural properties beyond preventing the dictionary atoms from becoming arbitrarily large.

Gram-Schmidt / step-wise orthonormality constraints. Akin to ICA-type methods, one can take C_j = the orthogonal complement of the linear span of the first $j-1$st atoms, namely

$$C_j = \text{span}\{\mathbf{d}^l | l < j\}^\perp \cap \mathbb{B}_2. \tag{10.7}$$

The dictionary updates are then simply the Gram-Schmidt orthonormalization of the ordered sequence of vectors $\mathbf{z}^{-1}, \dots, \mathbf{z}^{-j}$, namely[2]

$$\tilde{\mathbf{d}}^j \leftarrow \mathbf{z}^{-j} - \sum_{l<j} \text{proj}_{\mathbf{d}^l}(\mathbf{z}^{-j}), \;\; \mathbf{d}^j \leftarrow \frac{\tilde{\mathbf{d}}^j}{\|\tilde{\mathbf{d}}^j\|_2}, \tag{10.8}$$

where

$$\text{proj}_{\mathbf{d}^l}(\mathbf{z}^{-j}) := \begin{cases} \mathbf{0}, & \text{if } \mathbf{d}^l = \mathbf{0}, \\ \frac{\langle \mathbf{z}^{-j}, \mathbf{d}^l \rangle}{\langle \mathbf{d}^l, \mathbf{d}^l \rangle} \mathbf{d}^l, & \text{otherwise} \end{cases}$$

is the orthogonal projection of $\mathbf{z}^{-j}$ onto the line generated by the atom $\mathbf{d}^l$.

[2] The version of the Gram-Schmidt process presented here is not to be implemented as stated, as is it known to suffer from numerical instability errors in finite-precision arithmetic. There exists equivalent versions (e.g Golub & Van Loan 1996] which alleviate these instabilities.

10.2.2 "Social" sparsity: simultaneous sparsity and smoothness via windowed group-Lasso

The first non-trivial results of our ramblings this far is obtained by considering the *social sparsity* [Kowalski et al., 2013, Kowalski and Torrésani, 2009] prior. In this model, a weakly activated voxel v in the middle of strongly activated voxels will be saved (as if rescued by the clan), whilst a strongly activated voxel in the middle of weakly activated voxels will be eliminated (as if killed by isolation). "Strongness" and "weakness" are measured with respect to a specified threshold $\alpha > 0$, which plays a rule similar to the

regularization parameter in Group-Lasso. The penalty imposes both sparsity and structure simultaneously! Formally, social sparsity corresponds to a penalty $g_{\text{social}} : \mathbb{R}^p \to \mathbb{R}$ defined implicitly via its proximal operator

$$\begin{aligned}(\text{prox}_{\alpha g_{\text{social}}}(\mathbf{z}))_v &:= z_v \begin{cases} 1 - \frac{\alpha}{\|\boldsymbol{\gamma}_v \bullet \mathbf{z}\|_2}, & \text{if } \|\boldsymbol{\gamma}_v \bullet \mathbf{z}\|_2 > \alpha, \\ 0, & \text{otherwise} \end{cases} \\ &= z_v \left(1 - \frac{\alpha}{\left(\sum_{s \in \mathcal{N}(v)} (\gamma_v^s)^2 z_s^2 \right)^{1/2}} \right)_+ , \end{aligned} \tag{10.9}$$

where $\boldsymbol{\gamma}_v \bullet \mathbf{z} := (\gamma_v^1 z_1, \gamma_v^2 z_2, \ldots, \gamma_v^p z_p) \in \mathbb{R}^p$ for weights $(\gamma_v^s)_{v, s \in [\![p]\!]}$ satisfying $\sum_v |\gamma_v^s|^2 = 1$ for all s, and $\mathcal{N}(v) := \{s \in [\![p]\!] | \gamma_v^s \neq 0\}$ is the *neighborhood* of the vth voxel, assumed to be non-empty. Thus each $\boldsymbol{\gamma}_v$ can be thought of as (normalized) mean-filter supported on a patch $\mathcal{N}(v)$ around the voxel v. Examples include rectangular filters, truncated Gaussians, etc.

One notes the following facts

- $\|\mathbf{E}\mathbf{z}\|_2^2 \equiv \|\mathbf{z}\|_2^2$, where $\mathbf{E}\mathbf{z} := (\boldsymbol{\gamma}_v \bullet \mathbf{z})_{v \in [\![p]\!]} \in \mathbb{R}^{p^2}$ is the *expansion operator* associated with the weights w_v^s. In other words, $\mathbf{E}$ is a *linear isometry*.
- social sparsity is related to Group-Lasso (GL) by the formula

$$\text{prox}_{\alpha g_{\text{social}}} = \mathbf{F} \circ \text{prox}_{\alpha \text{GL}} \circ \mathbf{E}, \tag{10.10}$$

  where $\mathbf{F}$ is the right pseudo-inverse of $\mathbf{E}$.

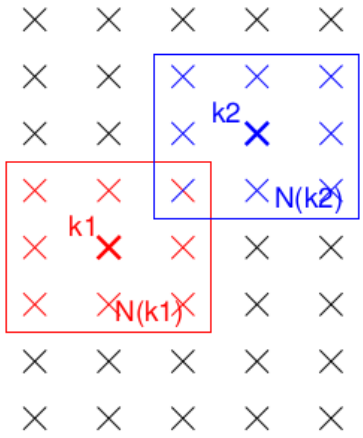


Figure 10.1: **Social sparsity** illustrated in 2D. The neighborhood of the coefficient k_1 is given by the red window, and the neighborhood of the coefficient k_2 by the blue one. These two neighborhoods share one coefficient. When considering the red group, coefficients are weighted by some weights $\boldsymbol{\gamma}_{k'}^{k_1} \neq 0$, $k' \in \mathcal{N}(k^1)$. Outside the red group, the weights are equal to zero. When considering the blue group, coefficients are weighted by some weights $\boldsymbol{\gamma}_{k'}^{k_2} \neq 0$, $k' \in \mathcal{N}(k^2)$. Adapted from [Kowalski et al., 2013].

10.3 Conclusion

These ideas have a great potential to extend the classical dictionary-learning technology providing the practitioner with a modeling framework incorporating a much larger class of constraints –namely proximable penalty functions– than is currently being done. As regards convergence of our proposed proximal online dictionary-learning scheme, direct application of [Fercoq and Richtárik, 2015] seems to suffice, since the general DL algorithm constructs unbiased estimates of $\nabla_j f(\mathbf{D}_t)$, where $f(\mathbf{D}) := \mathbb{E}_{\mathbf{x}} \min_{\mathbf{c} \in \mathbb{R}^k} \ell(\mathbf{D}\mathbf{c}, \mathbf{x})$. However, a more careful treatment is warranted, and is left for future work.

Bibliography

Olivier Fercoq and Peter Richtárik. Accelerated, parallel and proximal coordinate descent. *SIAM Journal on Optimization*, 25, 2015.

Matthieu Kowalski and Bruno Torrésani. Structured sparsity: From mixed norms to structured shrinkage. In *SPARS'09*, 2009.

Matthieu Kowalski, Kai Siedenburg, and Monika Dörfler. Social Sparsity! Neighborhood Systems Enrich Structured Shrinkage Operators. *Signal Processing, IEEE Transactions on*, 61(10), 2013.

J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online dictionary learning for sparse coding. *ICML*, 2009.

J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research*, 11, 2010.

Predicting task activation maps from task-free resting-state data

Contents

11.1 Introduction

Across-subject variability in organization is a hallmark of the human brain, that reflects genetic variability and is in turn reflected in behavioral differences. It has resisted so far modeling attempts, leading to blurred population-level anatomical templates and high-variance in functional representations across individuals. The only solution to defeat this variability is actually to condition individual representations on other data, for instance, mapping functional organization subject to anatomical constraints, or relevant features of brain organization, such as structural or functional connectivity [Saygin et al., 2011].

In neuroimaging and cognitive neuroscience, it is widely believed that the functional connectivity (FC) structure at rest remains grossly unchanged during task-stimulus presentation. This makes sense by least-action principle considerations: the brain does not need to rewire the functional links between regions upon presentation of a stimulus: it conserves the same networks as during rest, except that some edges are strengthened while others are weakened, to support the cognitive load of the particular task. Pushing this even further, one can claim that the resting-state FC of the brain predictively modulates the functional responses of the brain in the presence of task. Indeed, recently, resting-state fMRI has been shown to provide relevant constraints for functional mapping, opening the possibility to capture in standardized and cheaper acquisition most of the inter-individual differences [Tavor et al., 2016, Cole et al., 2016, Bzdok et al., 2016]. Possible applications include the improvement of population-level analyses, e.g. by finding better imputation schemes when dealing with missing data, detecting outlier data, and clarifying between-subjects similarities in comparison with genotyping or behavioral data. An important practical question has become how to optimize information transfer across these modalities to boost the chance of capturing the essence of inter-individual differences.

Our main contributions. In this work, we propose a general framework for the problem of predicting task fMRI activation maps from resting-state-only features. We present 2 main contributions: *(a)* the stacking of data across different random subsets subjects to reduce model-complexity and improve the prediction on held-out subjects, and *(b)* a multi-target regression approach to the predictive problem which better captures the functional inter-dependencies between different cognitive tasks. This generalizes and improves on the ideas in [Tavor et al., 2016]. We demonstrate the empirical gains brought by this approach through experiments on real datasets.

11.2 Feature extraction

The goal is to extract from resting-state data, pertinent features that encode the functional connectivity information in each voxel. A naïve choice would be to use the adjacency vector of each voxel in the whole-brain functional connectivity matrix. This is not practical due to the large number of (noisy) voxels, as it leads to enormous feature matrices. However, due to the inherent local correlations of data from different voxels, all this information is captured in the affinity of each voxel to a set of brain regions or networks. One way to get such profiles is to automatically learn a low-dimensional reduction of the resting-state data $\mathbf{X}_s \in \mathbb{R}^{n_s \times p}$ of each subject s into a common latent space, of dimension $k \ll \min(\min_s(n_s), p)$, as proposed in [Tavor et al., 2016]. Here, n_s is the number of TRs (Repetition Time) and p is the number of voxels.

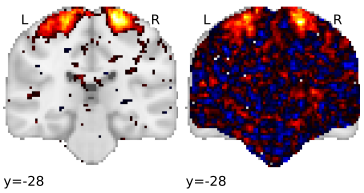


Figure 11.1: FC feature-extraction. **Left:** A component of the group dictionary. **Right:** Corresponding component for an individual subject's dictionary estimated using the proposed formula (11.2.2).

11.2.1 Dual regression

As before, let the n_s-by-p matrix $\mathbf{X}_s$ be the resting-state data for subject s and $\mathbf{D}$ be the k-by-p group-level dictionary (aka *topographic basis*) obtained

by stacking together resting-state time-series data from all the subjects and decomposing into k components of p voxels each by running a multi-subject decomposition algorithm like PCA, ICA, or dictionary-learning, etc. (more details on obtaining $\mathbf{D}$ later). We assume $\mathbf{D}$ to be under-complete – i.e $k \ll \min(\min_s(n_s), p)$– and therefore full-rank (i.e $\hat{\mathbf{D}}\hat{\mathbf{D}}^T$ is invertible).

The standard "dual-regression" procedure[Tavor et al., 2016] then proceeds as follows:

- Compute the n_s-by-k matrix of subject-to-components temporal dynamics $\mathbf{C}_s$ by regressing the group-level dictionary $\hat{\mathbf{D}}$ onto subject data $\mathbf{X}_s$:

$$\hat{\mathbf{C}}_s \in \text{argmin}_{\mathbf{C}_s \in \mathbb{R}^{n_s \times k}} \|\mathbf{X}_s - \mathbf{C}_s\hat{\mathbf{D}}\|^2_{\text{Fro}} \tag{11.1}$$

- Compute individual dictionary $\hat{\mathbf{D}}_s = (\hat{\mathbf{d}}^v_{s,j})_{1\le j\le k, 1\le v\le p} \in \mathbb{R}^{k\times p}$ by regressing the subject's resting-state data $\mathbf{X}_s \in \mathbb{R}^{n_s\times p}$ onto her subject-to-components temporal dynamics $\hat{\mathbf{C}}_s \in \mathbb{R}^{n_s\times k}$:

$$\hat{\mathbf{D}}_s \in \text{argmin}_{\mathbf{D}\in\mathbb{R}^{k\times p}} \|\mathbf{X}_s - \hat{\mathbf{C}}_s\mathbf{D}\|^2_{\text{Fro}} \tag{11.2}$$

The end result is that for each subject s and each voxel v, we obtain a k-dimensional encoding $\hat{\mathbf{d}}_{s,v} \in \mathbb{R}^k$ of the voxel's time-series $\mathbf{x}_{x,v} \in \mathbb{R}^{n_s}$ in a common group-level space. These are the features (see Fig. 11.1).

11.2.2 Using only a single regression step

For the standard "dual-regression" feature-extraction method[Tavor et al., 2016], a total of 2 regression steps are done (hence the name of the procedure). As a first (conceptual) improvement, we note that the individual dictionary $\hat{\mathbf{D}}_s = \hat{\mathbf{C}}_s^\dagger\mathbf{X}_s$ can be rewritten as

$$\begin{aligned}
\hat{\mathbf{D}}_s &= (\hat{\mathbf{C}}_s^T\hat{\mathbf{C}}_s)^{-1}\hat{\mathbf{C}}_s^T\mathbf{X}_s \\
&= ((\hat{\mathbf{D}}\hat{\mathbf{D}}^T)^{-1}\hat{\mathbf{D}}\mathbf{X}_s^T\mathbf{X}_s\hat{\mathbf{D}}^T(\hat{\mathbf{D}}\hat{\mathbf{D}}^T)^{-1})^{-1}(\hat{\mathbf{D}}\hat{\mathbf{D}}^T)^{-1}\hat{\mathbf{D}}\mathbf{X}_s^T\mathbf{X}_s \\
&= \hat{\mathbf{D}}\hat{\mathbf{D}}^T(\hat{\mathbf{D}}\mathbf{X}_s^T\mathbf{X}_s\hat{\mathbf{D}}^T)^{-1}\hat{\mathbf{D}}\mathbf{X}_s^T\mathbf{X}_s \\
&= \hat{\mathbf{D}}\hat{\mathbf{D}}^T\underbrace{\hat{\mathbf{D}}\mathbf{X}_s^T(\mathbf{X}_s\hat{\mathbf{D}}^T\hat{\mathbf{D}}\mathbf{X}_s^T)^\dagger\mathbf{X}_s}_{\text{OLS}(\mathbf{X}_s\hat{\mathbf{D}}^T,\mathbf{X}_s)}.
\end{aligned}$$

That is, we regress the subject's resting-state time-series data $\mathbf{X}_s$ onto n_s-by-k matrix $\hat{\mathbf{C}}_s := \mathbf{X}_s\hat{\mathbf{D}}^T$ and then reweight the result by the component-to-component covariance matrix $\hat{\mathbf{D}}\hat{\mathbf{D}}^T$ of the group-level dictionary. All in all, only a 1 regression step is needed.

11.2.3 Obtaining the global dictionary $\hat{\mathbf{D}}$

Since the resting-state time-series data are large (for example 1200 3D volumes of 2×10^5 voxels in each subject in the HCP –Human Connectome Project– dataset [van Essen et al., 2012]), a decomposition method that scales well is required. We use a variant of online dictionary-learning method [Mairal et al., 2010], a very fast implementation of which has been proposed in [Mensch et al., 2016], based on random *matrix sketching / sub-sampling*. Incremental PCA/ICA-based methods[Smith et al., 2014, Varoquaux et al., 2010] are also a competitive choice.

11.2.4 Relationship between dual-regression and hyper-alignment

It turns out that the *shared-response* "hyper-alignment" (HA) framework [Haxby et al., 2011] and the "dual regression" (DR) scheme [Tavor et al., 2016] we just presented are very closely related. Indeed, [Haxby et al., 2011] considers the following problem

$$\begin{aligned} &\text{minimize } \frac{1}{N}\sum_{s=1}^{N} \|\mathbf{X}_s - \mathbf{C}_s\mathbf{D}\|_{\text{Fro}}^2 \\ &\text{over } \mathbf{D} \in \mathbb{R}^{k\times p},\ \mathbf{C}_s \in \mathbb{R}^{n_s\times k}, \\ &\text{subject to } \mathbf{C}_s^T\mathbf{C}_s = \mathbf{I}_k,\ \forall s \in [1\dots N]. \end{aligned} \tag{11.3}$$

Without the orthonormality constraints "$\mathbf{C}_s^T\mathbf{C}_s = \mathbf{I}_k$", this problem is precisely the DR problem. (11.3) is usually solved via an alternating minimization scheme. Viz,

- **Update rotations (orthonormal Procrustes analysis):**

  $$\mathbf{C}_s^{(t+1)} = \mathbf{U}_s^{(t)}{\mathbf{V}_s^{(t)}}^T\ \forall s \in [1\dots N],$$

  where $\mathbf{U}_s^{(k)}\Sigma_s{\mathbf{V}_s^{(t)}}^T$ is the SVD of the n_s-by-k matrix $\mathbf{X}_s{\mathbf{D}^{(t)}}^T$.

- **Update shared-dictionary:**

  $$\mathbf{D}_s^{(t+1)} = \frac{1}{N}\sum_{s=1}^{N}\mathbf{D}_s^{(t+1)},$$

  where $\mathbf{D}_s^{(t+1)} := {\mathbf{C}_s^{(t+1)}}^T\mathbf{X}_s$.

  However,

- DR is much more attractive due to its low cost: HA performs an SVD per subject per iteration.
- HA is usually done parcel-wise (i.e locally) because, the orthonormality conditions are unreasonable globally (i.e full-brain).

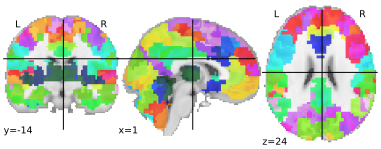


Figure 11.2: A **parcellation** is simply a collection of contiguous / simply-connected masks $\mathcal{P}$ called **parcels** (the colored patches) which cover the brain. Each **voxel** of the brain belongs / is assigned to exactly one parcel. In the parcellation shown here, each parcel contains approximately 1000 voxels.

11.3 Bags of low-rank multi-target linear models

We now develop our model for predicting subject-specific activation maps $\mathbf{Y}_s$ from resting-state features $\mathbf{D}_s$ (refer to section 11.2). Our model considers bootstraps of sub-samples of subjects instead of on a subject-by-subject basis enforces a reduction in the complexity of the model without loss in capacity. The idea is that the regression coefficients from predicting task activations from resting state should be partly shared across subjects. This reflects the hypothesis that the global cognitive organization of the brain should share some similarities across different subjects. Also, stacking across subjects as such corrects for *covariate-shift* [1] between different subjects, and facilitates *transfer-learning* from one-subject to another at test time.

Thus, for a bootstrap sub-sample $\mathcal{S}$ of b ($1 \leq b \leq N$) subjects, let $\mathbf{Z}_{\mathcal{S}} = [\mathbf{Z}_1|\dots|\mathbf{Z}_b] \in \mathbb{R}^{k\times p'b}$ be functional features masked over a parcel $\mathcal{P}$

[1] Grossly speaking, covariate-shift is a situation in which the distribution of the test set is not the same as the distribution of the training set, and so the model learned on the training set may not generalize to unseen (i.e test) data.

of $p' \leq p$ voxels (see Fig. 11.2) and horizontally stacked matrix of functional features. Similarly, let $\mathbf{Y}_{\mathcal{S}} \in \mathbb{R}^{p'b \times c}$ be corresponding activation maps to the c task contrasts, masked over the same parcel, and stacked vertically. The goal is to link these functional features with activations maps corresponding to c task-activation contrasts e.g. "Story-vs-Math", "Faces-vs-Houses", “2Back-vs-0Back”, etc.

11.3.1 Low-rank Ridge regression

Intra-subject activation maps for so-called different experimental stimuli may be correlated to one another. Indeed one would expect all brain activations to any conceivable experiment to be driven by a restricted set of latent causes, which is much less than the number of possible experiments. Thus, in an experiment with a sufficiently large bail of experimental conditions, one would expect that corresponding action maps would be correlated across different experimental conditions. Fitting a separate model per experimental condition would therefore be statistically inefficient due to model over-specification. We need a principled way to incorporate the covariance structure of the intra-subject activation maps into our predictive model. Low-rank linear models do just this. It produces a much smaller model (i.e few number of free parameters) which best explains the covariance structure between activation maps $\mathbf{Y}$ for the different conditions. This can be written as

$$\begin{aligned} &\text{Find } \boldsymbol{\beta}_{\mathcal{S}} \in \mathbb{R}^{k \times c}, \text{ with } \operatorname{rank}(\boldsymbol{\beta}_{\mathcal{S}}) \leq r, \\ &\text{s.t } \mathbf{Y}^j_{\mathcal{S}} \approx \mathbf{Z}^T_{\mathcal{S}} \boldsymbol{\beta}^j_{\mathcal{S}} \ \forall j \in [1 \dots c], \end{aligned} \tag{11.4}$$

for a chosen rank bound r, with $1 \leq r \leq \min(c, k)$. Here, $\mathbf{Y}^j_{\mathcal{S}} \in \mathbb{R}^{|\mathcal{S}|p \times 1}$ denotes the activation maps for contrast j, for the subjects in the bootstrap sub-population $\mathcal{S}$. The model (11.4) can be captured by the following convex program

$$\begin{aligned} &\text{minimize } \frac{1}{2} \|\mathbf{Y}_{\mathcal{S}} - \mathbf{Z}^T_{\mathcal{S}} \boldsymbol{\beta}_{\mathcal{S}}\|^2_{\text{Fro}} \text{ w.r.t } \boldsymbol{\beta}_{\mathcal{S}} \in \mathbb{R}^{k \times c} \\ &\text{subject to } \operatorname{rank}(\boldsymbol{\beta}_{\mathcal{S}}) \leq r. \end{aligned} \tag{11.5}$$

This defines a low-complexity linear model

$$\hat{f}_{\mathcal{S}} : \mathbf{Z}_s \mapsto \mathbf{Z}^T_s \hat{\boldsymbol{\beta}}_{\mathcal{S}} \tag{11.6}$$

for predictively linking resting-state data to individual activation maps over the parcel $\mathcal{P}$. The full-rank case $r = \min(c, k)$ together with the choice $b = 1$ (no bagging) corresponds to the subject-wise contrast-wise single-output linear regression model proposed in [Tavor et al., 2016].

Now, let $\hat{\mathbf{Y}}^{\text{OLS}}_{\mathcal{S}} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$ be the SVD (singular-value decomposition) of the least-squares prediction $\hat{\mathbf{Y}}^{\text{OLS}}_{\mathcal{S}} := \mathbf{Z}_{\mathcal{S}} \hat{\boldsymbol{\beta}}^{\text{OLS}}_{\mathcal{S}}$ where $\hat{\boldsymbol{\beta}}^{\text{OLS}}_{\mathcal{S}} := (\mathbf{Z}^T_{\mathcal{S}} \mathbf{Z}_{\mathcal{S}})^{\dagger} \mathbf{Z}^T_{\mathcal{S}} \mathbf{Y}_{\mathcal{S}}$ is the ordinary least-squares (OLS) solution to the unconstrained version of (11.5). Of course (11.5) may fail to have a unique solution. The following elementary lemma, whose proof (Supp. Mat.) follows directly from the *Eckart-Young-Mirsky theorem* [Carl and Gale, 2000] and the orthogonality property of the OLS fit, produces a solution for model (11.5). Viz,

Lemma 3. *A solution to* (11.5) *is given by* $\hat{\boldsymbol{\beta}}_{\mathcal{S}} = \hat{\boldsymbol{\beta}}_{\mathcal{S}}^{OLS}\boldsymbol{\Pi}_{\mathcal{S}}(r)$ *where* $\boldsymbol{\Pi}_{\mathcal{S}}(r) = \sum_{i=1}^{r}\mathbf{v}_i\mathbf{v}_i^T$ *is the orthogonal projector onto the subspace spanned by the first* r *principal singular vectors* $\mathbf{v}_{i\leq r}$ *of the OLS prediction* $\hat{\mathbf{Y}}_{\mathcal{S}}^{OLS}$.

It should be noted that the form of the solutions provided by the above lemma is particularly appealing: We only need to do a single fit to obtain a solution to problem (11.5) from solutions to the unconstrained OLS version of [Tavor et al., 2016].

Proof. Indeed, by the orthogonality property of least-squares estimates, we have the decomposition

$$\|\mathbf{Y}_{\mathcal{S}} - \mathbf{Z}_{\mathcal{S}}\boldsymbol{\beta}_{\mathcal{S}}\|_{\text{Fro}}^2 = \|\mathbf{Y}_{\mathcal{S}} - \hat{\mathbf{Y}}_{\mathcal{S}}^{\text{OLS}}\|_{\text{Fro}}^2 + \|\hat{\mathbf{Y}}_{\mathcal{S}}^{\text{OLS}} - \mathbf{Z}_{\mathcal{S}}\boldsymbol{\beta}_{\mathcal{S}}\|_{\text{Fro}}^2,$$

with the first summand being independent of the model parameters $\boldsymbol{\beta}_{\mathcal{S}}$. Thus (11.5) can be rewritten as

$$\begin{aligned} &\text{minimize } \frac{1}{2}\|\hat{\mathbf{Y}}_{\mathcal{S}}^{\text{OLS}} - \mathbf{Z}_{\mathcal{S}}\boldsymbol{\beta}_{\mathcal{S}}\|_{\text{Fro}}^2 \text{ w.r.t } \boldsymbol{\beta}_{\mathcal{S}} \in \mathbb{R}^{k\times c} \\ &\text{subject to } \text{rank}(\boldsymbol{\beta}_{\mathcal{S}}) \leq r. \end{aligned} \tag{11.7}$$

It is clear that $\hat{\boldsymbol{\beta}}_{\mathcal{S}}(r) := \hat{\boldsymbol{\beta}}_{\mathcal{S}}^{\text{OLS}}\boldsymbol{\Pi}_{\mathcal{S}}(r)$ is of rank at most r. One computes

$$\begin{aligned} \mathbf{Z}_{\mathcal{S}}\hat{\boldsymbol{\beta}}(r) &= \mathbf{Z}_{\mathcal{S}}\hat{\boldsymbol{\beta}}_{\mathcal{S}}^{\text{OLS}}\boldsymbol{\Pi}_{\mathcal{S}}(r) = \mathbf{Z}_{\mathcal{S}}\hat{\boldsymbol{\beta}}_{\mathcal{S}}^{\text{OLS}}\sum_{i=1}^{r}\mathbf{v}_i\mathbf{v}_i^T \\ &= \hat{\mathbf{Y}}_{\mathcal{S}}^{\text{OLS}}\sum_{i=1}^{r}\mathbf{v}_i\mathbf{v}_i^T, \end{aligned}$$

which, by the Eckart-Young-Mirsky theorem [Carl and Gale, 2000] for the Frobenius norm, is the best rank r approximation of $\hat{\mathbf{Y}}_{\mathcal{S}}^{\text{OLS}}$ w.r.t the Frobenius norm. □

11.4 Algorithms

11.4.1 Learning

Low-rank multi-output regression. For the estimation of the predictive model linking resting-state features to activation maps, the template model (11.5) is solved for each parcel and each bootstrap sub-sample of subjects, to obtain the coefficients for predicting individual subject activations for the different task contrasts, jointly. The estimation is **massively parallel**: it is done per bootstrap and per parcel.

11.4.2 Hyper-parameter tuning

The rank bound r can be selected via K-fold cross-validation: we would retain the smallest value or $r = \hat{r}_{\mathcal{S}}$ in the range $[1, \min(k, c))]$ which produces a cross-validation score within 1 standard deviation of the best score (the so-called *1 standard error rule*), or alternatively via leave-one-out (LOO) cross validation. However, cross-validation is very costly as multiple models must be fitted on different splits of the training data. Moreover it might not even be possible in the case limited training data.

Algorithm 5: Training model for predicting activation maps from resting-state

Require: • Data from N_{train} subjects. For each subject s we have precomputed spatial features $\mathbf{Z}_s \in \mathbb{R}^{k \times p}$.

• A set of brain parcellations (defined by sets of brain masks).

Ensure: Distributed sets of fitted models $\{\{\hat{f}_{\mathcal{S}} | \hat{f}_{\mathcal{S}} \in \mathcal{F}_{\mathcal{P}}\} | \mathcal{P} \in \text{ parcels}\}$, i.e one model $\hat{f}_{\mathcal{S}}$ per bootstrap sub-sample $\mathcal{S}$ per parcel $\mathcal{P}$.

1: **parallel for** each parcel $\mathcal{P}$ **do**
2: **parallel for** each bootstrap sub-sample of subjects $\mathcal{S}$ **do**
3: **Fit** a model $\hat{f}_{\mathcal{S}}$ from (11.5), for predicting $\mathbf{Y}_{\mathcal{S}}$ from $\mathbf{Z}_{\mathcal{S}}$ restricted on the parcel $\mathcal{P}$
4: **end pararell for**
5: **end pararell for**

Generalized cross-validation. A very attractive alternative to cross-validation is the so-called *generalized cross-validation (GCV)* [Golub et al., 1979], whereby one attempts to directly minimize (an unbiased estimate of) the generalization error, which in our case reduces to

$$GCV(r) := \frac{\|\mathbf{Y} - \hat{\mathbf{Y}}(r)\|^2_{\text{Fro}}}{(nc - \hat{df}(r))^2} \tag{11.8}$$

as a function of the rank parameter r. Hee, $\hat{df}(r)$ is an unbiased estimate of the number of *degrees of freedom*, that is, the number of free parameters needed to completely specify the linear prediction model given by the coefficients $\hat{\boldsymbol{\beta}}(r)$. One can show that GCV is the consistent asymptotic limit of *leave-one out (LOO)* cross-validation. However the advantage of GCV is that only one model needs to be fitted per value of the hyper-parameter (cf. LOO cross-validation, where as many models as sample points need to fitted per value of the hyper-parameter).

In our case of reduced rank linear regression, the approximation error term $\|\mathbf{Y} - \hat{\mathbf{Y}}(r)\|^2_{\text{Fro}}$ in (11.8) reduces to

$$\begin{aligned} \|\mathbf{Y} - \hat{\mathbf{Y}}(r)\|^2_{\text{Fro}} &= \|\mathbf{Y} - \hat{\mathbf{Y}}^{\text{OLS}}\|^2_{\text{Fro}} + \|\hat{\mathbf{Y}}^{\text{OLS}} - \hat{\mathbf{Y}}(r)\|^2_{\text{Fro}} \\ &= \|\mathbf{Y} - \hat{\mathbf{Y}}^{\text{OLS}}\|^2_{\text{Fro}} + \sum_{l=r+1}^{r_0} \sigma_l^2, \end{aligned} \tag{11.9}$$

where $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_{r_0}$ are nonzero singular values of $\hat{\mathbf{Y}}^{\text{OLS}}$ and $r_0 = \text{rank}(\hat{\mathbf{Y}}^{\text{OLS}})$. In [Mukherjee et al., 2015], finite-sample unbiased estimates for the degrees of freedom of rank-penalized models where derived. The authors established the formula

$$\hat{df}(r) = \hat{df}_{\text{naïve}}(r) + \underbrace{2 \sum_{k=1}^{r} \sum_{l=r+1}^{r_0} \frac{\sigma_l^2}{\sigma_k^2 - \sigma_l^2}}_{\text{bias correction}} \geq \hat{df}_{\text{naïve}}(r), \tag{11.10}$$

for any $r \in [1, r_0]$ with $\sigma_r > \sigma_{r+1}$ if $r < r_0$. The naïve estimate $\hat{df}_{\text{naïve}}(r) := kc - (k - r)(c - r)$ is simply the number of free parameters needed to com-

pletely specify a matrix of rank $r \in [1, r_0]$. Noting that

$$\frac{\sigma_l^2}{\sigma_k^2 - \sigma_l^2} \geq \frac{\sigma_l^2}{\sigma_k^2} \geq \frac{1}{\kappa(\hat{\mathbf{Y}}^{\text{OLS}})^2},$$

we get the trivial bound

$$\hat{df}(r) \geq \hat{df}_{\text{naïve}}(r) + \frac{2r(r_0 - r)}{\kappa(\hat{\mathbf{Y}}^{\text{OLS}})^2}, \tag{11.11}$$

where $\kappa(\hat{\mathbf{Y}}^{\text{OLS}}) := \sigma_1 / \sigma_{r_0}$ is the *condition number* of $\hat{\mathbf{Y}}^{\text{OLS}}$. Albeit, this bound is not very interesting for even mildly ill-conditioned $\hat{\mathbf{Y}}^{\text{OLS}}$ where $\kappa(\hat{\mathbf{Y}}^{\text{OLS}})^2 \gg 1$. Finally, we note that though $\hat{df}_{\text{naïve}}(r)$ under-estimates $\hat{df}(r)$ in (11.10) and (11.11), the former is already a very good approximation in practice, and in fact equals the latter (almost surely) in the asymptotic limit $n \to \infty$.

11.4.3 Inference

At prediction time, these different models are queried on held-out subjects and their results are aggregated by averaging. Such a divide-and-conquer approach allows us to learn complementary aspects of the data landscape, boosting prediction scores, while reducing the variance of the individual component models of which its is made. This is a well-known statistical property of bagging ensembles. The inference be done by making a single pass in Alg. 6.

Algorithm 6: Predicting activation maps from resting-state features

Require:
- Data from N_{test} subjects. For each subject s, we have precomputed spatial features $\tilde{\mathbf{X}}_s \in \mathbb{R}^{k \times p}$ using their resting-state data.
- Sets of fitted models $\hat{f}_{\mathcal{S}}$ (see Alg. 5).

Ensure: Predictions $\hat{\mathbf{Y}}_s \in \mathbb{R}^{p \times c}$, for each test subject s.

1: $\hat{\mathbf{Y}} \leftarrow \mathbf{0} \in \mathbb{R}^{N_{\text{test}} \times p \times c}$
2: **parallel for** each parcel $\mathcal{P}$ **do**
3: **parallel for** each trained model $\hat{f}_{\mathcal{S}}$ on $\mathcal{P}$ **do**
4: **parallel for** each test subject s **do**
5: **Predict** the activation maps of subject s with model $\hat{f}_{\mathcal{S}}$:

$$\hat{\mathbf{Y}}_{\mathcal{S}|\mathcal{P}} \leftarrow \hat{\mathbf{Y}}_{\mathcal{S}|\mathcal{P}} + \underbrace{\hat{f}_{\mathcal{S}}(\tilde{\mathbf{X}}_s|_{\mathcal{P}})}_{\text{contribution of } \hat{f}_{\mathcal{S}}}$$

6: **end pararell for**
7: **end pararell for**
8: **end pararell for**

11.5 Experiments

Our experiments were done on task fMRI data from 200 subjects from the HCP –Human Connectome Project– dataset [van Essen et al., 2012]. These task fMRI data were acquired in an attempt to assess major domains that sample the diversity of neural systems , including language processing (semantic and phonological processing) and working memory.

The activation maps $\mathbf{Y}$ to predict. This data includes task activation maps from General Linear Models (GLMs) [Friston et al., 1994] that show the activation of different brain voxels to different cognitive conditions / task contrasts, for each subject. For example of these conditions include "Math vs Story" (part of language task), and "2Back-vs-0Back" –or "2BK-vs-0BK" for short– (part of working memory task). For example, there are about 19 task contrasts activation maps –each containing $p = 2 \times 10^5$ voxels– per subject for the working memory protocol. For each subject s, this gives an output matrix $\mathbf{Y}_s \in \mathbb{R}^{p \times c}$, were c is the total number of contrasts considered. In our experiments, we only considered the language (3 contrasts) and working memory (19 contrasts), giving a total of $c = 22$ task contrasts.

Extracted resting-state only features. The data also comes shipped with resting-state fMRI data consisting of $n_s = 1200$ 3D volumes of $p = 2 \times 10^5$ voxels each, per subject, forming an n_s-by-p matrix $\mathbf{X}_s$. The feature extraction described in section 11.2 was then applied to transform each $\mathbf{X}_s$ into low-dimensional functional connectivity features $\tilde{\mathbf{X}}_s \in \mathbb{R}^{k \times p}$, with $k = 100$.

The setup. $N_{\text{train}} = 100$ subjects were used in Alg. 5 to fit an ensemble of models (section 11.3). We used parcellations in which each parcel was worth about 4000 voxels, for a total of about 60 parcels. $N_{\text{test}} = 100$ subjects were held out for evaluating the models predictions, computed via Alg. 6.

11.6 Results

Quantitative metrics. Fig. 11.3 shows confusion matrices (via Pearson correlation) of predicted against true activation maps. We see that a subject's predicted activation maps are consistently more similar to their true activation than to other subjects', reflected by the fact that the confusion matrices are strongly diagonal-dominant. This is even more true for our proposed method. The Fig. 11.4 shows box-plots of prediction R^2-score and Pearson correlation for the 47 distinct contrast of the HCP task fMRI dataset [van Essen et al., 2012]. We see that both the reference method [Tavor et al., 2016] and our proposed method successfully predict the subjects' activation maps well above chance, with our method doing much better.

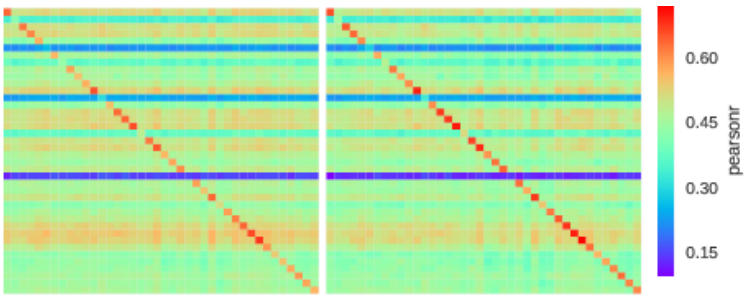


Figure 11.3: **Confusion matrices** for predicted versus true activation maps for the "Story vs Math" task contrast. The left plot corresponds to the reference method [Tavor et al., 2016] while the right one is for our proposed method. Higher diagonal values is better. The strong diagonal dorminance of these matrices reveals that the predicted maps of the subjects are more similar to their true maps than to other subjects.

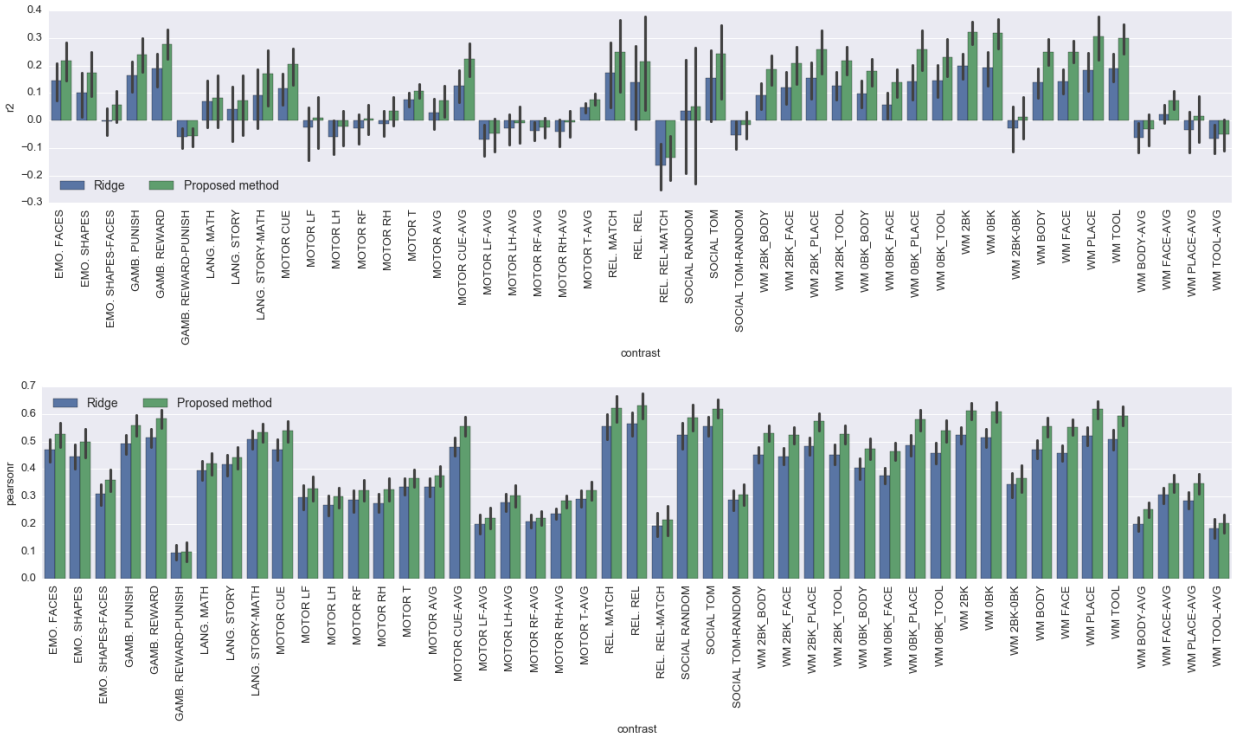


Figure 11.4: **Top:** R^2-score for predicting subject-specific activation maps for different task contrasts, from their resting data. These results are for the different contrasts of the HCP dataset [van Essen et al., 2012] are shown. Results for the reference method [Tavor et al., 2016] are also shown. **Bottom:** Pearson correlation for the same prediction problem.

Qualitative metrics. In Fig. 11.5, we display level-curves of the population mean (magenta) of activation maps for the "Story-vs-Math" and "2BK-vs-0BK" task contrasts, superimposed on the true activation maps of the subjects (the background image). The population mean activation map (magenta) is shown as a baseline (dummy predictor). We see that the contours for the predicted activation maps using our proposed method (green) faithfully follow topography of the true activation maps, indicating that the model successfully predicted the topography of the subjects' activation patterns for the contrasts.

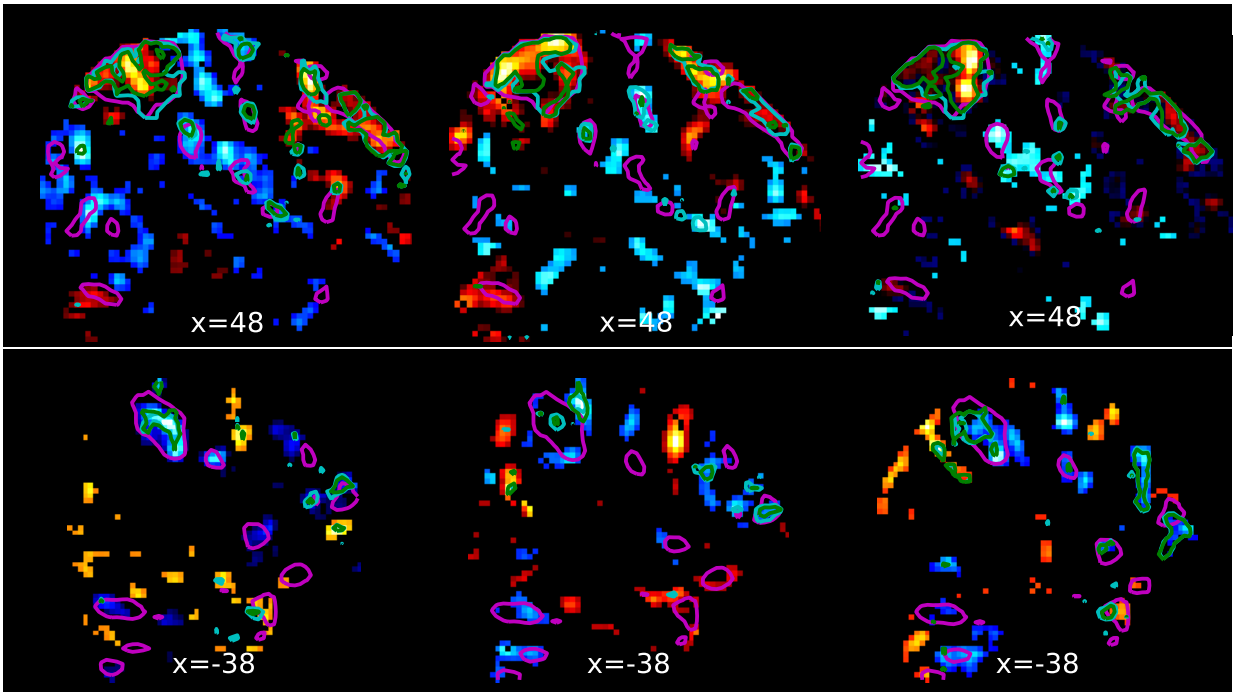


Figure 11.5: Level-curves of the population mean (magenta), predicted activation maps using our proposed method (green) and the reference method [Tavor et al., 2016] (cyan) for different contrasts. Each column represents a different subject (here 3), while each row represents a task contrast (here 2): first row is for "2BK-0BK" and second row is "Story-vs-Math".

11.7 Concluding remarks

We have proposed a general framework for the problem of predicting task fMRI activation maps from resting-state-only features. Our method creates an ensemble of parcel-wise low-rank multi-target linear models, over different random sub-populations of the training subjects to leverage the full richness of the data and jointly predict activation maps to different cognitive hypotheses (task contrasts). This is a major improvement over the state of the art [Tavor et al., 2016], as confirmed by extensive experiments on real data.

A practical implication of our results is that, for population studies, a large amount of information can be captured solely by a T1 image + resting-state fMRI: faster, cheaper scanning OR more control on data quality (imputation, outlier control). This explores new avenues for exploring the human

brain via resting-state data, in patients and healthy subjects alike.

Possible extensions with general approximators. Our work and also the previous works [Tavor et al., 2016, Cole et al., 2016, Bzdok et al., 2016] has shown beyond doubt, that there exists a predictive mapping from resting-state fMRI data $\mathbf{X}$ to task activation maps $\mathbf{Y}$. That is to say, the activation patterns in a person's brain during task are pre-determined by its background functional organization at rest. These works (including ours) have been limited to linear regression models, largely due to the simplicity of the latter. A priori, there is no reason why such a presumably complex relation should be accurately captured with a straight line, since there are probably many different layers of abstraction between functional connectivity patterns all the way through to activation patterns observed during task.

A simple extension would therefore be to replace the linear (Ridge) regression used to predict task activations from resting-state features, with a small *multi-layer perceptron (MLP)* [2], a cascade of linear transformations L_l, merged via simple non-linear functions σ_l like rectifier linear units (ReLU) or sigmoids

$$\mathbf{X} \overset{\sigma_1 \circ \mathbf{L}_1}{\longrightarrow} \dots \overset{\sigma_H \circ \mathbf{L}_H}{\rightarrow} \mathbf{Y}.$$

[2] An MLP can in principle approximate any "reasonable" function up to within arbitrary precision.

The intermediate representations extracted by such a model would be important in their own right. There would be enough data to fit such a model since voxels are the samples in this prediction problem, and there 2×10^5 voxels per subject. Indeed, a preliminary implementation of this generalization appears to give even much better prediction results (not presented here) than the improvements presented here over the state-of-the-art. This excursion will be continued in a future work.

Software. The code for the models presented in this chapter will be made publicly available online soon.

Bibliography

Danilo Bzdok, Gaël Varoquaux, Olivier Grisel, Michael Eickenberg, Cyril Poupon, and Bertrand Thirion. Formal Models of the Network Co-occurrence Underlying Mental Operations. *PLoS Comput Biol*, 12, 2016.

Eckart Carl and Young Gale. The Approximation of One Matrix by Another of Lower Rank. 2000.

Michael W. Cole, Takuya Ito, Danielle S. Bassett, and Douglas H. Schultz. Activity flow over resting-state networks shapes cognitive task activations. *Nat Neurosci*, 19, Dec 2016.

Karl J Friston, Andrew P Holmes, Keith J Worsley, J-P Poline, Chris D Frith, and Richard SJ Frackowiak. Statistical parametric maps in functional imaging: a general linear approach. *Human brain mapping*, 2(4):189–210, 1994.

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

J. V. Haxby, J. S. Guntupalli, A. C. Connolly, Y. O. Halchenko, B. R. Conroy, M. I. Gobbini, M. Hanke, and P. J. Ramadge. A common, high-dimensional model of the representational space in human ventral temporal cortex. *Neuron*, 72(2):404–416, Oct 2011.

J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research*, 11, 2010.

Arthur Mensch, Julien Mairal, Bertrand Thirion, and Gaël Varoquaux. Dictionary Learning for Massive Matrix Factorization. *arXiv:1605.00937*, 2016.

Ashin Mukherjee, Kun Chen, Naisyin Wang, and Ji Zhu. On the degrees of freedom of reduced-rank estimators in multivariate regression. *Biometrika*, 2015.

Z. M. Saygin, D. E. Osher, K. Koldewyn, G. Reynolds, J. D. Gabrieli, and R. R. Saxe. Anatomical connectivity patterns predict face selectivity in the fusiform gyrus. *Nat. Neurosci.*, 15, 2011.

Stephen M Smith, Aapo Hyvärinen, Gaël Varoquaux, Karla L Miller, and Christian F Beckmann. Group-PCA for very large fMRI datasets. *NeuroImage*, pages 738–749, 2014.

I Tavor, O Parker Jones, RB Mars, SM Smith, TE Behrens, and S Jbabdi. Task-free MRI predicts individual differences in brain activity during task performance. *Science*, 352(6282):216–220, 2016.

D.C. van Essen et al. The human connectome project: A data acquisition perspective. *NeuroImage*, 62(4), 2012.

Gaël Varoquaux, Sepideh Sadaghiani, Philippe Pinel, Andreas Kleinschmidt, Jean-Baptiste Poline, and Bertrand Thirion. A group model for stable multi-subject ICA on fMRI datasets. *Neuroimage*, 2010.

Part

Conclusion

Concluding remarks

Contents

12.1 Summary of main contributions

This thesis kicked-off with the goal of developing methods for modeling inter-subject functional variability, the aim being to enhance the estimation of functional connectomes –data-driven regions of interest, connectivity matrices, etc.– across populations of subjects. Below we summarize some of my major contributions.

12.1.1 Scientific contributions

The quest led to the study of, and proposal of methods for, structured (sparsity, smoothness, etc.) multi-variate models for brain encoding / decoding [Dohmatob et al., 2015b, Abraham et al., 2014, Eickenberg et al., 2015, Pellé et al., 2016]. Nonlinear registration of functional brain images also came up as a natural concern, and we contributed a method for direct registration of functional brain images [Dohmatob et al., 2016a] (submitted to *Neuroimage* journal).

We also improved the current state-of-the-art in ROI extraction and dimensionality reduction by combining techniques from online learning and structured sparsity (like TV-L1) to propose a novel scalable dictionary-learning framework for obtaining decompositions of brain images, which are closer to known neuro-biological organization of the brain: networks made of spatially localized smooth components with sharp boundaries.

The ultimate indicator for having understood a phenomenon is being able to recreate it, at least approximately. Indeed, Feynman once said, *"What I cannot (re)create, I do not understand!"* By combing techniques in generative modeling and ensembles, we improved state-of-the-art methods for predicting task-based activation maps (at the individual level!) from resting-state fMRI data, with accuracy well above chance. This work is being prepared for journal submission.

Due to the intimate relationship between modeling and optimization, the bulk of this work was made possible by development of new or improvement of existing methods of optimization, with scalability and robustness at heart [Dohmatob et al., 2014, 2015a, Varoquaux et al., 2015, Dohmatob, 2016].

Finally, some of the work done in the thesis have cross-fertilized other collaborative papers like [Rahim et al., 2015, Thirion et al., 2014].

12.1.2 Software contributions

While preparing this PhD project, I have made contributions to numerous open-source projects, including:

- *Nilearn* `http://nilearn.github.io/index.html`: Python package for leveraging machine learning algorithm in neuro-imaging. For example, the multi-variate models presented in chapter 3 are implemented as part of this package.
- *Pypreprocess* https://github.com/neurospin/pypreprocess: Python scripts scripts for preprocessing and QA of MRI data.
- *Nistats* `https://github.com/nistats/nistats`: Python package for statistical analyis (GLM, permutation tests, etc.) on MRI data.

12.2 Ongoing work and future directions

Unified view on structured models for brain data. We are preparing journal paper synthesizing all our contributions in the regarding structured models for brain decoding and segmentation presented in chapter 3. This will bring these methods to the doorsteps of the neuroscience practitioner.

Non-linear generative models for inter-subject brain data and prediction of task-fMRI activity from resting-state data. As concerns the modelling of inter-subject variability (chapters 9 and 11), most of the work done in this thesis can be cast in a more flexible framework of generative encoder-less models (see Fig. 9.6, for example)[1], with the space of parameters carefully constrained to ensure tractability and intepretability.

[1] Since the encoding representation is gotten by simply minimizing a reconstruction loss between the generated and the true brain image.

Synthèse en français

La thèse à démarrée avec l'objectif de développer des nouvelles méthodes pour la modélisation de la variabilité inter-sujet fonctionnelle, le but ultime étant l'amélioration de l'estimation de connectômes fonctionnelles sur des populations de sujets (chapitre 2).

Cette quête à conduit à la proposition des méthodes de pénalisation structurée (parcimonie, variation totale, etc.) multi-variées pour l'*encoding / decoding* [Dohmatob et al., 2015b, Abraham et al., 2014, Eickenberg et al., 2015, Pellé et al., 2016]. Le récalage fonctionnel est aussi souvenu naturellement, est nous avons contribué une méthode pour le récalage directe des images fonctionnelles (EPI) vers un cerveau standard (*template*) [Dohmatob et al., 2016a] (soumit au *Frontiers*). Voir chapitres 3, 4, 5, 6, 8, et 7.

Nous avons aussi amélioré l'état de l'art sur l'extraction de régions d'intérêt et la réduction de dimension en neuro-imagérie, par des méthodes combinant des techniques d'apprentissage en ligne et de parcimonie structurée (par exemple avec les pénalités TV-L1). Notre proposition est une nouvelle technique d'apprentissage de dictionnaire pour la décomposition d'images de cerveau plus conformes avec des a priori neurobiologique sur l'organisation fonctionnelle du cerveau: des réseaux spatialement localisés avec des contours bien délimités. Il s'agit d'un modèle génératif de base dimension, encodant succinctement la variabilité inter-sujet. Nous referons le lecteur aux chapitres 9 et 10.

Finalement, nous nous sommes intéressés à l'utilisation de techniques d'apprentissage supervisé pour expliquer la relation entre l'activité spontanée (activité au repos) et les enregistrements avec stimulations (activité évoquée dans des conditions précises). Nous avons proposé une méthode couplant un apprentissage non-supervisé (de type réduction de dimension par apprentissage de dictionnaire partagé) et des modèles prédictifs de faible rang pour exploiter les interdépendances entre les différentes fonctions cognitives. Les expériences numériques réalisées (200 sujets du projet HCP [van Essen et al., 2012]) montrent que nous apportons une amélioration considérable à l'état de l'art. Voir chapitre 11.

Les travaux réalisés on donné lieu à des nombreuses publication à des conférences et journaux tels que NIPS, ICASSP, MICCAI, Frontiers in Neurosciences, etc. Une liste complète des publications peut être consulter ma page Google scholar `https://scholar.google.fr/citations?user=FDWgJY8AAAAJ&hl=fr`. En chiffres

- Citations ≥ 194.
- Nombre total de publications ≥ 15.
- h index ≥ 4.
- 110 index ≥ 3,

dont

- Parcimonie et régularisation spatiale: [Dohmatob et al., 2014], [Dohmatob et al., 2015b], [Abraham et al., 2014], [Eickenberg et al., 2015], [Pellé et al., 2016]
- Récalage: [Dohmatob et al., 2016a]
- Optimisation: [Dohmatob et al., 2015a], [Varoquaux et al., 2015], [Dohmatob, 2016]
- Modelisation de variabilité fonctionnelle inter-sujet: [Dohmatob et al., 2016b]
- Neurosciences: [Rahim et al., 2015], [Thirion et al., 2014]

Il y a aussi des manuscrits en cours de préparation pour être publier dans des journaux:

- Vue globale sur la parcimonie et régularisation spatiale en neuro-imagérie: *"Structured penalties for brain decomposition and decoding: a unified view"*, pour *Neuroimage*
- *"Inter-subject registration of functional images: do we need anatomical images ?"* , pour *Frontiers*
- *"Enhanced prediction of task-based activation maps from resting-state data"*, pour *Neuroimage*

Logicielles contribuées. Pendant la préparation de la thèse, des nombreuses contributions dans de projets open-source on été réalisées. Pour en citer quelques unes:

- *Nilearn* `http://nilearn.github.io/index.html`: Librairie Python pour l'apprentissage statistique pour la neuro-imagerie. Par exemple les méthodes multi-variées présentées au chapitre 3 font partir des modules de cette librairie.
- *Pypreprocess* https://github.com/neurospin/pypreprocess: Des scripts Python pour le pré-traitement d'images IRMf.

- *Nistats* `https://github.com/nistats/nistats`: Outils d'analyse statistique en Python, pour les données la neuro-imagérie.

Bibliography

Alexandre Abraham, Elvis Dohmatob, Bertrand Thirion, Dimitris Samaras, and Gael Varoquaux. Region segmentation for sparse decompositions: better brain parcellations from rest fMRI. In *Sparsity Techniques in Medical Imaging*, 2014.

Elvis Dohmatob. A simple algorithm for computing Nash-equilibria in incomplete information games. In *OPT2016 – NIPS workshop on optimization for machine learning*, 2016.

Elvis Dohmatob, Alexandre Gramfort, Bertrand Thirion, and Gaël Varoquaux. Benchmarking solvers for TV-l1 least-squares and logistic regression in brain imaging. In *PRNI*. IEEE, 2014.

Elvis Dohmatob, Michael Eickenberg, Bertrand Thirion, and Gael Varoquaux. Local Q-Linear Convergence and Finite-time Active Set Identification of ADMM on a Class of Penalized Regression Problems. In *ICASSP 2016*, 2015a.

Elvis Dohmatob, Michael Eickenberg, Bertrand Thirion, and Gaël Varoquaux. Speeding-up model-selection in GraphNet via early-stopping and univariate feature-screening. In *Pattern Recognition in NeuroImaging (PRNI), 2015 International Workshop on*. IEEE, 2015b.

Elvis Dohmatob, , Gaël Varoquaux, and Bertrand Thirion. Inter-subject highres EPI-to-EPI direct nonlinear registration outperforms classical T1-based method. In *Annual meeting of the Organization for Human Brain Mapping - 2016*, 2016a.

Elvis Dohmatob, Arthur Mensch, Gaël Varoquaux, and Thirion Bertrand. Learning brain regions via large-scale online structured sparse dictionary-learning. In *NIPS*, 2016b.

Michael Eickenberg, Elvis Dohmatob, Bertrand Thirion, and Gaël Varoquaux. Total Variation meets Sparsity: statistical learning with segmenting penalties. In *MICCAI*. 2015.

Hubert Pellé, Philippe Ciuciu, Mehdi Rahim, Elvis Dohmatob, Patrice Abry, and Virginie Van Wassenhove. Multivariate Hurst exponent estimation in fMRI. Application to brain decoding of perceptual learning. In *13th IEEE International Symposium on Biomedical Imaging*, 2016.

Mehdi Rahim, Bertrand Thirion, Alexandre Abraham, Michael Eickenberg, Elvis Dohmatob, Claude Comtat, and Gael Varoquaux. Integrating Multimodal Priors in Predictive Models for the Functional Characterization of Alzheimer's Disease. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer International Publishing, 2015.

Bertrand Thirion, Gaël Varoquaux, Elvis Dohmatob, and Jean-Baptiste Poline. Which fMRI clustering gives good brain parcellations? *Frontiers in neuroscience*, 8, 2014.

D.C. van Essen et al. The human connectome project: A data acquisition perspective. *NeuroImage*, 62(4), 2012.

Gaël Varoquaux, Michael Eickenberg, Elvis Dohmatob, and Bertand Thirion. FAASTA: A fast solver for total-variation regularization of ill-conditioned problems with application to brain imaging. *arXiv:1512.06999*, 2015.