Neuroimaging Signatures of Brain Disorders: Fighting Overfitting in Predictive Models

Edouard Duchesnay

universite PARIS-SACLAY

Memoire d'Habilitation à Diriger des Recherches (HDR)

09 July 2020 – version 1.0

Abstract

This manuscript summarizes research in designing machine learning models to discover brain imaging signatures of mental disorders. We explore dimension reduction and regularization strategies to overcome the “curse of dimensionality” caused by a large number of neuroimaging measurements. Given the limitations of sparse models to produce stable and interpretable predictive signatures, we propose to push forward regularization by integrating spatial constraints. Evaluations on experimental data demonstrated that those constraints force the solution to adhere to biological priors, producing a more plausible interpretable predictive brain signature of clinical status. To bridge the gap between biological processes and brain imaging, we present multivariate latent variable sparse models to investigate the genetic influence on the brain.

Résumé

Ce manuscrit résume les recherches sur la conception de modèles d’apprentissage automatique pour découvrir les signatures en imagerie cérébrale des troubles mentaux. Nous explorons les stratégies de réduction de dimensions et de régularisation pour résoudre la "malédiction de la dimensionnalité" causée par le grand nombre de mesures de neuroimagerie. Étant donné les limites des modèles parcimonieux à produire des signatures prédictives stables et interprétables, nous proposons de pousser la régularisation en intégrant des contraintes spatiales. Les évaluations sur données expérimentales ont montré que ces contraintes obligent la solution à adhérer à des a priori biologiques, produisant une signature cérébrale prédictive de l’état clinique plus plausible et plus interprétable. Pour combler le fossé entre les processus biologiques et l’imagerie cérébrale, nous présentons des modèles à variables latentes multivariées parcimonieux pour étudier l’influence génétique sur le cerveau.

This manuscript was written during the unprecedented and bright spring of 2020. To my family: Julie, Margaux, Manon and Augustine.

Research summary

Brain anatomy and functioning are modeled by the individual's genetic and environmental background. Specialized behavior or certain genes may be associated with a commensurately greater allocation of the neural circuitry in the corresponding brain centers (Draganski et al., 2004; Maguire et al., 2000; Zatorre, Fields, and Johansen-Berg, 2012). Conversely, in neurodegenerative disorder (Frisoni et al., 2010), the clinical symptoms are thought to reflect differential patterns of atrophy. While neuroimaging is now widely accepted in neurology, its potential in psychiatry (Goodkind et al., 2015) is still emerging: brain developmental effect of genes (Leonard, Eckert, and Kuldau, 2006) or environment (Teicher et al., 2014) could be observed in neuroimaging. Those findings open a way to advance a biologically grounded re-definition of major psychiatric disorders.

Chap. 1: Design of machine learning algorithms for neuroimaging

My goal, 15 years ago, was to design artificial intelligence algorithms that learn to predict an output (the clinical outcome) given the hundreds of thousands of brain imaging features. This chapter provides the background of neuroimaging and Multivariate machine learning models. Those models leverage the capacity to capture complex brain patterns offering new opportunities such as making predictions at the individual level.

Chap. 2: Fighting overfitting

The core problem emerges from the unfavorable ratio between the small number of subjects ($\approx 10^2$) and a large number of neuroimaging features ($\geq 10^5$). This situation, a.k.a as the "curse of dimensionality," misleads the algorithm resulting in overfitting the training data and producing near-chance predictions on an independent sample.

This issue became the common thread of my research: Chapter 2 presents and evaluates the predictive performances of several designs combining regularized models with feature selection. Those designs were used to predict (i) sex and brain asymmetries from the cortical folding patterns (Duchesnay et al., 2007); (ii) visual stimuli from functional activation maps (Thirion et al., 2006) and (iii) the clinical status of patient with autism using PET imaging (Duchesnay et al., 2011).

Supervision: In 2007-2010, together with J.B. Poline, I supervised Cécilia Damon's PhD (entitled: *Réduction de Dimension et Régularisation* *pour l'Apprentissage Statistique et la Prédiction Individuelle en Irmf*) on the problems of regularization and size reduction applied to fMRI.

Adjusting research strategy

This application to psychiatry inspired the purpose of my methodological developments: machine learning has the potential to retrieve cerebral signatures of mental illnesses. This experience also drew two methodological limitations:

1. The biological (millimetric) scale of neuroimaging is not informative about the underlying physiopathological process.
2. Standard models analyze brain images as they would do for customer indicators to predict internet purchases. They do not permit the integration of any biological knowledge about the brain, even the simple spatial organization of the measurements from a scanner is utterly ignored.

Chap. 3: Imaging genetics

This Chapter presents multivariate models to address the first limitation, i.e., investigating the potential of neuroimaging to provide useful information to understand the biological processes that underpin the disorders. Brain imaging is increasingly recognized as an intermediate phenotype to understand the complex path between genetics (molecular) and behavioral or clinical phenotypes. In this context, the first goal was to propose ML algorithms to identify the part of genetic variability that explains some neuroimaging variability. The identified genetic variability would point to biological processes.

In Le Floch et al., 2012, we investigated the efficacy of different strategies of regularization and dimension reduction techniques combined with multivariate latent variable models to face the very high dimensionality of imaging genetics studies. A comparison of the strategies on a simulated dataset showed that univariate filtering combined with (ℓ_1, ℓ_2) regularized Partial Least Squares (PLS), outperformed other approaches. Then, we applied the chosen strategy on a real dataset composed of 94 subjects, around 600,000 genetic measurements (Single Nucleotide Polymorphisms, SNPs) and 34 functional MRI lateralization indexes measured during reading and speech comprehension tasks. We identified a genetic signature that explains the brain activation involved in language processing.

Supervision: In 2009-2012, I supervised Edith Le Floch's PhD (entitled: *Multivariate methods for the joint analysis of neuroimaging and genetic data*).

Chap. 4: Integrating spatial regularization

Then, I addressed the second limitation of classical sparse algorithms to integrate biological structure to produce stable and interpretable predictive signatures. I initiated a research program (funded by ANR BrainOmics, obtained together with V. Frouin) that pushed forward the regularization approaches by extending models with structural constraints issued from the known biological structure (spatial structure of the brain and the linkage disequilibrium or pathways of OMICs data). The aim is to constraint the solution to adhere to biological priors, producing more plausible interpretable solutions.

Adding those new penalties raised complex minimization problems. Existing solvers were either limited in the functions they can minimize or in their practical capacity to scale to "real life" high-dimensional data and to process meshes of the cortical surface. One outcome of this research program was a solver, called CONESTA, (Hadj-Selem et al., 2018) and the corresponding Python library (ParsimonY), which extends scikit-learn, for high dimensional structured input data such as 3D images and meshes of the cortical surface. The first result of this methodological effort was underwhelming: we could not demonstrate the relevance of such an approach to DNA data. It could be a promising track to follow with functional genomic measurements (such as RNA). Nevertheless, the second result demonstrated, beyond expectation, the relevance of spatial regularization (using Total Variation, TV) for neuroimaging (3D images and cortical meshes). Given the versatility of the proposed solver, we use it to integrate structured sparsity in other multivariate analysis methods. In de Pierrefeu et al., 2018c we demonstrated that the popular PCA (Principal Component Analysis) could be extended with spatial regularization to identify interpretable patterns of the neuroimaging variability in either functional or anatomical meshes of the cortical surface.

Supervision:

- 2013-2015: Fouad Hadj Selem, post-doc, mathematical foundations of CONESTA.
- 2013-2015: Tommy Lofstedt, post-doc, implementation of the Parsimony library.
- 2013-2015: Mathieu Dubois, post-doc, application to ADNI dataset.

Chap. 5: Identification of predictive signatures of brain disorders

Application to psychiatry:

- In de Pierrefeu et al., 2018b a spatially regularized supervised model (ElasticNet-TV) could identify an interpretable functional predictive signature (clusters in speech-related brain regions) of the upcoming hallucinations in patients with schizophrenia, offering perspectives in bio-feedback. Here, classical linear Support Vector Machine produced a useless signature (Fig 5.9), with signal across the whole-brain and significantly lower predictive power.

- In de Pierrefeu et al., 2018a, we demonstrated that spatial regularization (ElasticNet-TV), working at a voxel level, could identify a neuroanatomical signature of Schizophrenia (SCZ), reproducible across sites and more importantly to an earlier stage of the disorder. We demonstrated that spatial regularization provides a qualitative breakthrough (Fig 5.3) in terms of support recovery of the predictive brain regions.

Application to neurology:

- In Duchesnay et al., 2018 we used spatially regularized PCA to identify spatial patterns of white matter hyperintensities (WMH) in patients with CADASIL syndrome. We found two distinct patterns: a first subcortical pattern of WMH associated better clinical outcomes and a second pattern lesions in the deep white matter (pyramidal tracts and forceps minor) that is associated with clinical worsening. This finding suggests different mechanisms of small vessel disease (CADASIL) and clinical consequences.

Supervision:

- 2009-2012: Amicie de Pierrefeu, PhD: *Machine Learning with Structured Sparsity: Application to Neuroimaging-based Phenotyping in Schizophrenia*.

- 2017-2018: Pauline Favre, Post-doc, together with JF Mangin and J Houenou.

Contents

Background in machine-learning in neuroimaging

This chapter presents the background in neuroimaging (Sec. 1.1) and the processing to produce input features of the machine learning algorithms. Then, we summarize the principles of supervised Machine Learning algorithms (Sec. 1.2).

1.1 Neuroimaging features

The success of machine learning relies on the features used to represent the information contained in the brain images. Each MRI brain scan is composed of thousands of 3D volumetric units called voxels, in which the local anatomical or functional information is measured. However, a certain number of pre-processing steps are required before group analysis. Stacking subject's preprocessing outputs yields to a $(N \times P)$ data matrix $(\boldsymbol{X})$ containing the P features for N subject.

1.1.1 Structural MRI

sMRI uses the phenomenon of nuclear magnetic resonance (NMR) of the hydrogen atom to produce high-resolution, detailed images of internal brain structures and tissues. The strength of the magnetic field determines the resolution of the images. sMRI provides good contrast between gray matter and white matter. Three main preprocessing can be used: voxel-based grey matter density, vertex-based cortical thickness and region of interest-based measurements, each reflecting different aspects of the brain anatomy.

Voxel based morphometry (VBM)

VBM Fig. 1.1, described in (Ashburner, 2007), generally include 3 main steps: Segmentation (see Fellhauer et al., 2015 for a comparison), normalization (see Klein et al., 2009 for a comparison) and Modulation. Briefly, the sMRI images are first segmented into Gray Matter (GM), White Matter (WM), and Cerebrospinal Fluid (CSF). The second step is crucial to achieve spatial correspondence of voxels across subjects: All brain images are normalized into a common standard space. This normalization composes two transformations: (i) a linear transformation that accounts for global alignment (rotation, translation, and global brain size); (ii) and, a non-linear deformation that

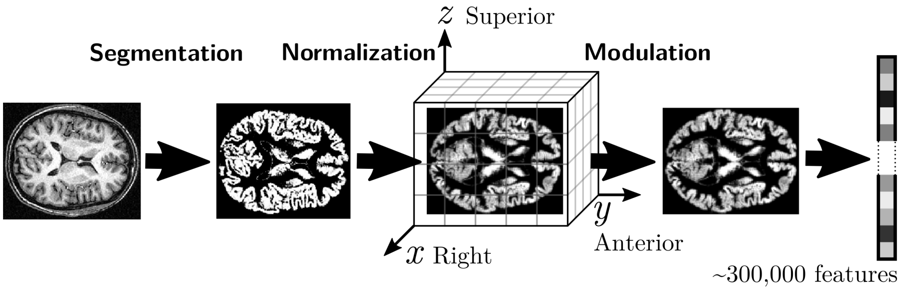


Figure 1.1: VBM-based feature extraction.

locally aligns brain structures. (iii) All the normalized images are finally modulated by the Jacobian of their transformation. This enables to preserve the quantity of tissue. No spatial smoothing is conducted. The validity of modulation to detect mesoscopic (i. e.between microscopic and macroscopic) abnormalities is discussed in Radua et al., 2014. If the global brain size is not of interest, one should modulate for non-linear effects only, or apply a proportional scaling according to the individual Total Intracranial Volume (TIV), as post-processing, to fully modulated images.

Recently, we decided to do the pre-processing with Computational Anatomy Toolbox (CAT). This toolbox of Statistical Parametric Mapping (SPM) uses a modified segmentation procedure reducing the role of tissue priors. Although, it uses DARTEL for the normalization, CAT, uses existing DARTEL templates in MNI space. Therefore, the creation of study-specific DARTEL templates is no longer necessary for most studies. This may seem somewhat sub-optimal, however, good performances have been reported Farokhian et al., 2017 and the use of the same template for all studies offers the possibility to pool data across studies (transfer learning, etc.).

VBM preprocessing produces hundreds of thousands (typically 300,000 GM voxels at 1.5mm^3 resolution) of features representing the local GM volume at each voxel. One advantage of VBM is that it is not restricted to a specific brain region, such as region-of-interest (ROI) analysis (described below) that requires a priori assumptions.

Cortical surface based morphometry

The goal is to obtain a measurement of the cortical thickness at each vertex of the cortical surface of the brain. The cortical thickness directly characterizes the amount of cortex atrophy. The measurements of cortical thickness are realized with Freesurfer software. All cortical thickness maps are registered on the default template of Freesurfer. Thus, the dimensionality of the vertex-based features is very high ($\approx 300,000$), since it corresponds to the number of vertices on the cortical mesh of the brain.

Regions-of-interest (ROIs)

Neuroimaging Softwares (Freesurfer, etc.) can segment the brain into cortical parcels and subcortical regions using atlases (Desikian, etc.). They automatically extract measurements on those ROIs: Cortical thickness and volume of subcortical regions. Compared to voxel-based and vertex-based approach, the number of features yielded by ROIs-based approaches is limited (≤ 100).

Cortical sulci based structural morphometry

Most automatic brain morphometry approaches are based on a point-by-point (voxel or mesh node) strategy in which each brain image is warped towards a reference coordinate system. Sulci based structural morphometry, provide by brainVISA software, proposes a complementary approach producing structural representations (6,000 morphometric features per brain) of cortical shapes to overcome the complex consequences of non-linear spatial normalization in population comparisons (Mangin et al., 2004).

1.1.2 Functional MRI

Functional Magnetic Resonance Imaging (fMRI) monitors local brain activity. fMRI exploits the local variations in the blood oxygen level. Indeed, it indirectly tracks brain activity by measuring the blood-oxygen-level-dependent (BOLD) signal (Poldrack, Mumford, and Nichols, 2011), which reflects the amount of brain activity. When a brain region becomes active, the amount of blood flow through that specific local area is increased. It subsequently leads to a relative surplus in local blood oxygen. This variation in the level of oxygenated blood induces a change in the local magnetic field and thus affects the MR signal. fMRI data is typically composed of temporal sequences of 3D images acquired every 2 to 3 seconds. Spatial resolution is usually $3mm^3$ when acquired with 3 Tesla (T) scanners.

Preprocessing steps

First step is the slice timing correction that temporally realigns the slices of each 3D volume. Second, the motion correction step allows spatial realignment between each 3D volume acquired at different point in time. It allows filtering out potential movement of the subject within the scanner. Third step, is the co-registration of each 3D fMRI volume acquires with the anatomical image of the subjects (the sMRI). The last step is the normalization of each subject in the common brain template.

General Linear Model

Once the fMRI time series are preprocessed, features can be extracted from the images. The most used approach is the General Linear Model (GLM) (Friston et al., 1994). The idea is to regress the signal of each individual voxel independently, onto a set of regressors explaining the experimental design (such as condition/task). Therefore, for each voxel, regression coefficients associated with each regressor are computed. Thus, different activation maps can be derived, corresponding to each condition/task. Those activation maps are used for subsequent statistical inferences. Usually, in fMRI studies, we want to test an effect of interest, to identify voxels that are significantly activated in condition A compared to condition B. This is answered by conducting a contrast between the activation map yielded under condition A, and the activation map yielded in condition B. Such (contrast)activation maps typically contain few tens of thousands of features.

1.2 Supervised machine learning

Machine learning (ML) is a term that encompasses a collection of methods to uncover patterns in data to perform trustworthy future predictions on new data. There are two mains methodological approaches: supervised and unsupervised machine learning algorithms. In supervised learning, the goal is to find a mapping from input data to an output target. In unsupervised learning, the objective is to identify inherent structures in the data to either: (i) classify data into homogeneous subgroups (clustering) or (ii) represent data with few features (feature extraction).

Given a training set of N samples, $D = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)\}$, where $\boldsymbol{x}_i$ is a multidimensional input vector with dimension P and y_i is the target to be predicted, class label in case of classification and a continuous value in case of regression. Input vectors are stacked in an input $N \times P$ matrix $\boldsymbol{X}$ and output N-dimensional vector $\boldsymbol{y}$.

1.2.1 Linear models

Linear models learn a predictive function $f(\boldsymbol{x}) = f(\sigma(\boldsymbol{x}^T\boldsymbol{w}))$ parameterized by a vector of weights or coefficients $\boldsymbol{w}$ that performs a linear combination of the input variables, $\boldsymbol{x}^T\boldsymbol{w}$ followed by an optional activation function $\sigma(.)$. This step performs a *projection* or a rotation of input sample into a good predictive one-dimensional sub-space.

The vector of parameters ($\boldsymbol{w}$), is obtained by minimizing an *objective function* $J(\boldsymbol{w})$ that is a sum of a *loss function* $L(\boldsymbol{w})$ and an optional regularization term (penalties on the weights vector) $\Omega(\boldsymbol{w})$ discussed in Sec. 2.2.

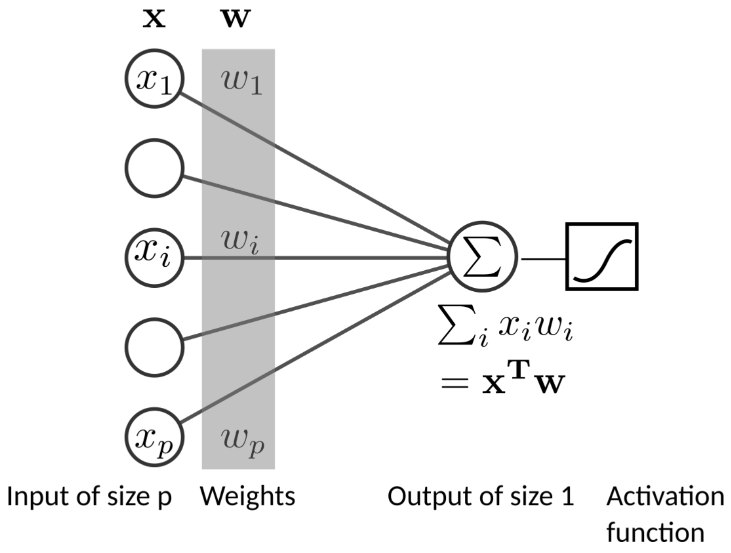


Figure 1.2: Linear models with an optional activation function.

$$
\begin{aligned}
\min_{\boldsymbol{w}} \; J &= \sum_{i=1}^{N} L_{\varepsilon}(y_i, f(\boldsymbol{x}_i, \boldsymbol{w})) + \Omega(\boldsymbol{w}) && (1.1)\\
&= L_{\varepsilon}(\boldsymbol{y}, f(\boldsymbol{X}, \boldsymbol{w})) + \Omega(\boldsymbol{w}) && (1.2)
\end{aligned}
$$

The sum of squared errors (SSE) loss

$$
\begin{aligned}
L(\boldsymbol{y}, f(\boldsymbol{X}, \boldsymbol{w})) &= \mathrm{SSE}(\boldsymbol{w})\\
&= (\mathbf{y} - \mathbf{X}\boldsymbol{w})^T(\boldsymbol{y} - X\boldsymbol{w})\\
&= \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2,
\end{aligned}
$$

Minimizing the SSE is the Ordinary Least Square (OLS) regression as objective function. This minimization has an analytic solution:

$$\boldsymbol{w}_{\mathrm{OLS}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

If needed, the gradient of the loss:

$$\partial \frac{L(\boldsymbol{y}, f(\boldsymbol{X}, \boldsymbol{w}))}{\partial \boldsymbol{w}} = 2\sum_i \boldsymbol{x}_i(\boldsymbol{x}_i \cdot \boldsymbol{w} - y_i)$$

The Logistic loss

The logistic regression for classification problems is a generalized linear models. It is a linear model with a link function that maps the output of linear multiple regression to the posterior probability of class 1 $p(1|x)$ using the logistic sigmoid function:

$$p(1|\boldsymbol{w}, \boldsymbol{x}_i) = \frac{1}{1 + \exp(-\boldsymbol{w} \cdot \boldsymbol{x}_i)}.$$

The *Loss function* for sample i is the negative log of the probability:

$$L(\boldsymbol{w}, \boldsymbol{x_i}, y_i) = \begin{cases} -\log(p(1|w, \boldsymbol{x_i})) & \text{if } y_i = 1 \\ -\log(1 - p(1|w, \boldsymbol{x_i}) & \text{if } y_i = 0. \end{cases}$$

For the whole dataset $\boldsymbol{X}, \boldsymbol{y} = \{\boldsymbol{x_i}, y_i\}$ the loss function to minimize $L(\boldsymbol{w}, \boldsymbol{X}, \boldsymbol{y})$ is the negative negative log likelihood (nll) that can be simplied using a 0/1 coding of the label in the case of binary classification:

$$\begin{aligned} L(\boldsymbol{y}, \boldsymbol{f}(\boldsymbol{X}, \boldsymbol{w})) &= -\log \mathcal{L}(\boldsymbol{w}, \boldsymbol{X}, \boldsymbol{y}) \\ &= -\log \Pi_i \{p(1|\boldsymbol{w}, \boldsymbol{x_i})^{y_i} * (1 - p(1|\boldsymbol{w}, \boldsymbol{x_i})^{(1-y_i)}\} \\ &= \sum_i \{y_i \log p(1|\boldsymbol{w}, \boldsymbol{x_i}) + (1 - y_i) \log(1 - p(1|\boldsymbol{w}, \boldsymbol{x_i}))\} \\ &= \sum_i \{y_i \boldsymbol{w} \cdot \boldsymbol{x_i}) - \log(1 + \exp(\boldsymbol{w} \cdot \boldsymbol{x_i}))\} \end{aligned}$$

This is solved by numerical method using the gradient of the loss:

$$\partial \frac{L(\boldsymbol{y}, f(\boldsymbol{X}, \boldsymbol{y}))}{\partial \boldsymbol{w}} = \sum_i \boldsymbol{x}_i (y_i - p(1|\boldsymbol{w}, \boldsymbol{x}_i)).$$

Logistic regression is a *discriminative model* since it focuses only on the posterior probability of each class $p(C_k|x)$. It only requires to estimate the P weights of the $\boldsymbol{w}$ vector. Thus, it should be favored over Linear Discriminant Analysis (LDA) with many input features. In small dimension and balanced situations it would provide similar predictions than LDA.

However, imbalanced group sizes cannot be explicitly controlled. It can be managed using a re-weighting of the input samples.

The hinge loss

This is the loss of Support Vector Machines which is given by:

$$L(\boldsymbol{y}, \boldsymbol{f}(\boldsymbol{X}, \boldsymbol{w})) = \sum_{i=1}^{N} (1 - y_i \boldsymbol{x}_i^T \boldsymbol{w})_+,$$

where $y_i \in \{-1, 1\}$ for binary classification problem.

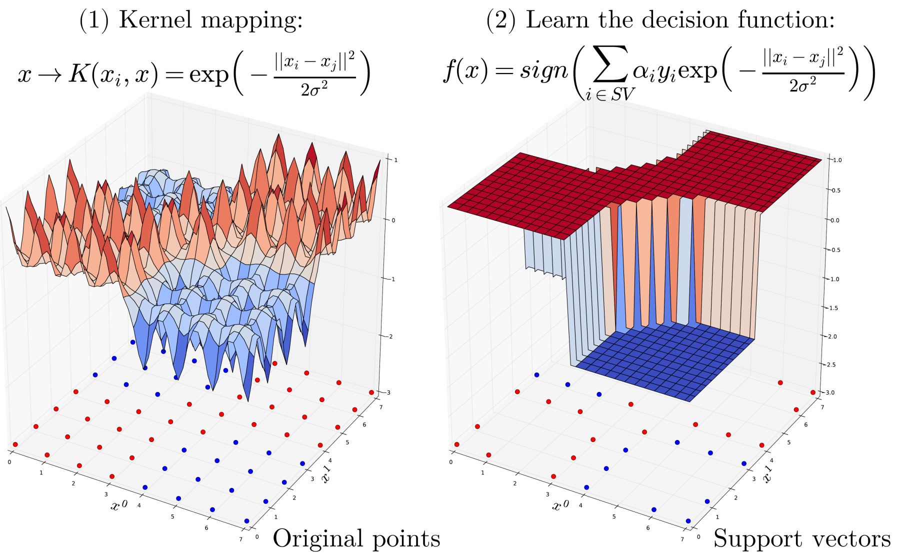
1.2.2 Non-linear kernel-based models: Support Vector Machines

Support Vector Machines (SVM)

SVMs belong to the family of kernel methods that rely on a kernel function $K(\boldsymbol{x}_i, \boldsymbol{x}_j)$ which measures the *similarity* over pairs of data points $(\boldsymbol{x}_i, \boldsymbol{x}_j)$. Input data are mapped into the kernel (dual) space on which learning algorithms operate linearly, i. e.every operation on points is a linear combination of $K(\boldsymbol{x}_i, \boldsymbol{x}_j)$. Outline of the SVM algorithm:

1. Map points $\boldsymbol{x}$ into kernel space using a kernel function: $\boldsymbol{x} \rightarrow K(\boldsymbol{x}, .)$.
2. Learning algorithms operate linearly by dot product into high-kernel space $K(., \boldsymbol{x}_i) \cdot K(., \boldsymbol{x}_j)$.
   - The kernel trick (Mercer's Theorem) replaces dot product in high dimensional space $K(., \boldsymbol{x}_i) \cdot K(., \boldsymbol{x}_j)$ by a simple similarity function $K(\boldsymbol{x}_i, \boldsymbol{x}_j)$. Thus, we only need to compute similarity for each pair of points and store them in a $N \times N$ Gram matrix.
   - Finally, The learning process consists of estimating the contribution of training samples (α_i in SVM, Eq. 1.3) in the decision function that maximizes the hinge loss plus some penalty when applied on all training points.
3. Predict a new point $\boldsymbol{x}$ using the decision function (here binary classification problem with class label $y_i \in \{-1, 1\}$):

$$f(\boldsymbol{x}) = \text{sign}\left(\sum_i \alpha_i y_i K(\boldsymbol{x}, \boldsymbol{x}_i)\right) \tag{1.3}$$



(1) Kernel mapping:

$x \rightarrow K(x_i, x) = \exp\left(-\frac{||x_i - x_j||^2}{2\sigma^2}\right)$

(2) Learn the decision function:

$f(x) = sign\left(\sum_{i \in SV} \alpha_i y_i \exp\left(-\frac{||x_i - x_j||^2}{2\sigma^2}\right)\right)$

Figure 1.3: Left: Kernel mapping using Gaussian kernel from red and blue dots (lower) to kernel space (upper). Right, Support Vector Machines decision function on the top and support vector on the bottom.

Gaussian kernel (RBF, Radial Basis Function)

RBF is the most commonly used kernel function. For a pair of points $\boldsymbol{x}_i, \boldsymbol{x}_j$ the RBF kernel is defined as:

$$\begin{aligned} K(\boldsymbol{x}_i, \boldsymbol{x}_j) &= \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{2\sigma^2}\right) \\ &= \exp\left(-\gamma \, \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2\right), \end{aligned}$$

where σ (or γ) defines the kernel width parameter. Basically, we consider a Gaussian function centered on each training sample $\boldsymbol{x}_i$. It has a ready interpretation as a similarity measure as it decreases with squared Euclidean distance between the two points. Non-linear SVM also exists for regression problems.

2 Fighting overfitting

This chapter presents two approaches to overcome the overfitting phenomenon (summarized Sec. 2.1): Sec. 2.2 regularization and Sec. 2.3 dimension reduction using feature selection. Using these approaches, we designed predictive pipelines combining feature selection and regularized models. Three contributions are presented in the following sections:

- Sec. 2.4 explores various strategies to predict sex from the cortical folding patterns (Duchesnay et al., 2007). This work laid the foundation for the following researches; it demonstrated the relevance of dimension reduction associated with regularization; however it raised the crucial issue of model selection.
- Sec. 2.5 presents an inverse inference model to predict the visual stimuli from functional activation map (Thirion et al., 2006).
- Sec. 2.6 proposes a strategy to predict ASD based on whole-brain PET imaging using regional features (Duchesnay et al., 2011). This work presents a preliminary proposition to extract interpretable predictive maps from whole-brain images.

2.1 The overfitting phenomenon

The estimation of the model parameters (coefficient vector $\boldsymbol{w}$ in linear models) is very sensitive to the conditioning of $\boldsymbol{X}$, and sometimes produces a dangerous situation of overfitting. In statistics and machine learning, overfitting occurs when a statistical model describes random errors or noise instead of the underlying relationships. In such situations, the model performs perfectly on the training data but will lead to poor performances and replicability on independents subjects. The overfitting phenomenon has three main explanations: excessively complex models, multicollinearity, and high dimensionality.

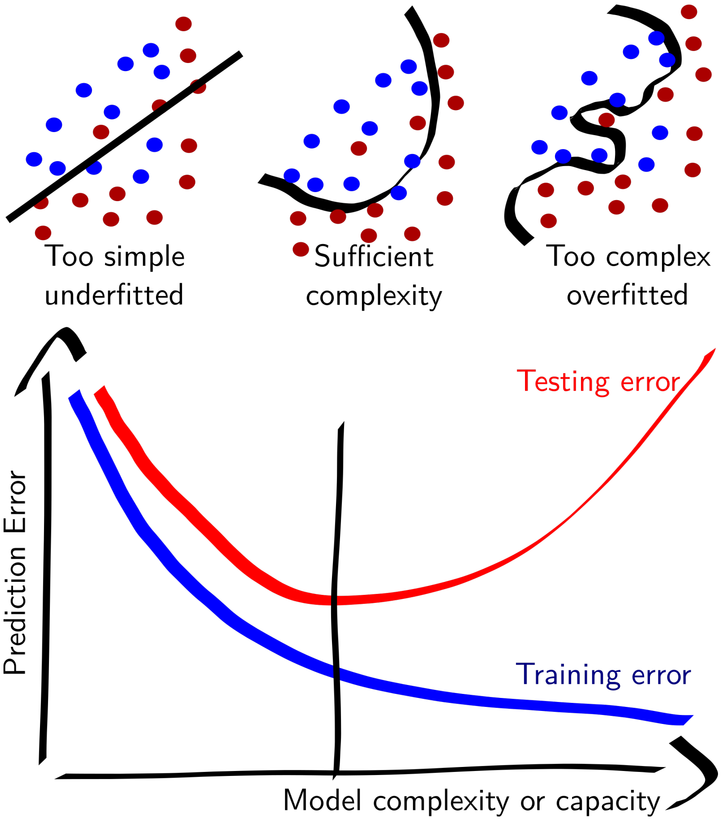


Figure 2.1: Top: three models of increasing complexity (capacity). Bottom: The increase of model complexity always reduces prediction error on the training dataset. However, beyond the necessary complexity, the error on an independent test set increases. The unnecessary complexity gives the model the capacity to fit the noise of the training data that will not replicate on test data.

Statistical learning theory (Vapnik, 1999) founded the theoretical background of overfitting. It provides measures to quantify the model capacity. Among these, the most well-known is the Vapnik-Chervonenkis dimension or VC dimension defined as the cardinality of the largest set of points that the algorithm can label arbitrarily.

The VC dimension states that a linear classifier with P input features has VC dimension of $P + 1$. Hence, as shown in Fig. 2.2, in dimension two ($P = 2$) any random partition of 3 points can be learned. Thus, any linear classifier can trivially obtain 100% of classification on N training sample with only $P = N - 1$ features.

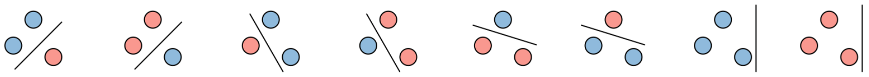


Figure 2.2: In two dimensions, we can shatter any three non-collinear points. VC dimension of a linear classifier in 2D equals 3.

The risk of overfitting is specifically high in the context of neuroimaging data, where the number of features (e.g., number of voxels/vertices) for a subject is much larger than the total number of subjects, resulting in high-dimensional data. This problematic situation leads to a high imbalance between the number of parameters to estimate and the number of available samples. It sometimes results in extremely complex models with low generalization capabilities. Moreover, neuroimaging measurements are frequently correlated. In this situation, the coefficient estimation in the multiple regression may fluctuate erratically in response to small changes in the model or the data. Multicollinearity does not reduce the predictive power or reliability of the model as a whole, at least not within the sample data set; it only affects computations regarding individual predictors. That is, a multiple regression model with correlated predictors can indicate how well the entire bundle of predictors predicts the outcome variable, but it may not give valid results about any individual predictor, or about which predictors are redundant for each other.

2.2 Regularization based on shrinkage of the coefficient vector

Regarding linear models, overfitting generally leads to excessively complex solutions (coefficient vectors), accounting for noise or spurious correlations within predictors. **Regularization** aims to alleviate this phenomenon by constraining (biasing or reducing) the capacity of the learning algorithm to promote simple solutions. Regularization penalizes "large" solutions forcing the coefficients to be small, i.e., to shrink them toward zeros.

The *objective function* $J(\boldsymbol{w})$ to minimize with respect to $\boldsymbol{w}$ is composed of a *loss function* $L_\varepsilon(\boldsymbol{w})$ for goodness-of-fit and a *penalty term* $\Omega(\boldsymbol{w})$ (regularization to avoid overfitting). This is a trade-off where the respective contribution of the loss and the penalty terms is controlled by the regularization parameter λ. Therefore the loss function $L_\varepsilon(\boldsymbol{w})$ is combined with a penalty function $\Omega(\boldsymbol{w})$ of the Eq. 1.1.

2.2.1 ℓ_2 (Ridge) regularization

ℓ_2 Regularization imposes a ℓ_2 penalty on the coefficients, i.e., it penalizes with the Euclidean norm of the coefficients while minimizing the loss function. The objective function of Eq. 1.1 becomes:

$$\min_{\boldsymbol{w}} \sum_{i=1}^{N} L_\varepsilon(y_i, f(\boldsymbol{x}_i, \boldsymbol{w})) + \lambda_2 \|\boldsymbol{w}\|_2^2 \tag{2.1}$$

with $\lambda_2 \geq 0$ and $\|\boldsymbol{w}\|_2 = \sqrt{\sum_{i=1}^{P} w_i{}^2}$

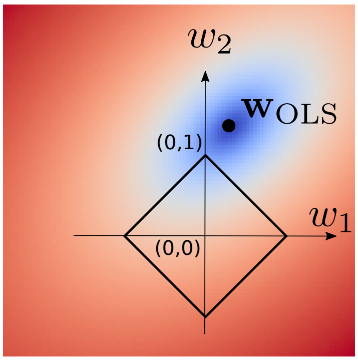
2.2.2 ℓ_1 (Lasso) regularization

However, the Ridge penalty does not assign exactly zero coefficients to predictors. Yet, with high dimensional features, such as with neuroimaging datasets, many variables are expected to be irrelevant for the prediction task. They should be removed from the model. One solution to conduct such variable selection is the use of ℓ_1 (a.k.a Lasso) penalty. The lasso (Least Absolute Shrinkage and Selection Operator) constraint (Tibshirani, 1996) penalizes the ℓ_1-norm of the coefficients vector. It enforces many coefficients to have zeros weights. The criterion to optimize becomes:

$$\min_{\boldsymbol{w}} \sum_{i=1}^{N} L_{\varepsilon}(y_i, f(\boldsymbol{x}_i, \boldsymbol{w})) + \|\boldsymbol{w}\|_1, \tag{2.2}$$

with $\lambda_1 \geq 0$ and $\|\boldsymbol{w}\|_1 = \sum_{i=1}^{p} |w_i|$

In contrast to the ridge regression, the lasso regression can perform variable selection. Indeed, due to the geometric properties of ℓ_1 norm, Lasso tends to reach its minimum along the axis. Indeed, as shown in the figure, minimizing a quadratic loss (in blue) under the constraint of the ℓ_1 norm of $\|\boldsymbol{w}\|_1 \leq 1$. ℓ_1 norm promotes optimal solution along axis. Here $w_1 = 1$. Thus, it produces sparse solution of $\boldsymbol{w}$ by selecting at most k non-null coefficients for $k \ll p$. This sparse configuration of the solution is desirable for the interpretability of prediction. However, in a set of correlated predictors, the lasso regression tends to select only one variable on the set. Such selection might be unstable, and thus interpretability is still limited The Lasso regression problem lacks an analytic solution. It is convex but not differential anymore due to the addition of the ℓ_1 penalty. It requires specific optimization algorithms such as FISTA: the fast iterative shrinkage-thresholding algorithm described in (Beck and Teboulle, 2009a).



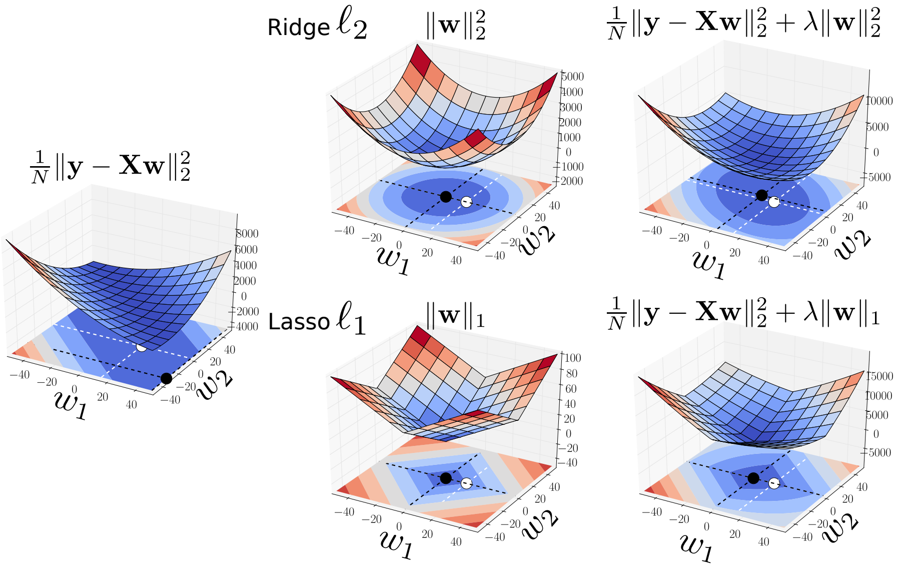


Figure 2.3: ℓ_1, ℓ_2 shrinkages.

2.2.3 ℓ_1, ℓ_2 (ElasticNet) regularization

The ElasticNet model combines both ℓ_1 and ℓ_2 penalties (Zou and Hastie, 2005):

$$\min_{w} \sum_{i=1}^{N} L_{\varepsilon}(y_i, f(x_i, w)) + \lambda_1 \|w\|_1 + \lambda_2 \|w\|_2^2. \tag{2.3}$$

ElasticNet associates the advantages of both Ridge (ℓ_2) and Lasso penalties by favoring sparse and stable configurations in case of correlated predictors. Elastic net encourages a grouping effect, where strongly correlated predictors tend to be in or out of the model together. As with Lasso, ElasticNet can be solved with FISTA algorithm.

2.3 Feature selection

The selection of a reduced set of input features may improve the predictive performance by reducing the risk of overfitting and producing simpler and more interpretable models, see (Guyon and Elisseeff, 2003; Guyon et al., 2006) for a review and NIPS 2003 Feature Selection Challenge (Guyon et al., 2005). Some approaches (filters wrappers) include two distinct steps: feature ranking then model selection to feed the predictive algorithms. Conversely, embedded methods combine feature selection in the learning process.

2.3.1 Univariate feature ranking: filters

Filters are a simple, widely used method for *supervised* dimension reduction (Guyon and Elisseeff, 2003). Filters are univariate methods that rank features according to their ability to separate populations, independently of other features. This ranking may be based on parametric (e.g. t-test, ANOVA) or non-parametric (e.g. Wilcoxon test) statistical methods. Filters are computationally efficient and more robust to overfitting than multivariate methods. However, they are blind to feature interrelations, a problem that can be addressed only with multivariate selection.

2.3.2 Embedded methods

We need to consider combinations of features for at least two reasons: (i) This makes it possible to remove redundant features, which is important as redundancy may weaken the classification. For example, if a neurodegenerative disease modifies the depth of a sulcus but not its length, then depth and area ($\approx$depth$\times$length) will be discriminant (i.e. with a high univariate rank), but area measurements simply increase classifier noise. A feature selection combining depth and area could be used to detect this kind of situation of redundancy. (ii) It also makes it possible to exploit informative interrelations.

Embedded methods are multivariate feature selection methods that combine feature selection in the learning process. Embedded methods can produce sets of selected feature of different size by playing on some regularization parameters. I. e.decreasing the ℓ_1 regularization parameter produces selected feature of increasing size. Therefore, model selection is also required with most of embedded algorithms.

Examples of Embedded methods:

- Automatic Relevance Determination (ARD) (MacKay and Laboratory, 1994) provides a Bayesian framework wheres priors allow the model to determine which of the features are most relevant.

- Recursive Feature Elimination (RFE) (Guyon et al., 2002) performs feature selection by iteratively training the model on a reduced set of features selected according to their importance in the current model. In linear models, the absolute value of the weight can be use as the feature importance.

- Random Forests (Breiman, 2001) is an ensembles learning of decision trees select features in the process of building classification or regression tree.

- ℓ_1 regularization may be used as embedded feature selection. Indeed, the use of the ℓ_1 norm constraint on the parameter leads to a sparse model.

2.3.3 Wrappers: forward-and backward-stepwise selection

Wrappers use the *objective function* of a learning algorithm as a "black box" to score subsets of features according to their predictive power (Guyon and Elisseeff, 2003). Wrappers explore the features space with greedy strategies and can work in two ways: forward or backward. Sequential *forward* selection (SFS) adds features that improve the objective function, whereas sequential *backward* selection (SBS) deletes features that weaken the objective function. SFS and SBS are sensitive to the nesting effect, as they never backtrack on their choices. A hybrid strategy, sequential floating forward selection (Pudil, Novovi*v*cová, and Kittler, 1994), limits this effect by nesting a backward loop, that deletes the worst feature only when the objective function is improved, within a forward loop.

The wrapper evaluates the objective function to select (forward) or remove (backward) a feature. The choice of the objective function is closely related to the classifier. For LDA, a Manova F-statistic (Pillai-Bartlett trace) is recommended (Hand and Taylor, 1987). SVMs provide many methods for estimating the generalization power on unseen data (the bounds of test error). Such methods are useful in implementing the objective function. We used the margin-based bound based on its simplicity and demonstrated efficiency in a previous comparative study (Rakotomamonjy, 2003). Finally, it is possible to use cross-validation (CV) of the loss function on training data (Schölkopf and Smola, 2002) (chap. 12.2).

Results from "NIPS 2003 Feature Selection Challenge" (Guyon et al., 2005) recommend the use of embedded methods over wrappers.

2.3.4 Model (dimension) selection

Most feature selection methods yield several nested feature sets of increasing dimensions. The user must select the size k of the subset to be used for further processing: either training of the classifier or an additional step of multivariate feature selection. Most dimension selection approaches are of one of three types:

Cross-validation

Choose k from cross-validation: In this classical solution, the classifier is trained on subsets of features of increasing size. The dimension yielding the best generalization rate is then chosen.

Select fewer features than samples

If k is to be selected on the basis of information about the classifier, then as the first rule of thumb is to take k smaller than that k the number of subjects (N). This choice can be justified by many theoretical arguments:

- A Linear Discriminant Analysis (LDA) classifier, estimates $\mathcal{O}\left(k^2\right)$ parameters ($k(k-1)/2+k$ for the covariance matrix and $2k$ for the 2 means). Therefore, we want $k \leq N$ such that the number of estimated parameters is smaller than the dataset size ($N \times k$).
- VC dimension Sec. 2.1 states that linear classifier in dimension k can reach 100% of classification rate on $N-1$ samples. Considering the Occam's razor principle that states that among competing hypotheses that explain known observations equally well, one should choose the "simplest," we should select models with $k \leq N-1$ input features.

Thresholding based on univariate p-values

In cases of univariate feature selection (filter), k can be obtained with a correction of multiple comparisons (Bonferroni correction or false discovery rate (FDR)).

In-sample prediction error: structural risk minimization

An alternative to cross-validation is to penalize the training error with the model complexity to calculate error bound as a function of the model complexity (the number of features) and use this bound for model selection. Structural risk minimization provides such error bounds.

The complexity of the class of functions performing classification or regression and the algorithm's generalizability can be evaluated by the Vapnik-Chervonenkis (VC, see Sec. 2.1) theory that provides a general measure of complexity and error bounds as a function of complexity. Structural risk minimization search for a model that minimizes this bound, which depends on the empirical risk (error on training sample) and the capacity of the function class (Vapnik, 1999).

For SVM-based classifiers the simplest bound (Schölkopf and Smola, 2002) is given by the proportion of support vectors (SVs):

$$\mathbb{E}[P(\text{error})] \leq \frac{\text{number of SVs}}{\text{Number of training samples}}$$

The selected model is the one that minimizes the proportion of SVs.

In-sample prediction error: penalized likelihood

Another way to get an in-sample prediction error is to estimate the optimism and then add it to the training error using a penalized likelihood framework. Many criteria have been proposed ((Burnham and Anderson, 1998)), and the two most commonly used are the Bayesian (BIC, Schwarz, 1978) and the Akaike information criteria (AIC, Akaike, 1974). Despite their different theoretical foundations, both yield a similar linear penalization of the log-likelihood with the number of parameters. The BIC is given by

$$BIC = -2\mathcal{L}(\boldsymbol{y}, f(\boldsymbol{X}, \boldsymbol{w})) + k\ln(N),$$

where $\mathcal{L}(\boldsymbol{y}, f(\boldsymbol{X}, \boldsymbol{w}))$ is the likelihood of the fitted model $f(., \boldsymbol{w})$ on data $\boldsymbol{X}, \boldsymbol{y}$.

2.4 Feature selection and classification based on cortical folding patterns

In Duchesnay et al., 2007, we performed an extensive comparison of strategies combining feature selection with multivariate classifiers to deal with the overfitting problem in high dimensional settings.

The aim was to predict the sex of a subject from its cortical folding. 151 structural MRIs from the ICBM database (65 women and 86 men) were processed with brainVISA to identify 116 sulci for a total of 6747 (shapes and coordinates) descriptors. Preliminary spatial affine normalization prevented the classifier from using global differences in volume between the sexes.

This compares main strategies of predictive pipeline with feature selection:

1. *Filter+SVM-Lin*: Univariate filter (based on t-test) combined with Linear SVM (ℓ_2 regularization + hinge loss).

2. *Filter+SVM-RBF*: Univariate filter (based on t-test) combined with non linear SVM-RBF.

3. *RFE+SVM-Lin*: Recursive Feature Elimination combined with Linear SVM.

4. *Elasticnet+LinSVM*: Feature selection obtained with ℓ_1-based regularization of the linear SVM with ℓ_2 penalty.

2.4.1 General behavior of feature ranking

Whatever the feature ranking strategy, we observed a similar pattern of behavior:

- Feature ranking must be performed on training data only, i.e., withing the cross-validation loop. As shown in Fig. 2.4, feature selection on the whole dataset leads to severely optimistically biased performances that are close to the performance obtained on the training dataset.

- Performances with feature ranking can be divided into four distinct phases (Fig. 2.4): after having selected few features, (i) we reach a phase (green shaded region) where best-ranked features improve the performance compared to baseline ℓ_2 regularized model (using all features). (ii) Then (yellow region), performances stop to increase and become unstable. In a third phase, (iii, red region) feature selection induces overfitting until it reaches a final phase (iv, blue region) where it has no more effect of prediction since almost all the features are selected.

- Overfitting starts to occur around $P \approx N = 151$ input features, which generally marks the end of the second phase (green range). Here, prediction on train data reaches almost 100% of accuracy, and test performance becomes irregular (yellow range).

- Feature selection can improve the performance: in Fig. 2.4 at the end of the first phase ($10 \leq k \leq 20$ green region), this improvement is significant: the baseline scores are bellow the lower confidence interval.

- Model selection (of the k number of feature) is a difficult task. Selection using nested cross-validation might not help due to large errors (see confidence intervals of the blue curve in Fig. 2.4).

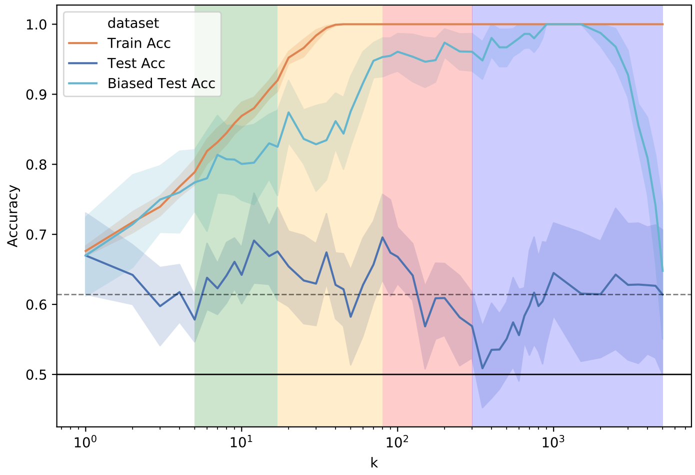
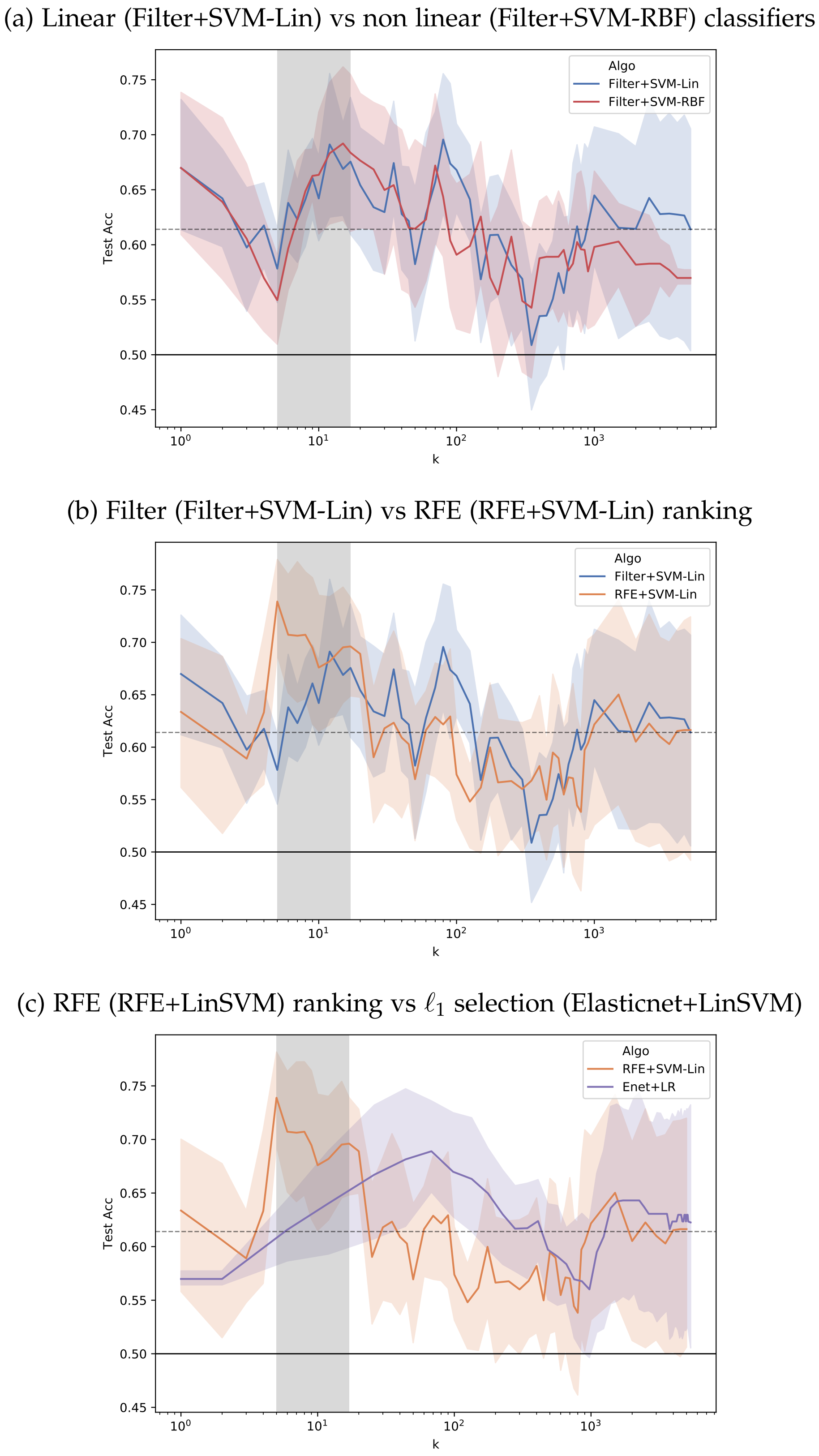


Figure 2.4: Accuracies on training and testing (measured with a 10 folds CV) data as a function of an increased number of (k) selected features, using a univariate filter. The "Test Biased" curve shows biased performances obtained with feature selection performed outside of the CV loop, i.e., on all data. The classifier is a linear SVM with $C = 1$. The shades surrounding the lines depict the 95% confidence intervals (CIs). The black horizontal lines show the chance level (plain line) and the predictive performances without the feature selection (dotted line). The colored shaded ranges of k depict the four phases of feature ranking: (i) green: improvement, (ii) yellow: instability (iii) red: overfitting, and (iv) blue: neutral effect.

2.4.2 Results

Comparing feature ranking strategies

- Fig. 2.5 (a) shows that non-linear SVM-RBF model does not outperform linear models. This suggests that the performances are driven by the selected input features.
- Fig. 2.5 (b) shows that sophisticated multivariate feature ranking (RFE) and simple filter produce a comparable performance profile in four phases. Nevertheless, the best result was obtained with RFE (around 7 features), demonstrating the potential to identify a predictive pattern. However, this peak performance is useless. Indeed, the model (dimension) selection has not been made yet, and it is unlikely that a model selection procedure could pick this precise location.
- Fig. 2.5 (c) shows that (i) ℓ_1-based feature selection does not outperform RFE. It has a simpler profile of performance based only on 3 phases: (i) an increase, (ii) an overfiting and (iii) a neutral effect.



(a) Linear (Filter+SVM-Lin) vs non linear (Filter+SVM-RBF) classifiers

(b) Filter (Filter+SVM-Lin) vs RFE (RFE+SVM-Lin) ranking

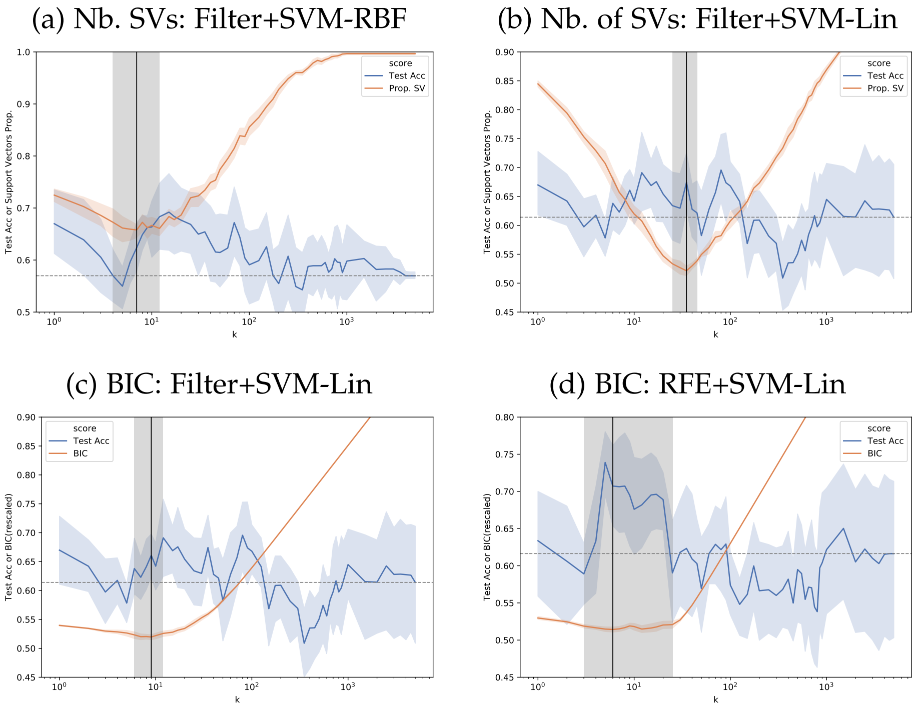
(c) RFE (RFE+LinSVM) ranking vs ℓ_1 selection (Elasticnet+LinSVM)

Figure 2.5: Test accuracies as a function of k the number of selected features for three strategies of feature selection combined with a multivariate classifier: Univariate filter with linear SVM, Univariate filter with non-linear SVM-RBF and RFE with linear SVM. The black horizontal lines show the chance level (plain line) and the baseline predictive performances without the feature selection (dotted line). The gray shaded region shows the range where all features selection strategies outperform the baseline model.

The problem of model selection

Efficient model selection is the key point to benefit of feature selection . Fig. 2.6 presents results obtained with methods presented in Sec. 2.3.4.

- As a rule of thumb, we should use fewer features than subjects ($k \leq N$). This recommendation is inspired from VC dimension and experiments corroborated this statement: when $k \geq N$ all strategy started to overfit the training data (Fig. 2.5).
- CV should be used carefully due to large error bars, see Fig. 2.5 and Fig. 2.6 for a simplified figure. This recommendation is no more valid for highly correlated features, like brain images.
- Fig. 2.6 shows that in in-sample model selection criteria (such as the proportion of support vectors or BIC) provide acceptable solutions with low variability and a negligible computational burden.


(a) Nb. SVs: Filter+SVM-RBF

(b) Nb. of SVs: Filter+SVM-Lin

(c) BIC: Filter+SVM-Lin

(d) BIC: RFE+SVM-Lin

Figure 2.6: Test accuracy (blue line) and model selection criteria (orange line) as a function of k. The vertical line is the solution selected by minimizing criteria (proportion of Support Vectors or BIC), the shaded gray region shows its confidence interval.

2.4.3 Discussion and conclusion

- All feature selection strategies significantly outperformed baseline ℓ_2-regularized models (without feature selection) on a small range of input features.

- Simple filter is a good default candidate to build a predictor in high dimensional space. Experiments demonstrate that filters are as efficient as sophisticated multivariate feature selection. In Guyon et al., 2006 (p.242), I. Guyon summarizes the results obtained from the NIPS2003 challenge, and she states: "Filter methods are powerful."

- Concerning sophisticated multivariate feature selection, ℓ_1-regularization should be favored. Indeed, it minimizes a well-defined convex problem leading to smoother performance profile (Fig. 2.6 (c)) where CV may help to select a reasonable model.

- To be useful, feature selection requires an efficient model selection procedure to select a model somewhere at the end of phase 1. We suspect that this major difficulty led to disappointing results reported by Chu et al., 2012, and Kerr et al., 2014.

- CV-based model selection is not efficient on small dataset du large confidence intervals Varoquaux, 2018.

- In-sample criteria provide an efficient solution. However, they may require some calibration (of the predicted probability to compute the likelihood in the BIC, or the model complexity term).

The predictive features include (Fig. 2.7):

- *Central and pre-central regions*: In men, the right central sulcus of men folds deeper into the internal face (closer to the interhemispheric plane). Conversely, The right inferior pre-central sulcus seems deeper and closer to the interhemispheric plane in women.

- *Temporal lobe*: The anterior part of the right occipito-temporal sulcus follows different orientations in the two sexes.

- *Occipital lobe*: The right lingual sulcus is deeper and larger in men.

- *Parietal lobe*: The post-central and intra-parietal sulci are shifted upward in men.

- *Cingulate*: The left posterior cingulate sulcus anterior extremity is higher in women.

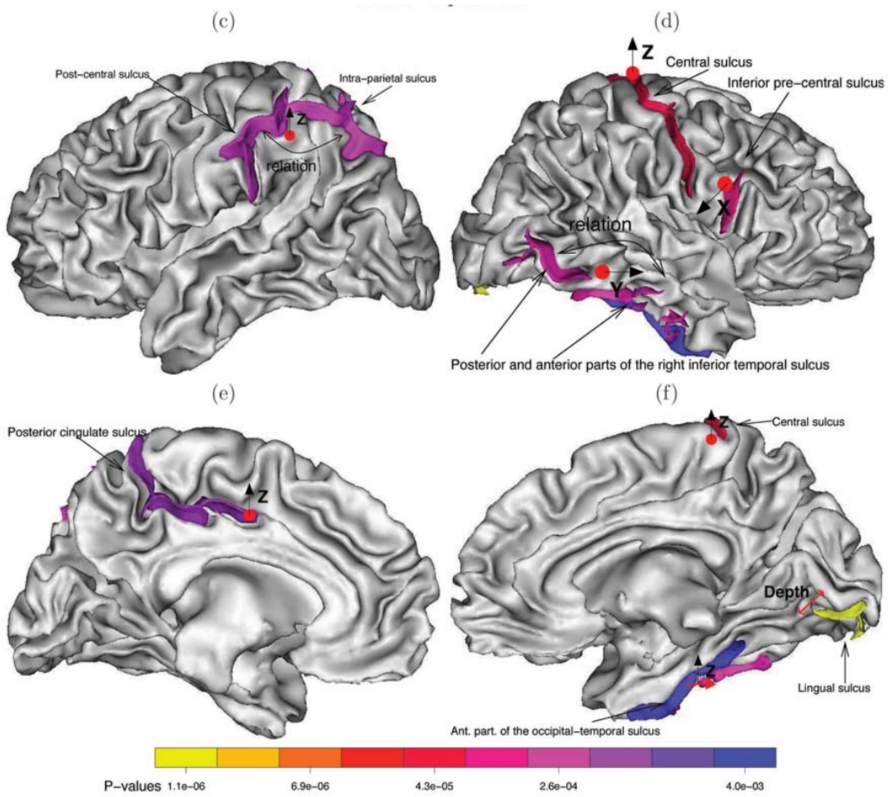


Figure 2.7: From (Duchesnay et al., 2007): Sulci features that are significantly associated with sex. Taken together, they provide a significant prediction of the sex.

2.5 Predicting stimuli from visual cortex activation using voxels selection and SVM

2.5.1 Retinotopic functional activation maps of "dominos" visual stimuli

Traditional inference in neuroimaging consists of describing brain activations elicited and modulated by different kinds of experimental conditions such as visual stimuli. In (Thirion et al., 2006), we addressed the inverse problem of predicting the visual stimuli given functional MRI activation images. Such brain decoding exploited the well-known retinotopic mapping of the visual cortex (Sereno et al., 1995) to infer the visual content. We used univariate feature selection (filter) combined with linear SVMs to perform the prediction.

In the *domino* experiment, two grids, situated on the left and right parts of the visual field, and a central fixation cross, were presented to the subjects. Every 8s, a flickering pattern appeared in several sectors of the grid. These patterns belonged to a set of 6 possible dominos shapes. The dominos were presented simultaneously to the bilateral visual fields for a total of 36 combinations. Each subject performed the experiment four times. For each subject and each hemisphere, 144 (24 maps for each of the 6 domino shape) trial-specific activation maps were extracted using classical first-level analysis based on General Linear Model (GLM). Maps were masked by the retinotopic regions estimated from a retinotopic mapping experiment yielded to $\approx 10,000 - 15000$ voxels, according to the subject.

2.5.2 Decoding: prediction of visual stimulus from activation map

We combined univariate feature selection (ANOVA-based) filter with a linear SVM. As model (dimension) selection, we retained only voxels that were significantly associated with the domino category using a threshold at p-value ≤ 0.1 after a False Discovery Rate (FDR) correction for multiple comparison (Benjamini and Hochberg, 1995).

All 16 datasets (8 subjects both hemispheres) were classified (accuracy between 60 and 96%, 82% in average) significantly (p-value $\leq 10^-3$) above the chance level (1/6 or 16.7% correct responses). Across subjects and hemisphere, we found that 60-70% of the most discriminative voxels were in V1, while only 5-20% were in V2 (ventral and dorsal). We did not try to study other visual areas since their delineation was not reliable enough from our retinotopic maps.

2.6 Feature selection and classification based on PET images of children with autistic spectrum disorders

Autism Spectrum Disorders (ASD) are typically characterized by impaired social interaction, narrow interests, and repetitive behaviors, with high variability in expression and severity. The numerous findings revealed by brain imaging studies suggest that ASD is associated with an intricate and distributed pattern of abnormalities that makes the identification of a shared and common neuroimaging profile a problematic task.

In Duchesnay et al., 2011, we aim to identify the rest functional (PET scans perfusion) brain imaging abnormalities pattern associated with ASD and to validate its efficiency in the individual classification. The dataset was small and highly imbalanced: 45 low-functioning children with ASD and only 13 non-ASD low-functioning children. The small number of control is inherent to the difficulties to obtain PET scan of low-functioning children. Additionally to those difficulties, clinicians requested for an interpretable predictive brain map organized in brain regions. Indeed activation abnormalities map is neither expected to be scattered into isolated voxels (a solution that would be produced by a sparsity promoting penalty) nor allowed to show rapid changes of positive (hyperperfusion) and negative (hypoperfusion) values (produced by a ℓ_2 penalty).

2.6.1 Regional feature extraction

To take advantage of such biological priors to limit overfitting, we designed a simple pipeline the first stage of which extracts regional feature: Given to simplicity and the good performances of univariate feature filtering obtained in (Duchesnay et al., 2007) and especially on whole brain images (Thirion et al., 2006), we ranked voxels based on t-tests (Fig. 2.9 step 1.1). Such parametric statistics rely on some assumptions (independence, normality of the residuals, and homoscedasticity). However, they do not favor the most numerous class, which was a risk in such an imbalance setting. We selected voxels with a p-value < 0.001, uncorrected for multiple comparisons. Thresholded-connected voxels were grouped into regions (or clusters), and the PET signal was averaged within each region, producing a set of new regional features (Fig. 2.9 step 1.3). The p-value threshold was empirically chosen among the three possible values of 10^{-2}, 10^{-3} and 10^{-4} as the value that produces clusters of reasonable size.

2.6.2 Regional feature ranking

The resulting regional features were ranked using univariate (filter) or multivariate (wrappers, RFE, and ℓ_1-based regularization) approaches that all produce similar results.

2.6.3 Model selection: calibration of penalized likelihood using randomization

The third stage (step 3 in Fig. 2.9) selects the optimal subset of features to be used by the final classifier . This model selection problem is solved with a penalized likelihood framework such as BIC or AIC (Sec. 2.3.4). Such an approach would not be appropriate for feature selection in high-dimension Giraud, 2014. Nevertheless, it is conceivable in the low dimensional space of regional features. Nevertheless, our two-stages feature selection procedure prevents a straightforward application of a fixed penalty as a function of the number of regional features. Indeed, such a penalty would ignore the overfitting induced by the first step of region extraction. We demonstrated this under-penalization of fixed penalties criteria in Fig. 2.8. Those limitations motivated the development of data-driven methods to calibrate criteria whose penalties are known up to a multiplicative factor, e.g. “the slope heuristics” proposed by (Birgé and Massart, 2007).

Adaptive log-likelihood penalization

Likewise, “the slope heuristics” of (Birgé and Massart, 2007), we retain the linear penalization as function the number regional features (BIC like) but we loosen the fixed link by adding a free parameter (noted “a” in Eq.2.4):

$$\text{aPena} = -\mathcal{L} + \boldsymbol{a}\,\frac{1}{2}k\ln(N). \tag{2.4}$$

Penalization calibration based on randomized data

The penalization value (a) is calibrated to the overfitting caused by the feature selections, using an estimation under the null hypothesis obtained from randomized datasets, i.e., randomly permuted $\boldsymbol{y}$. We repeated the whole algorithm and measured the increase of the log-likelihood ($\ln\mathcal{L}$), purely due to the overfitting of the training, data and compared it to the theoretical log-evidence, which is supposed to be constant and equal to $\ln(1/2)^N$. A good penalization is supposed to fit this increase. Fig 2.8 clearly shows that the feature selection algorithm dramatically increases the overfitting, which is not balanced with the pure BIC or AIC penalization criteria. However, this experiment also suggests that a satisfying linear approximation can be estimated. This calibration is conducted within each cross-validation fold, excluding the test sample. The estimated adaptive penalization values (a) (average across folds $2.62, \pm 0.03$) were then plugged in Eq. 2.4). The features subset that maximized this adaptive penalized log-likelihood was selected for the classification step.

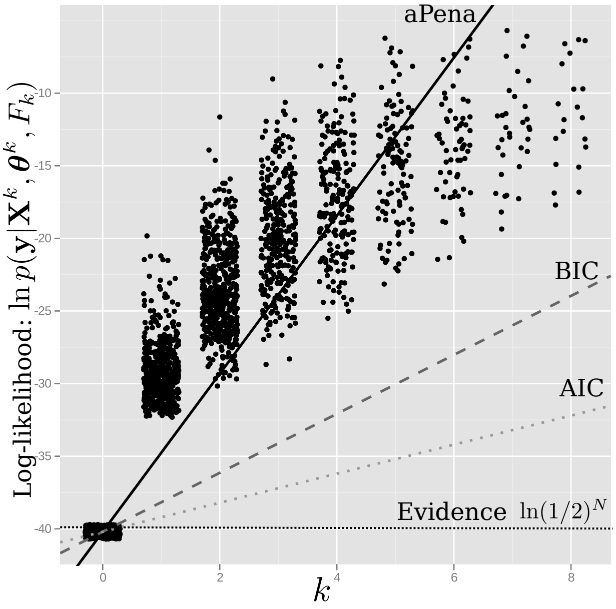


Figure 2.8: Calibration of the penalization with randomization: The multi-stages feature selection algorithm has been repeated on randomly permuted datasets. We can observe the increase of the log-likelihood for a varying number of regional features (k). We reported the theoretical log-evidence (baseline), which is supposed to be constant and equal to $\ln(1/2)^N$. We also reported the penalizations obtained with the BIC and AIC criteria. This experiment shows that those fixed penalty criteria lead to a severe under-penalization of the log-likelihood. However, it also demonstrates that a good linear approximation can be obtained, leading to an adaptive penalization criterion noted (aPen) with a penalty term of $2.67\ \frac{1}{2}\ k \ln N$ as noted in (2.4).

2.6.4 Performances validation and comparison methodology

The classification accuracy of the entire pipeline (including the feature selection) was evaluated by leave-one-out cross-validation (LOO-CV).

By combining a selected choice of those alternatives, we formed four strategies:

1. No feature selection combined with a ℓ_2 linear (reweighted) SVM classifier. This strategy acts as a baseline to highlight the specific contributions of feature selection.

2. Univariate t-test feature subset ranking, CV-based model selection with a linear (reweighted) SVM classifier. This strategy acts as a baseline to highlight the specific contributions of the two last strategies based on multivariate feature selection.

3. RFE based feature ranking, CV-based model selection with a linear (reweighted) SVM. This strategy is commonly used as

a "state-of-the-art" multivariate feature selection with a kernel-based classifier (Ecker et al., 2010; Guyon et al., 2006).

4. Lasso-based feature selection (CV-based model selection) with (reweighted) logistic regression classifier. This strategy is a "state-of-the-art" representative of recent advances in ℓ_1-regularized based methods.

Those four strategies were first directly applied on entire brain images (hundreds of thousands of voxels), and then on the regional features (≈10). The low dimension space made of regional features (≤ 10) allowed us to test a Linear Discriminant Analysis (LDA) classifier, replacing discriminative models (SVM, logistic regression) by a generative model. Indeed, generative models estimate the parameters of the classes' conditional distributions independently limiting the bias toward the most numerous class. Class imbalance can be nicely controlled with the classes' prior probabilities where discriminative models have to use a sample reweighting "trick" to re-balance the load of class in the minimization problem.

2.6.5 Results

Prediction performances

- ℓ_2 penalty on whole-brain images obtained a significant AUC of 0.64. However, predictions were strongly biased toward the most numerous class (ASD) (93% of sensitivity and 23% of specificity).

- Generative model (LDA), using a reduced set of regional features, obtained comparable AUC of 0.65 with the benefice of balanced predictions (82% of sensitivity and 53% of specificity)

- Model selection based on adaptive (calibrated) penalized likelihood improves performances of all the feature selection methods: AUCs of 0.81 with ℓ_1 and 0.74 with RFE-LDA, with better balance in predictions: sen./spe. of 75%/69% with ℓ_1 and 88%/69% with RFE-LDA.

Predictive regions of ASD

The pipeline identified a final characteristic pattern in the ASD group that featured a hypoperfused region in the right superior temporal sulcus and a hyperperfused region in the left post-central (Fig. 2.9.

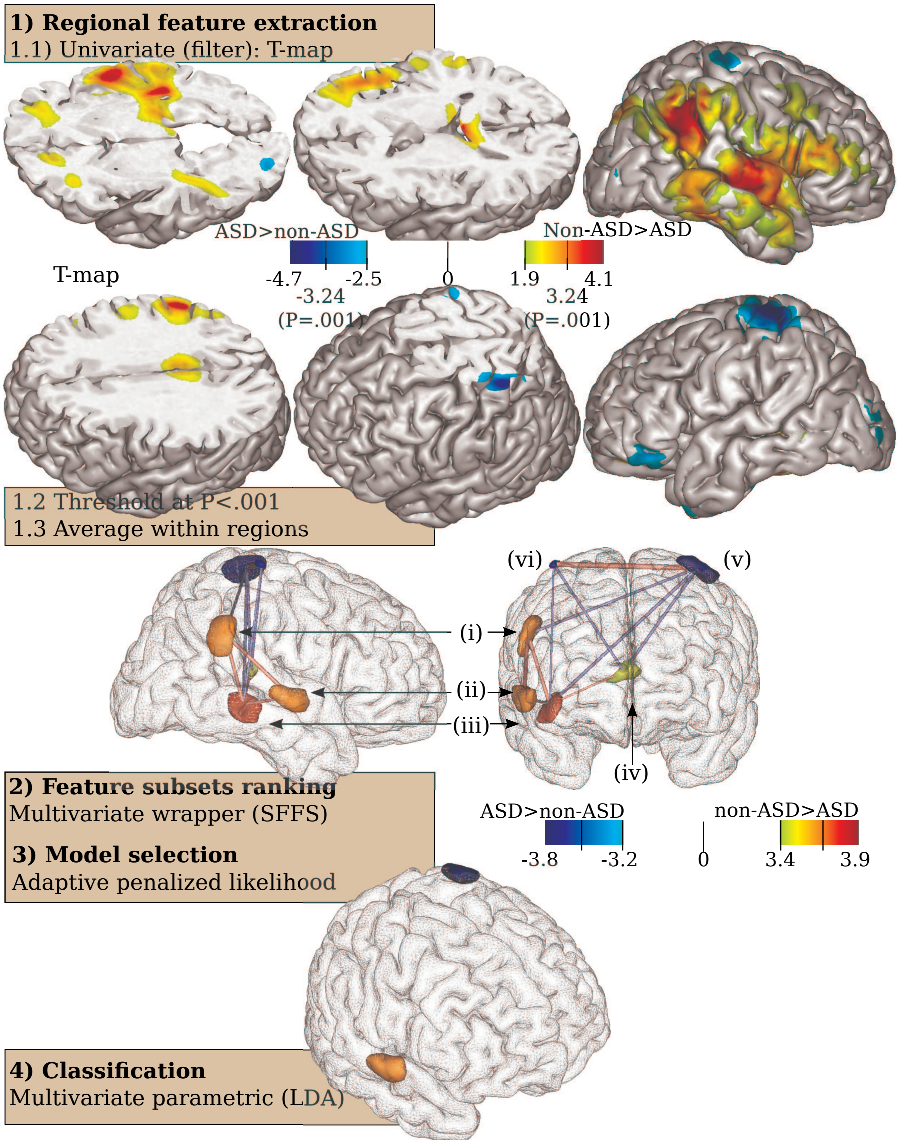

1) Regional feature extraction
1.1) Univariate (filter): T-map
T-map
ASD>non-ASD
-4.7
-2.5
-3.24
(P=.001)
0
Non-ASD>ASD
1.9
4.1
3.24
(P=.001)
1.2 Threshold at P<.001
1.3 Average within regions
(vi)
(v)
(i)
(ii)
(iii)
(iv)
2) Feature subsets ranking
Multivariate wrapper (SFFS)
3) Model selection
Adaptive penalized likelihood
ASD>non-ASD
-3.8
-3.2
0
non-ASD>ASD
3.4
3.9
4) Classification
Multivariate parametric (LDA)

Figure 2.9: In a first step: univariate *t*-statistics selected regions of hypoperfusion in the ASD group: (i) the right temporo-parietal junction (RTPJ); (ii) the right Superior Temporal Sulcus (STS); (iii) middle temporal gyrus; and (iv) the posterior zone of the corpus callosum where it overlaps with the right posterior cingulum and bilateral thalami. Two hyperperfused regions in the ASD group were identified in (v) the left post-central and (vi) the right pre-central areas. A second step of multivariate feature selection identified two most predictive regions (ii) the right STS and (v) the left postcentral whose provide more than 80% of balanced prediction accuracy.

2.7 Feature selection and classification based on whole brain VBM

We compared various regularization and features selection strategies applied to whole-brain VBM Gray matter image containing $\approx 360{,}000$ voxels within a brain mask. The first dataset (SCZ vs. CTL) contains

Gray Matter (from VBM) maps from 605 participants: 330 Controls (CTL) and 275 with patients with chronic schizophrenia (SZC). The second dataset (BD vs. CTL) contains Gray Matter (from VBM) maps from 662 participants: 356 Controls (CTL) and 306 patients with Bipolar Disorder (BD). Fig. 2.10 shows the comparison of the following strategies:

- ℓ_2 regularized Logistic Regression (LR). This setting acts as the baseline strategy.
- Lasso (ℓ_1 regularized) LR.
- ElsaticNet with $\ell_1\ell_2$ regularized LR.
- Filter with ℓ_2 regularized LR.
- RFE with ℓ_2 regularized LR.

2.7.1 Results

SCZ vs CTL

BD vs CTL

(a) Filter vs. RFE + LR

(b) Filter vs. RFE + LR

(c) ℓ_1, ℓ_2, ElasticNet LR

(d) ℓ_1, ℓ_2, ElasticNet LR

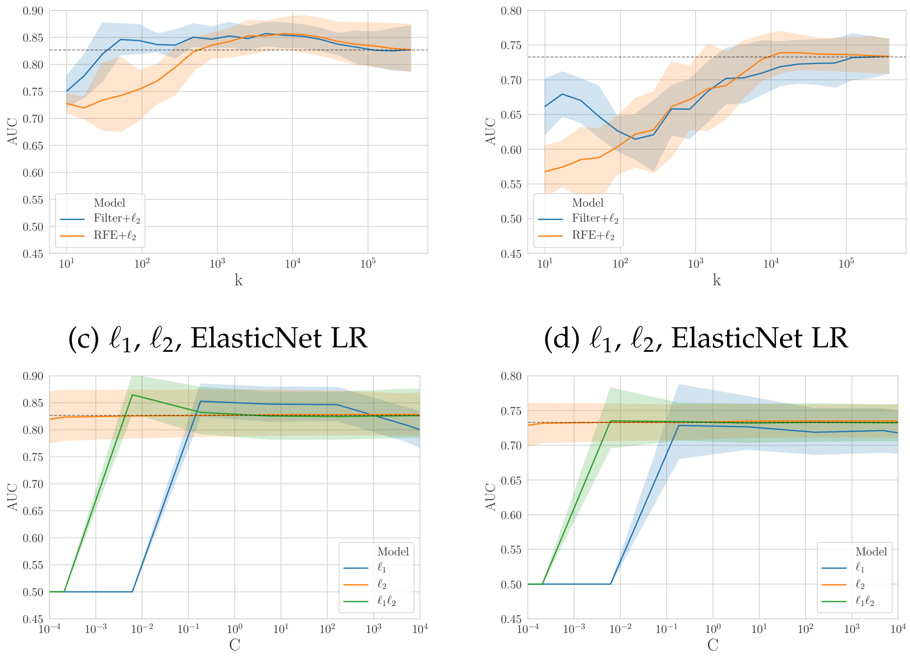


Figure 2.10: Sensitivity analysis: AUC as a function of model complexity parameter for two datasets, left column: SCZ vs. CTL, right column BD vs. CTL. (a, b) AUCs of Filter vs. RFE + ℓ_2 regularized Logistic Regression (LR) as a function of k the number of selected features. (c, d) AUCs of ℓ_1 (Lasso) vs ℓ_2 vs. ElasticNet- regularized LR as a function of the regularization parameter C. The dotted horizontal line provides the baseline performance of aℓ_2 regularized LR with $C = 1$.

Feature selection

- Fig. 2.10(a) shows that with SCZ vs. CTL dataset, feature selection (FS) provides a slight improvement ($\geq +0.05$) which is significant enough to be detected (i. e.larger that CV standard errors ± 0.05) by CV-based model selection. On BD dataset (Fig. 2.10(b)) the FS is dominated by baseline ℓ_2 LR, nevertheless CV model selection would pick a configuration (around 10^4 features) with similar performance.

- RFE does not out-perform a simple filter.

Regularization

- Similar conclusions can be drawn with ElasticNet or Lasso regularization: small improvement could be obtained on SCZ (Fig. 2.10(a)). Note that CV-based model selection would pick a parameter of regularization (C) with better or at least similar performances compared to baseline ℓ_2-regularization. On BD dataset (Fig. 2.10(d)), CV model selection will easily pick a value of C where ElasticNet, or Lasso, get the same performances than baseline ℓ_2 with the benefice of parsimony.

- Note that, we have explored an overly broad range. The common practices limit to a range between $[10^{-2}, 10^2]$, which matches the range where the sparse model either outperformed or compete with baseline ℓ_2. AUC equals 0.5 correspond to over-penalized settings (small C), leading to all coefficients being exactly null, which could be trivially avoided.

Conclusion

ℓ_2 regularization must be used in the first place, providing a baseline performance target. Then, some improvements could be obtained with ElasticNet/Lasso or basic univariate filtering. Here such strategy provides improvements on SCZ and similar performances on BD. Again, feature selection requires a stable model selection procedure with small standard error. Here, (5 folds)CV would achieve this goal thanks to the large size ≥ 600 of the dataset. On smaller datasets, the baseline ℓ_2 should be favored.

Model selection based on priors like the "VC dimension" would propose $k \approx 600$. This tendency to select too low k is due to the strong spatial correlation between the voxels. It would be necessary to calibrate such a procedure.

3

Imaging-genetics

This chapter presents the background in imaging-genetics Sec. 3.1. Then, Sec. 3.2 reviews multivariate analysis models. Finally, Sec. 3.4 presents our contribution (Le Floch et al., 2012) based on dimension reduction and ℓ_1-regularized multivariate latent variable models to face the very high dimensionality of imaging genetics studies.

3.1 Background in imaging genetics

Imaging genetics studies rely on the idea that neuroimaging data may be considered as a relevant intermediate phenotype (or endophenotype) to understand the complex path between genetics and behavioral or clinical phenotypes. Compared to the final phenotypes, this endophenotype is expected to be closer to genetics. In this context, we aimed to propose multivariate models to identify the part of genetic variability that explains some neuroimaging variability.

While other OMICs modalities can now be measured, here we address the processing of (Single Nucleotide Polymorphisms (SNPs)) issued from DNA (Deoxyribonucleic acid) arrays. SNPs are the few millions of nucleotides with some variability across the population. There exists a correlation structure within genetic data, called linkage disequilibrium (LD), which refers to the non-random association between the alleles of two genetic polymorphisms at two distinct positions (loci). It means that some pairs of alleles corresponding to the two loci are seen more often together on the same chromosome than expected by chance. This phenomenon is due to the physical link existing between neighboring loci on the same chromosome, called genetic linkage, and the fact that their alleles are inherited together from one generation to another. One cause of LD decay is recombination occurring during the crossing-over in meiosis. Thus, the larger the physical distance between two polymorphisms is, the higher is the probability of recombination and, thus, the lower the LD would be. There exist some regions with a higher recombination rate, called hot spots, and some others with a low recombination rate, called haplotype blocks.

LD reduces the number of genetic polymorphisms necessary to capture most of the genetic variability. Only a reduced number of independent and highly informative SNPs, called tagSNPs, should be genotyped. For instance, 99% of common SNPs (with a minor allele frequency > 5%) are tagged with an LD of $r^2 \geq 0.8$ by only one million of tagSNPS. Most of the commercial chips use these tagSNPs. A consequence is that in most cases, when a genotyped SNP is found to be associated with a given phenotype, it is not the causal SNP itself but in LD with the causal SNP.

3.2 Analysis strategies

Imaging genetics studies that include a large amount of data in both the imaging and the genetic components are facing challenges for which the neuroimaging community has no definitive answer so far.

Candidate genes approach

Current imaging genetics studies are often either limiting the brain imaging endophenotype studied to a few candidate variables but testing their relationship with a large number of SNPs as one usually proceeds during gene screening (e.g., Furlanello et al., 2003), or limiting the number of candidate SNPs or genes to be tested on the whole brain or some significant portion of it (e.g., Glahn, Thompson, and Blangero, 2007; McAllister et al., 2006; Roffman et al., 2006). When faced with both a large number of SNPs, and a large number of voxels, one has to design an appropriate analysis strategy that should be as sensitive and specific as possible.

Massive univariate linear model (MULM)

Without any priors on genetic or brain regions involved, exploratory methods can be used. The most straightforward approach to exploratory imaging genetics studies is clearly to apply a massive univariate analysis on both genetic and imaging data (Stein et al., 2010), which may be called Mass-Univariate Linear Modelling (MULM). However, while univariate techniques are simpler, they encounter a multiple comparison problem in the order of 10^{11}. Moreover, the link between genetic and imaging data is likely to be in part multivariate: epistasis or pleiotropy are likely phenomena in common traits or diseases. Indeed, brain imaging endophenotypes are probably influenced by the combined effects of several SNPs and different brain regions may also be influenced by the same SNP(s). A way to partially take into account epistasis may be to use a gene-based method for associating the joint effect of the different SNPs within each gene across the voxels of the whole brain (Hibar et al., 2011).

Two-block multivariate models

To address the limitations of univariate analysis, multivariate methods models have been proposed in imaging genetics studies. Such a joint analysis of two blocks can take advantage of two multivariate structure in the data:

- Within block correlation that stems from linkage disequilibrium.
- Between blocks multivariate relation that comes from epistasis or pleiotropy.

Two-block multivariate methods aim to discover "associations" between one genetic block (noted $\boldsymbol{X}$) and a second neuroimaging block (noted $\boldsymbol{Y}$). Where $\boldsymbol{Y}$ (of size $N \times q$) contains the q imaging phenotypes and $\boldsymbol{X}$ (of size $N \times p$) contains p genetic measurements (e. g.SNPs), N being the number of subjects.

Partial Least Squares (PLS) regression (Wold, Martens, and Wold, 1983) and Canonical Correlation Analysis (CCA) (Hotelling, 1936) appear to be good candidates to look for associations between two blocks, as they extract pairs of covarying/correlated latent variables from a linear combination of the variables of each block. Another approach has also been proposed by (Calhoun, Liu, and Adali, 2009) based on parallel Independent Component Analysis to combine functional MRI data and SNPs from candidate regions. Nevertheless, all these multivariate methods encounter critical overfitting issues due to the very high dimensionality of the data.

Fighting overfitting

To face these issues, methods based on dimension reduction Sec. 2.3 or regularisation Sec. 2.2 can be used.

As for regularization, multivariate methods based on ℓ_1 -penalization, like sparse Partial Least Squares (Chun and Keleş, 2010; Lê Cao et al., 2009, 2008; Parkhomenko, Tritchler, and Beyene, 2007, 2009; Waaijenborg, Witt Hamer, and Zwinderman, 2008; Witten and Tibshirani, 2009) or regularised CCA (Soneson et al., 2010), provided good results in correlating two blocks of data such as transcriptomic and metabolomic data, gene expression levels and gene copy numbers, or gene expression levels and SNP data. Vounou, Nichols, and Montana (2010) also introduced a similar promising method, called sparse Reduced-Rank Regression (sRRR) and based on ℓ_1-penalization, that they applied to a simulated dataset made of 1000s of SNPs and brain imaging data. The implementation of the method becomes equivalent to sparse PLS in high dimensional settings since they make the classical approximation that its diagonal elements may replace the covariance matrix of each block. However, whether these multivariate techniques can resist even higher dimensions remains an open question.

3.3 Regularized multi-blocks latent variables models

3.3.1 PLS

Partial Least Squares (PLS) regression is modeling the associations between two blocks of variables with the assumption that blocks are linked through unobserved latent variables. A latent variable (or component) corresponding to one block is a linear combination of the observed variables of this block.

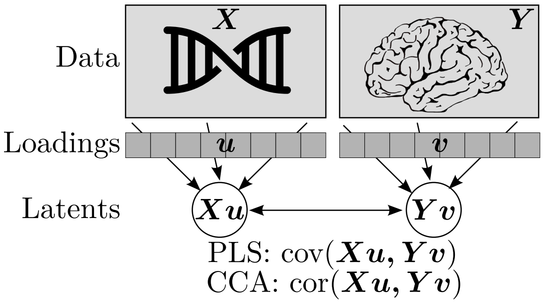


Figure 3.1: Two blocks latent variables models: PLS or CCA.

PLS builds a sequence of orthogonal latent variables for each block such that, at each step, the covariance between the pair of latent variables is maximal. For each step h in $1..H$, where H is the maximal number of pairs of components, it optimizes the following criterion:

$$\max_{\|\boldsymbol{u}_h\|_2=\|\boldsymbol{v}_h\|_2=1} \mathrm{cov}(\boldsymbol{X}_{h-1}\boldsymbol{u}_h, \boldsymbol{Y}_{h-1}\boldsymbol{v}_h) \tag{3.1}$$

$$\max_{\|\boldsymbol{u}_h\|_2=\|\boldsymbol{v}_h\|_2=1} \boldsymbol{u}_h'\boldsymbol{X}_{h-1}'\boldsymbol{Y}_{h-1}\boldsymbol{v}_h, \tag{3.2}$$

where $\boldsymbol{u}_h$ and $\boldsymbol{v}_h$ are the weight vectors for the linear combinations of the variables of blocks $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively. $\boldsymbol{X}_{h-1}$ and $\boldsymbol{Y}_{h-1}$ are the residual (deflated) $\boldsymbol{X}$ and $\boldsymbol{Y}$ matrices after their regression on the $h-1$ previous pairs of latent variables, starting with $\boldsymbol{X}_0 = \boldsymbol{X}$ and $\boldsymbol{Y}_0 = \boldsymbol{Y}$ (whose columns have been standardized).

There exist two ways of deflation: an asymmetric way (the original PLS regression) and a symmetric way (canonical-mode PLS). With asymmetric deflation, both blocks are deflated on the same the latent variables of block $\boldsymbol{X}$ (which becomes the predictor block), while with symmetric deflation, each bock is deflated on its latent variable. In our case, we are more interested in symmetric PLS as we investigate exploratory methods trying to extract covarying networks among a tremendous amount of neuroimaging and SNP data, many of which are very likely to be irrelevant. Note that, on the first pair of components, the original PLS regression and symmetric PLS give the same results.

This optimization problem is solved using the iterative algorithm called NIPALS (Wold, 1966) and, more precisely, its inner loop. The outer loop of NIPALS is the iteration over the number of pairs of components. The optimal vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ are the first pair of singular vectors of the matrix $\boldsymbol{X}'\boldsymbol{Y}$.

3.3.2 CCA

A similar method is Canonical Correlation Analysis (CCA) which maximizes the correlation between the two latent variables:

$$\max_{\|\boldsymbol{u}_h\|_2=\|\boldsymbol{v}_h\|_2=1} \operatorname{corr}(\boldsymbol{X}_{h-1}\boldsymbol{u}_h, \boldsymbol{Y}_{h-1}\boldsymbol{v}_h) \tag{3.3}$$

$$\max_{\|\boldsymbol{u}_h\|_2=\|\boldsymbol{v}_h\|_2=1} \frac{\boldsymbol{u}_h'\boldsymbol{X}'\boldsymbol{Y}\boldsymbol{v}_h}{\sqrt{\boldsymbol{u}_h'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{u}_h}\sqrt{\boldsymbol{v}_h'\boldsymbol{Y}'\boldsymbol{Y}\boldsymbol{v}_h}} \tag{3.4}$$

The solution may be obtained by computing the SVD of $(\boldsymbol{X}'\boldsymbol{X})^{-1/2}\boldsymbol{X}'\boldsymbol{Y}()\boldsymbol{Y}'\boldsymbol{Y})^{-1/2}$. The successive pairs of weight vectors $\boldsymbol{u}_h$ and $\boldsymbol{v}_h$ are obtained by: $\boldsymbol{u}_h = \boldsymbol{X}'\boldsymbol{X}^{-1/2}\mathbf{e}$ and $\boldsymbol{v}_h = \boldsymbol{Y}'\boldsymbol{Y}^{-1/2}\mathbf{f}$, where the columns of $\mathbf{e}$ and $\mathbf{f}$ are the left and right singular vectors respectively.

CCA is more prone to overfitting than PLS: it requires the inversion of the scatter matrices $\boldsymbol{X}'\boldsymbol{X}$ and $\boldsymbol{Y}'\boldsymbol{Y}$, which are ill-conditioned in our high-dimensional settings with very large p and q (numbers of variables for blocks $\boldsymbol{X}$ and $\boldsymbol{Y}$ respectively) and a small N (number of observations or individuals).

3.3.3 ℓ_2 Regularization of CCA

ℓ_2-regularization alleviates the risk of overfitting and the non-invertibility issues of CCA. Assuming data scaling (to zero mean and unit standard deviation), ℓ_2-regularization of CCA consists of modifying the correlation matrices $\boldsymbol{X}'\boldsymbol{X}$ and $\boldsymbol{Y}'\boldsymbol{Y}$ by $\boldsymbol{X}'\boldsymbol{X} + \lambda_2\boldsymbol{I}$ and $\boldsymbol{Y}'\boldsymbol{Y} + \lambda_2\boldsymbol{I}$ respectively. However, in high-dimensional space, the approximation is often made that the covariance matrices may be replaced by identity matrices. Such extreme ℓ_2-regularization of CCA shrinks the loading coefficients and makes CCA equivalent to PLS-SVD and thus to PLS regression as well on the first component.

3.3.4 ℓ_1 Regularization of PLS

PLS is a ℓ_2 regularized CCA, moreover, authors (Lê Cao et al., 2008; Tibshirani, 1996) proposed to add a sparsity promoting ℓ_1 penalty to deal with overfitting issue. It should be noted that ℓ_1 penalization may not be easily implemented on PLS-SVD without loosing the orthogonality constraint on weight vectors (Zou, Hastie, and Tibshirani, 2006). Sparse PLS (sPLS) minimization problem is given by:

$$\min_{\|\boldsymbol{u}\|_2=\|\boldsymbol{v}\|_2=1} -\boldsymbol{u}'\boldsymbol{X}'\boldsymbol{Y}\boldsymbol{v} + \lambda_{1X}\|\boldsymbol{u}\|_1 + \lambda_{1Y}\|\boldsymbol{v}\|_1 \tag{3.5}$$

where λ_{1X} and λ_{1Y} are ℓ_1-penalization parameters for the weight vectors (loadings) of blocks $\boldsymbol{X}$ and $\boldsymbol{Y}$ respectively. The sPLS criterion is bi-convex in $\boldsymbol{u}$ and $\boldsymbol{v}$ and may be solved iteratively for $\boldsymbol{u}$ fixed or $\boldsymbol{v}$ fixed, using soft-thresholding of variable weights at each iteration of the NIPALS inner loop. Weight vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ are computed using the following Algo. 1:

Algorithm 1 Sparse PLS $(\boldsymbol{X}, \boldsymbol{Y}, \lambda_{1X}, \lambda_{1Y})$

```
Initialize $u$ and $v$ using for instance the first pair of singular vectors of the matrix $X'Y$ and normalise them.
repeat
    for fixed $v$ do
        $\widehat{u} = \arg\min_{\|u\|_2=1} -u'X'Yv + \lambda_{1X}\|u\|_1 = g_{\lambda_{1X}}(X'Yv)$   ▷ where $g_\lambda(y) = \mathrm{sign}(y)(|y| - \lambda)_+$ is the soft-thresholding function.
        Normalise $u = u/\|u\|_2$.
    end for
    for fixed $v$ do
        $\widehat{v} = \arg\min_{\|v\|_2=1} -u'X'Yv + \lambda_{1Y}\|v\|_1 = g_{\lambda_{1Y}}(Y'Xu)$
        Normalise $v = v/\|v\|_2$
    end for
until Until convergence of $u$ and $v$
return $u, v$
```

Sparse versions of CCA have also been proposed by Parkhomenko, Tritchler, and Beyene (2007, 2009), Waaijenborg, Witt Hamer, and Zwinderman (2008), and Witten and Tibshirani (2009). However, to solve the non-invertibility issue, they make the approximation that the covariance matrices $\frac{1}{n-1}\boldsymbol{X}'\boldsymbol{X}$ and $\frac{1}{n-1}\boldsymbol{Y}'\boldsymbol{Y}$ may be replaced by their diagonal elements, which makes sparse CCA equivalent to sparse PLS.

3.4 Feature selection and sparse PLS reveal the genetic polymorphisms that explain brain acrivation

In Le Floch et al., 2011, we proposed latent based multivariate models to identify to identify the part of genetic variability (SNPs) that explains some neuroimaging functional variability (fMRI).

In this work, we address the overfitting risk in a very high dimension by combining dimension reduction on SNPs, either by PCA or univariate filtering, before applying (sparse) PLS or (regularized) CCA. We first use a simulated dataset mimicking fMRI and genome-wide SNP data and compare the performances of the different methods, by assessing their positive predictive value, as well as their capacity to generalize the link found between the two blocks on unseen data with a cross-validation procedure. Indeed, we first compared PLS and CCA, and then we investigated the influence of ℓ_2 regularization on CCA and ℓ_1 regularization on PLS. Finally, we evaluated the potential benefice of dimension reduction using either PCA or filtering.

Finally, we apply these different methods with the same cross-validation procedure on a real dataset made of fMRI and genome-wide SNP data. The statistical significance of the link obtained on "test" subjects is assessed with randomization techniques.

3.4.1 Methods

Univariate feature selection consisted of (i) $p \times q$ pair-wise linear regressions based on an additive genetic model; (ii) ranking the SNPs according to the minimal p-value each SNP gets across all phenotypes, and (iii) keeping the set of SNPs with the lowest "minimal" p-values.

In this study, $N = 94$ subjects were genotyped and participated in a general cognitive assessment fMRI task described in Pinel et al. (2007). With fMRI data we focused only on two activation contrasts: *reading minus checkerboard viewing* and *speech comprehension minus rest*. After the usual subject-level processing, we selected 34 brain regions of interest (ROIs): 19 from the "reading" contrast and 15 from the "speech comprehension" contrast. We identified the 34 mirror ROIs by symmetry for the inter-hemispheric plane. Finally, 34 lateralization indices (normalized right-left) were derived from those regions.

Preprocessing provided two blocks of data $\boldsymbol{Y}$ (fMRI) and $\boldsymbol{X}$ (genetics) of size 94×34 and $94 \times 622,534$ respectively.

We compared the following models: (i) *MULM*: Mass Univariate Linear Modelling; (ii) *PLS*: Partial Least Squares; (iii) *KCCA*: Kernel Canonical Correlation Analysis; (iv) *sPLS*: sparse PLS with a re-parametrization of the sparsity parameter: The raw λ_{1X} and λ_{1X} (used in Algo. 1) have been replaced by selection rates, $s_{\lambda_{1X}}$ and $s_{\lambda_{1Y}}$, as the number of selected variables from each block out of the total number of variables of that block. In our case, we chose to apply sparsity on SNPs only and to set $s_{\lambda_{1Y}}$ to 1 for imaging phenotypes, as we had a very large number of SNPs and only a few (34) imaging phenotypes. (v) *rKCCA*: ℓ_2 penalized KCCA with λ_2 being the regularization parameter: (Full CCA is obtained with $\lambda_2 = 0$, large λ_2 leads to PLS); (vi)

PCPLS: Principal Component Analysis + PLS; (vii) *PCKCCA*: Principal Component Analysis + KCCA; (viii) *f(s)PLS*: Filtering + (sparse) PLS; (ix) *f(r)KCCA*: Filtering + (regularised) KCCA.

Generalization performances were assessed through cross-validation by computing the correlation between the pair of latent $\boldsymbol{X}_{test}\boldsymbol{u}$, $\boldsymbol{Y}_{test}\boldsymbol{v}$ obtained by multiplying the test sample $\boldsymbol{X}_{test}$, $\boldsymbol{Y}_{test}$ by the coefficients (and any other dimension reduction) estimated on the training data only.

Finally, in the case of simulated data, since the ground truth was known, we could evaluate the precision of methods by computing the Positive Predictive Value (PPV=number of true positives/number of positive calls).

3.4.2 Results on simulated data

Influence of regularization

Fig. 3.2 compares PLS, CCA and the influence of regularization when the number of SNPs p increases, from 200 (mostly made of the 198 informative features) up to 85,772 SNPs (mostly made of noise).

1. Fig. 3.2 (a): pure CCA, (rKCCA without regularization $\lambda_2 = 0$), suffers from overfitting as soon as irrelevant features are added in the model. Such a result highlights the limits of pure CCA to deal with situations where the number of training samples (100) is smaller than the dimension (p = 200).

2. Fig. 3.2 (b): with low-dimensional datasets $p \leq 100$ containing only informative features and with suitable ℓ_2-regularization ($\lambda_2 = 100$), rKCCA outperformed other methods, notably all (sparse) PLS. This results with an "optimal" dataset, was expected since the evaluation criterion (correlation between factorial scores) is exactly the one which is maximized by CCA.

3. Fig. 3.2 (c): Superiority of PLS over their CCA counterparts is observed when the dimensionality increases, adding irrelevant features. More notably, sPLS dominates rKCCA: the performance of rKCCA rapidly decreases while sPLS ($s_{\lambda_{1X}} = 0.1$, i.e., 10% of feature are selected) tolerates an increase of the dimensionality up to 1,000 features before its performance starts to decrease. One may note that as expected theoretically, along with the increase of penalization (λ_2), rKCCA curves smoothly converge toward PLS.

On the second component pair, the results are less interpretable. However, (s)PLS curves are above the rKCCA ones.

The four graphs on the right panel of Fig. 3.2 shows precision (PPV) curves computed for each pattern separately.

1. Fig. 3.2 (d): precision on the first genetic component appears to be much higher for the first pattern than for the second pattern, especially in low dimensions, while the opposite trend is observed on the second genetic component. This demonstrates that the first causal pattern is captured by the first component, while the second pattern is captured by the second component.
2. Fig. 3.2 (e): An increase of ℓ_2 penalty on CCA improves the precision on respective components.
3. Fig. 3.2 (f): Adding some ℓ_1 penalty to PLS shows little improvement in precision.

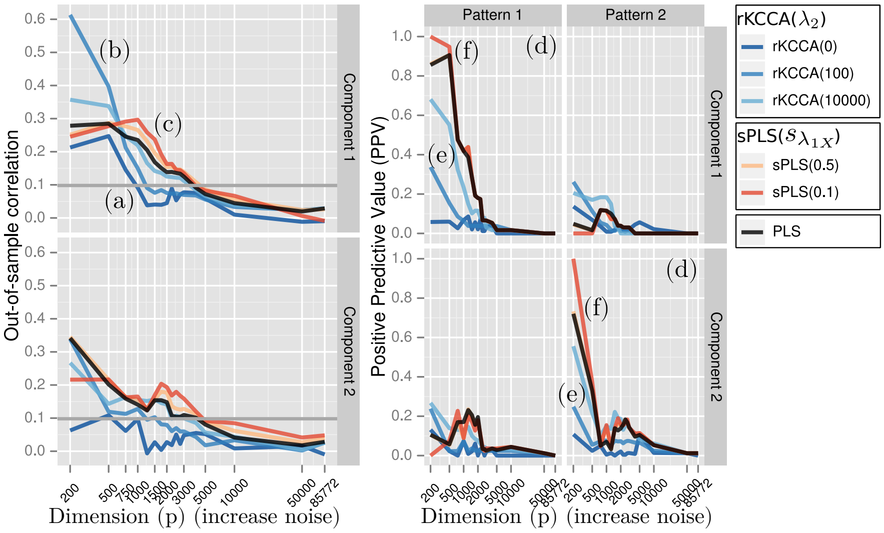


Figure 3.2: Simulated dataset: comparison of regularization methods to deal with increasing irrelevant features. The total number of features varies along the x-axis between 200 (mostly informative) to 85,772 non-causal SNPs. We compared: (i) in blue, regularized kernel CCA (rKCCA) with various ℓ_2 regularization values (λ_2) ranging from 0 (pure CCA) to 10,000; (ii) in black, PLS; (iii) in red, sparse PLS (sPLS) with various ℓ_1 regularization values parameterized as sparsity rate ($s_{\lambda_1 X}$) ranging from 0.5 (50% of input features have a non null weight) to 0.1. The y-axis of the two left panels shows the (5-fold CV) average out-of-sample correlation coefficients between the two first component pairs. The four right panels present the power of the methods to identify causal SNPs implied in the two causal patterns. The y-axis depicts the precision (PPV) for each of the two first genetic components ($\boldsymbol{u}_1, \boldsymbol{u}_2$).

Influence of the dimension reduction step

Then we investigated the influence of the first step of dimension reduction. Fig. 3.3 presents the influence of different dimension reduction strategies: Principal Component Analysis (PC), filter (f), sparse (s), and combined filter+sparse (fs) methods. Here the parameter setting, 50 selected SNPs, was derived from the known ground truth (56 true causal SNPs). The 50 SNPs were either the 50 best-ranked SNPs for (f) methods or the 50 non-null weights for sparse PLS or, a combination of both (10% of the 500 best-ranked SNPs or 50% of 100 for fsPLS).

1. Fig. 3.3 (a): shows that all PC-based methods (green curves) failed to identify generalizable covariations between blocks when the number of irrelevant features increases.

2. Fig. 3.3 (b): Filtering slightly improved the performance of CCA and greatly those of PLS: simple fPLS($k = 50$) is the second-best approach in our comparative study. Filtering with regular PLS outperformed sparse PLS.

3. Fig. 3.3 (c): Finally the best performance is obtained by combining filtering and ℓ_1 regularization: fsPLS($k = 100$, $s_{\lambda_{1X}} = 0.5$), which keeps 100 SNPs after filtering and selects 50% of those SNPs by sPLS.

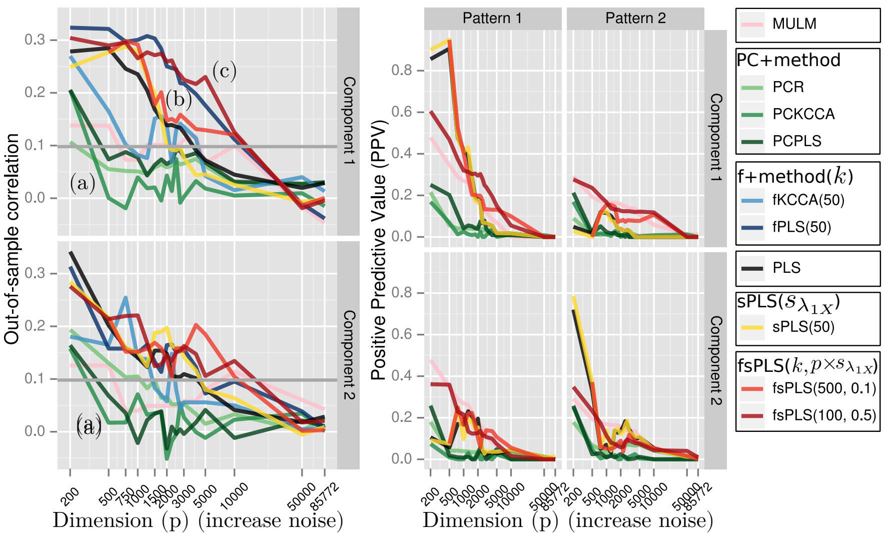


Figure 3.3: Simulated dataset: comparison of regularization methods to deal with increasing irrelevant features. The total number of features varies along the x-axis between 200 (mostly informative) to 85,772 non-causal SNPs. In greens, Principal Component (PC) Analysis based methods: PC regression (PCR), PCA+KCCA (PCKCCA), PCA+PLS (PCPLS). In blues, filter (f) based methods: f+KCCA (fKCCA), f+PLS (fPLS). We selected only the 50 best SNPs, while according to ground truth, 56 SNPs were identified as causal. In black, PLS. In yellow, sparse PLS (sPLS) where selection rate, $s_{\lambda_{1X}}$, is such that 50 features have a non-null weight. In reds, filter + sparse PLS (fsPLS) with settings both leading to 50 selected features: fsPLS($k = 500$, $s_{\lambda_{1X}} = 0.1$) (resp. fsPLS($k = 100$, $s_{\lambda_{1X}} = 0.5$)) keeps the 500 (resp. 100) best ranked features and then 10% (50%) get a non-null weight. Finally, in pink, we add MULM. The y-axis of the two left panels shows the (5-fold CV) average out-of-sample correlation coefficients between the component pairs. The four right panels present the power of the methods to identify causal SNPs implied in the two causal patterns. The y-axis depicts the precision (PPV) when 50 SNPs are selected for each of the two first genetic components: $(\boldsymbol{u}_1, \boldsymbol{u}_2)$.

3.4.3 Results on experimental data

Comparative analysis

Tab. 3.1 summarizes the two first average correlation between pairs of latent variable for the different methods tested.

	Correlation between PC pairs			
Methods	ρ^1_{test}	ρ^2_{test}	$\rho^1_{training}$	$\rho^2_{training}$
PLS	-0.092	0.218	0.990	0.984
sPLS ($s_{\lambda_{1X}}$=0.1%)	0.008	0.201	0.938	0.922
fPLS (k=1000)	0.236	**0.268**	0.962	0.953
fsPLS (k=1000,$s_{\lambda_{1X}}$=5%)	**0.432**	0.210	0.772	0.788

Table 3.1: The correlation (averaged across CV-folds) between pairs $h \in \{1, 2\}$ of components: $\text{corr}(\boldsymbol{X}_{h-1}\boldsymbol{u}_h, \boldsymbol{Y}_{h-1}\boldsymbol{v}_h)$, where $\boldsymbol{X}_0, \boldsymbol{Y}_0$ are the original datasets. Correlation are given for left-out "test" sample (ρ^h_{test}) and "training" sample $\rho^h_{training}$.

- Both PLS and ℓ_1-regularized sPLS failed to identify a generalizable imaging-genetic link in such high dimensions and captured only noise. Overfitting occurred in training data.

- As found with simulated dataset, univariate filtering combined with PLS (fPLS) significantly improved the performance of PLS.

- As found with simulated dataset, the best performance is obtained with a combination of filtering ℓ_1-regularized sPLS. This strategy succeeds in identifying a generalizable imaging-genetic link.

Sensitivity analysis of fsPLS and significance assessment

Tab. 3.2 shows sensitivity analysis to assess the influence of the sparse PLS penalization parameter $s_{\lambda_{1X}}$ and the number k of SNPs kept by the filter. The significance was assessed with random permutation and Westfall and Young (maxT) correction for multiple testing (**westfall_resampling-based_1993**; Dudoit, Shaffer, and Boldrick, 2003). The performances are driven by the filter parameter k: once fixed to 1,000, good performances are obtained on a wide range of values for the sparsity parameter.

	Ratio of sparsity: $s_{\lambda_{1X}}$						
k	1%	5%	10%	25%	50%	75%	100%
10	0.041	0.041	0.041	0.041	0.144	0.112	0.112
100	0.182	0.074	0.085	0.057	0.069	0.188	0.243
1000	0.151	**0.432***	**0.414***	0.400	0.317	0.285	0.236
10000	0.004	0.120	0.130	0.027	-0.006	-0.031	-0.061

Table 3.2: Out-of-sample (test) correlation on the first component pair as a function of k and $s_{\lambda_{1X}}$. Empirical p-values still significant ($p < .05$) after correction are shown here as: *.

Imaging genetics findings

Among the 50 SNPs selected by fsPLS, some of them were located within a gene. Eighteen genes were identified including PPP2R2B and RBFOX1, that were reported to be linked with ataxia and poor coordination of speech and body movements, and PDE4B, which is associated with schizophrenia and bipolar disorder.

3.4.4 Discussion

1. The overwhelming superiority of PLS over CCA demonstrates that ℓ_2-regularization is a prerequisite.
2. With such high-dimensional data, simple univariate filtering was essential.
3. ℓ_1 brings a valuable contribution.
4. The model selection problem has not been addressed in this work.

4

Integrating spatial regularization

Given the limitations of classical sparse algorithms to produce stable and interpretable predictive signatures, I initiated a research program to extend regularization with spatial constraints (Sec. 4.1). This objective required the design of a new solver that scales to "real life" high-dimensional data ($\leq 300,000$ input features).

- Sec. 4.2 presents *CONESTA* (Hadj-Selem et al., 2018) our original solver for high-dimensional structured input data such as 3D images, meshes of the cortical surface or genetic data (with LD structure). This solver was originally designed for supervised linear problems. This solver (applied to many classification problems) has been released through an open-source python library ParsimonY.
- Sec. 4.3 proposes an application of CONESTA to unsupervised Principal Component Analysis with spatial regularization (de Pierrefeu et al., 2018c).

4.1 Interpretable ML: spatial regularization

4.1.1 Interpretable predictive maps: toward brain signatures

Although good performances are achieved by linear classifiers working on the whole brain at the voxel level, it is difficult to interpret the contribution of brain regions in the prediction of the target variable. Indeed, most of the state-of-the-art classifiers, such as linear SVM (or any ℓ_2 regularized linear model), produce a dense coefficient map with rapid sign flipping. In other words, most of the whole brain contributes to the prediction, and more puzzling, nearby voxels (within the same brain regions) have an opposite effect on the prediction. Therefore, such predictors do not provide objective neuroanatomical markers on which a clinical decision is built: the solution must provide meaningful predictive patterns to reveal the neuroimaging markers of the pathology. In the context of predictive signature discovery, it is crucial to understand the brain's structural patterns that underpin the prediction. This absence of interpretability of the decision is ruling out the prospect of clinical application.

Our view of interpretable predictive pattern:

1. Pattern must be organized as clearly delimited regions, i. e.clusters of connected imaging measurements that can be interpreted by clinicians.
2. Pattern must be stable, i.e., another set of subjects sampled in the same population should produce a similar predictive pattern.

4.1.2 Limitations of regional features

A possibility would have been to use regional approaches (ROIs), based on the parcellation of cortical surface (Desikan et al., 2006) and the segmentation of subcortical structures. Indeed, such approaches reduce the dimensionality of the problem to several tens of measurements, simplifying the solution interpretability. Most of the multi-subject analyses require proper alignment of brain structures across subjects. However, ROIs add the assumption that the spatial extent of the searched imaging marker match with the size of the atlas-based pre-defined regions.

In a comparison study (Cuingnet et al., 2011), voxels based approaches outperformed region based approach. It suggests that the gain obtained by dimensionality reduction does not compensate for the loss of informative signal caused by averaging within pre-defined regions. Let's provide an illustrative scenario of the problem: Suppose that we look for the structural markers that best predict the response to antidepressant medication in patients with unipolar depression. A regional approach produces the volume of many structures, including the whole hippocampus. However, neurogenesis associated with antidepressants occurs entirely in a fairly small subsection of the hippocampus: the dentate gyrus (Sapolsky, 2001). Thus, the measure of the whole hippocampal volume mixes up the specific increase of the dentate gyrus with non-specific volume variation of the rest of the hippocampus. Such, partial volume effect, reduces the capacity to detect the informative signal and prevent to identify the relevant localized biomarker within the dentate gyrus. Hippocampal subfield segmentation (Van Leemput et al., 2008) could alleviate this problem, however, the segmentation of such small structures (Wisse, Biessels, and Geerlings, 2014) still an open debate on low resolution (1 mm^3) T1 images. More generally, the multiplication of many fine-grained pre-defined regions, to capture small effects, increases the risk of poorly matched (or defined) regions across individuals. Moreover, this brings us back to a situation similar to the one encounter at the voxel level. Thus, atlas-based regional approaches may overlook a searched effect with an unknown spatial extent.

4.1.3 Limitations of classical penalties applied to voxel-wise features

Thus, it seems appropriate to process whole-brain data at voxel/vertex level. On such high-dimensional space, ℓ_2 penalty (e.g., linear SVM) produces dense and, which is more questionable, irregular solutions with abrupt and high-frequency changes of values and coefficients sign. Although some methods exist to define thresholds to uncover brain regions than significantly contribute to the classification process (Gaonkar and Davatzikos, 2013; Wang et al., 2007), they do no produce interpretable weight maps per se.

ℓ_1 regularization produces sparse, scattered, and unstable solutions. In both cases, the weight maps are hard to interpret in terms of neuroanatomy. The combination of both penalties in ElasticNet, promotes sparse models while still maintaining the regularization properties of the ℓ_2 penalty. However, a major limitation of the ElasticNet penalty is that it does not take into account the spatial structure of brain images, leading to either the problem of ℓ_2 penalty or, more generally, the scattered patterns of ℓ_1.

The more straightforward approach to get interpretable maps could be to integrate spatial smoothing within the learning procedure. This option has been explored in Cuingnet et al., 2013. Authors show that a kernel-based method (SVM) that incorporates smoothing within the pre-computation of the similarity (Gram) matrix is equivalent to the trivial smoothing of the 3d images.

4.1.4 Spatial regularization: GraphNet and Total Variation

Without strong priors, the efficient scale (region or voxel) depends on the target to be predicted. Therefore, the scale should be determined during the learning process. Such automatic, data-driven identification of regions that best predict the target has motivated the use of structured sparsity.

One solution to extract brain regions at the relevant scale is to take benefit of the known structure of brain MRI images, to force the solution to adhere to biological priors, thereby producing more plausible and interpretable solutions. Indeed, MRI data is naturally encoded on a 3-dimensional grid where some voxels are neighbors, and others are not. Structured sparsity can be achieved with several penalties. Here, we propose to encode the spatial structure of the images or meshes as a penalty on the spatial gradient of the solution.

GraphNet penalty

One of the penalty is the Graph-constrained Elastic-Net, GraphNet (GN), described in (Dohmatob et al., 2015; Grosenick et al., 2013).

GraphNet closely resembles the Elastic-Net, but with a modification of the ℓ_2 -norm penalty term:

$$\min_{w} L_\varepsilon(w) + \lambda_1 \|w\|_1 + \lambda_2 \|w\|_2^2 + \lambda_G \|\nabla w\|_2^2, \tag{4.1}$$

where ∇ denotes a finite difference spatial gradient operator acting upon an image. For a 3D grid of size $p = p_x p_y p_z$, flattened into a long vector, we have $\nabla \in \mathbb{R}^{3p}$. It promotes local smoothness of the weight map by forcing adjacent voxels/vertices to have similar weights, and it does this by imposing a squared ℓ_2 penalty on the gradient of the weight map. The GN penalty induces smoothness by penalizing the size of the pairwise differences between coefficients that are adjacent in the graph. Therefore, *GN* promotes smooth change rather than piecewise constant structure in the non-sparse parts of the weight map. Such an outcome is of interest if we expect the magnitudes of nonzero coefficients to be different within a volume of interest. Concerning minimization, GN is differentiable, its combination with Lasso is solved with any proximal gradient method such as FISTA (Beck and Teboulle, 2009a).

However, to identify predictive regions of a clinical condition, it is desirable to produce clearly delineated piecewise constant structures. Therefore, we explored the potential of TV-ElasticNet penalty.

TV-Enet penalty

The Total Variation (TV) penalty is widely used in image denoising and restoration. It accounts for the spatial structure of images by encoding piecewise smoothness and enabling the recovery of homogeneous regions separated by sharp boundaries.

TV penalty forces sparsity on the spatial derivatives of the weight map (using an ℓ_{12}-norm), segmenting the weight map into spatially-contiguous parcels with almost constant values (Michel et al., 2011). TV can be combined with sparsity-inducing penalties (such as ℓ_1 (Lasso) (Tibshirani, 1996) to obtain segmenting properties that extracts predictive regions from a noisy background with zeros (Dohmatob et al., 2014; Dubois et al., 2014; Gramfort, Thirion, and Varoquaux, 2013). TV, together with Lasso, produces the desired foreground-background segmentation by imposing constant-valued parcels.

The ElastiNet-TV (Enet-TV) minimization problem is given by:

$$\min_{w} L_\varepsilon(w) + \lambda_1 \|w\|_1 + \lambda_2 \|w\|_2^2 + \lambda_s \|\nabla w\|_{2,1}, \tag{4.2}$$

where λ_1, λ_2 and λ_s are the hyper-parameters controlling the relative strength of each penalty.

4.1.5 Experimental evaluation of spatial penalties

This section compare spatial regularization: (i) Total Variation (TV) and (ii) GraphNet (GN) with classical ElascticNet (Enet) penalty of whole brain gray matter images.

3D T1 MRI Datasets

1. *Dep-TMS* dataset - Patients with treatment-resistant depression (TRD) where investigated with baseline structural MRI (sMRI) prior to a treatment with transcranial magnetic stimulation (TMS) Paillère Martinot et al., 2011. Followup response to active TMS was defined as having 50% depression score decrease. After quality control of the images, we could collect baselines sMRI of 18 responders and 16 non-responders. The aim was to learn a prognostic model of the treatment response from baseline 3D T1 MRI using (i) VBM 3D GM maps and (ii) cortical thickness.

2. *NUSDAST* dataset - Participants under 50 years old were selected from the NUSDAST cohort (Wang et al., 2013) leading to 97 patients with schizophrenia, according to DSM-IV criteria, and 139 healthy controls. The aim was to learn a diagnostic model of the clinical status using (i) VBM 3D GM maps, and (ii) cortical thickness.

3. *ADNI* dataset - 81 patients with a diagnosis of mild cognitive impairments (MCI) from the ADNI database who converted to AD within two years during the follow-up period where compare with 120 healthy controls elderly subjects. The aim was to learn a prognostic model of the conversion to AD from baseline 3D T1 MRI using (i) VBM 3D GM maps and (ii) cortical thickness.

Sensitivity analysis

GraphNet provides similar Fig. 4.1 performances than ElasticNet, whatever its contribution. The right panel shows that increasing GN slowly increases the stability of the coefficient maps. This result calls for the use of GraphNet instead of Enet without much risk. Nevertheless, there is no increase in prediction and a moderate improvement of stability (+10%).

TV-Enet is more sensitive to settle: too large global penalty ($\alpha = 1$, dark plain blue line) leads to a collapse of prediction performances but a tremendous increase of stability (+50%). An inspection of the coefficient maps shows vast regions covering almost the whole brain. This goes against the requirement of clearly delimited regions.

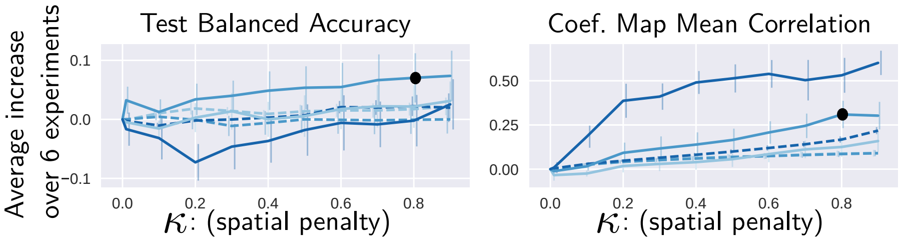


Figure 4.1: Sensitivity analysis of the spatial penalization parameter averaged across 6 experiments: the 3 datasets for both 3D GM and cortical thickness. Y-axis reports the differences of score (balanced accuracy in left panel or map stability in right panel) with the baseline ElasticNet that are averaged across the experiments for an increased spatial penalty in TV (plain lines) or GN (dashed lines). The models Eq. 4.1 and Eq. 4.2 were re-parameterized with only two parameters: α which determines the global penalty and κ which controls the ratio of ℓ_2 over the spatial penalty. The ElasticNet penalty has a fixed ℓ_1/ℓ_2 ratio of 0.1 with $\lambda_1 = \alpha/10$ and $\lambda_2 = \alpha$. Finally, $\lambda_{s\in\{TV,GN\}} = \alpha\kappa$ with $\kappa \in [0,1]$. The darkness of blue lines codes for the global strength of the penalty: dark $\alpha = 1$, medium $\alpha = 0.1$, and light blue $\alpha = 0.01$. Left panel: the balanced accuracy. Right panel: stability of the coefficient map as the average correlation (using Fisher's z-transformation) between pairs of coefficient maps computed across the 5 CV-folds.

On the other side, too small global penalty ($\alpha = 0.01$, light plain blue line) neither harm nor improve the performance and the stability compared to ElasticNet.

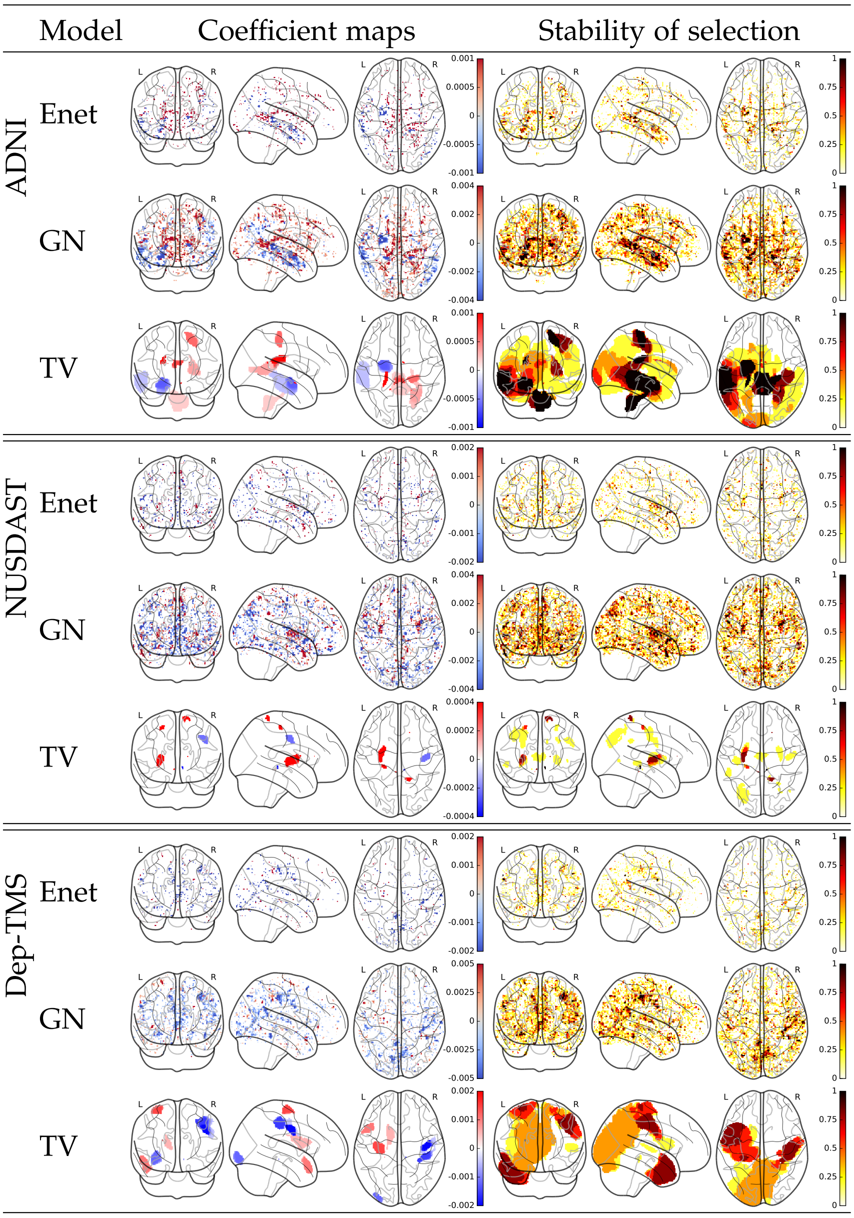
An intermediary global penalty ($\alpha = 0.1$, medium plain blue line) progressively improves both the prediction performances and the stability. Visual inspection show that with $\kappa = 0.8$ provides clearly delineated regions for all the experiments, as shown in Figs. 4.2 and 4.3, with a significant improvement of performances +7% and stability +30%.

Visual evaluation of the predictive maps

Left panels of Figs. 4.2 and 4.3 provides the coefficient maps obtained with ElasticNet (Enet), GraphNet (GN) and Enet-TV (TV) with the similar parameters settings for all models: $\alpha = 0.1$, ℓ_1/ℓ_2 ratio of 0.1 and $\kappa = 0.8$ for TV and GN. The right panel presents a measure of the stability of the coefficient maps as the proportion of selection across CV-folds.

Enet and GraphNet both provide scattered and unstable maps. Such a result was expected with Enet, but the solution provided by GraphNet was disappointing in terms of interpretation. Meanwhile, the predictive maps obtained with TV-Enet classifier appear much more interpretable, since it provides a smooth and stable map made of few identifiable regions.

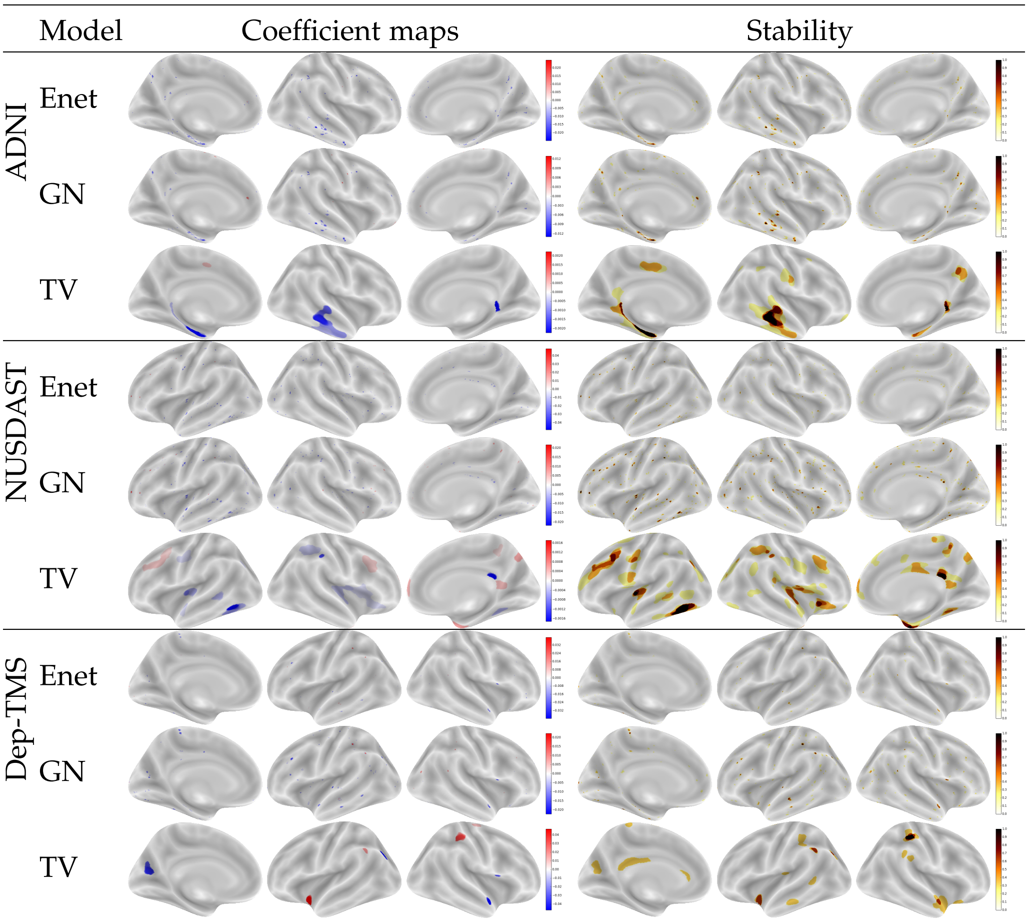
Figure 4.2: Left panel: coefficient maps from 3D gray matter images obtained by the three models: GM ElasticNet (Enet), GraphNet (GN), and Enet-TV (TV) on the three datasets: ADNI, NUSDAST, and Dep-TMS. Right: stability of the coefficient maps as the proportion $\in [0,1]$ of selection across CV-folds for each voxel, i. e., the dark value of 1 means that the voxel is always selected.



Conclusion

1. GraphNet can be safely used in replacement of ElasticNet.
2. GraphNet does not provide a breakthrough in term of performances and stability.

Figure 4.3: Left panel: coefficient maps from meshes of cortical thickness obtained by the three models: GM ElasticNet (Enet), GraphNet (GN), and Enet-TV (TV) on the three datasets: ADNI, NUSDAST, and Dep-TMS. Right: stability of the coefficient maps as the proportion $\in [0, 1]$ of selection across CV-folds for each vertex, i. e., the dark value of 1 means that the vertex is always selected.



3. TV-Enet is more sensitive to settle, however we propose a default settings $(\lambda_1, \lambda_2, \lambda_{TV}) = \alpha(0.1, 1, \kappa)$ with $\alpha = 0.1$ and $\kappa = 0.8$ that provides some increase of performance compare to ElasticNet and breakthrough in term of map stability and interpretability.

4.2 CONESTA: an efficient solver for structured sparsity in high-dimensionality

The previous section has demonstrated the potential of TV regularization. This section addresses the practical optimization problem incorporating structured penalties (TV) to deal with large scale "real life" datasets ($N \geq 100$ and $P \geq 300,000$ with both 3D image and meshes (of cortical surface). To our knowledge, there is still no solver to address these two constraints.

4.2.1 Reformulating TV as a linear operator

Before discussing the optimization strategy, we provide details on the encoding of the spatial structure within the TV penalty.

3D image

This section presents the formulation and the design of $\boldsymbol{A}$ in the specific case of the TV penalty applied to the parameter vector $\boldsymbol{w}$ measured on a 3-dimensional (3D) image.

A brain mask is used to establish a mapping, $\phi(i, j, k)$, from integer coordinates (i, j, k) in the 3D grid of the brain image, and an index $\phi \in \{1, \ldots, P\}$ in the collapsed (vectorized) image. We extract the spatial (forward) neighborhood at (i, j, k), of size ≤ 4, corresponding to a voxel and its three neighboring voxels, within the mask, in the positive i, j and k directions. The TV penalty is defined as

$$\mathrm{TV}(\boldsymbol{w}) \equiv \sum_{i,j,k} \left\| \nabla \left(\boldsymbol{w}_{\phi(i,j,k)} \right) \right\|_2, \tag{4.3}$$

where $\nabla(\boldsymbol{w}_{\phi(i,j,k)})$ denotes the spatial gradient of the parameter map, $\boldsymbol{w} \in \mathbb{R}^P$, at the 3D position (i, j, k) mapped to element $\phi(i, j, k)$ in $\boldsymbol{w}$. A first order approximation of the spatial gradient, $\nabla(\boldsymbol{w}_{\phi(i,j,k)})$, can be computed by applying the linear operator $\boldsymbol{A}'_\phi \in \mathbb{R}^{3\times 4}$ to the parameter vector $\boldsymbol{w}'_{\phi(i,j,k)} \in \mathbb{R}^4$ as

$$\nabla \left(\boldsymbol{w}_{\phi(i,j,k)} \right) \equiv \underbrace{\begin{bmatrix} -1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ -1 & 0 & 0 & 1 \end{bmatrix}}_{\boldsymbol{A}'_\phi} \underbrace{\begin{bmatrix} \boldsymbol{w}_{\phi(i,j,k)} \\ \boldsymbol{w}_{\phi(i+1,j,k)} \\ \boldsymbol{w}_{\phi(i,j+1,k)} \\ \boldsymbol{w}_{\phi(i,j,k+1)} \end{bmatrix}}_{\boldsymbol{w}'_{\phi(i,j,k)}}, \tag{4.4}$$

where $\boldsymbol{w}'_{\phi(i,j,k)}$ contains the elements at linear indices in the collapsed parameter map, $\boldsymbol{w}$, corresponding to the spatial neighborhood in the 3D image at $\boldsymbol{w}_{\phi(i,j,k)}$. Then, $\boldsymbol{A}'_\phi$ is extended, by zeros, to a large but very sparse matrix $\boldsymbol{A}_{\phi(i,j,k)} \in \mathbb{R}^{3\times P}$ such that $\boldsymbol{A}'_\phi \boldsymbol{w}'_{\phi(i,j,k)} = \boldsymbol{A}_{\phi(i,j,k)} \boldsymbol{w}$. If some neighbors lie outside of the mask, the corresponding rows in $\boldsymbol{A}_{\phi(i,j,k)}$ are removed (or set to zero). Approximating TV by

$$\mathrm{TV}(\boldsymbol{w}) = \sum_{i,j,k} \|\boldsymbol{A}_{\phi(i,j,k)} \boldsymbol{w}\|_2 \tag{4.5}$$

allows us to use the TV, as the structured penalty s, in Eq. 4.2. Finally, with a vertical concatenation of all the $\boldsymbol{A}_{\phi(i,j,k)}$ matrices, we obtain the full linear operator $\boldsymbol{A} \in \mathbb{R}^{3P\times P}$ that will be used in the following sections.

Mesh of cortical surface

The linear operator $\boldsymbol{A}'_{\phi(v)}$ used to compute a first order approximation of the spatial gradient can be obtained by examining the neighboring vertices of each vertex v. Where $\phi(v)$ establishes a mapping, from vertex v in the mesh to an index $\phi \in \{1, \ldots, P\}$ in the collapsed mesh. With common triangle-tessellated surfaces, the neighborhood size is ≤ 7 (including $\phi(v)$). In this setting, we have $\boldsymbol{A}'_{\phi(v)} \in \mathbb{R}^{3\times7}$, which can be extended and concatenated to obtain the full linear operator $\boldsymbol{A}$.

4.2.2 The TV-Enet problem

We propose a solver that addresses a very general class of optimization problems including many group-wise penalties (allowing overlapping groups) such as Group Lasso and TV. The function that we wish to minimize has the form

$$\min_{\boldsymbol{w}} f(\boldsymbol{w}) = \underbrace{\overbrace{L_\varepsilon(\boldsymbol{w}) + \lambda_2\|\boldsymbol{w}\|_2^2}^{\text{smooth}}}_{g(\boldsymbol{w})} + \overbrace{\lambda_1 \underbrace{\|\boldsymbol{w}\|_1}_{h(\boldsymbol{w})} + \lambda_s \underbrace{\sum_{g\in\mathcal{G}} \|\boldsymbol{A}_g\boldsymbol{w}\|_2}_{s(\boldsymbol{w})}}^{\text{non-smooth}}, \quad (4.6)$$

where $g(\boldsymbol{w})$ is the penalized smooth (i. e.differentiable) loss, $h(\boldsymbol{w})$ is a sparsity-inducing penalty whose proximal operator is known and $s(\boldsymbol{w})$ is a complex penalty (e.g., Group Lasso and TV) on the structure of the input variables with an unknown proximal operator. h (ℓ_1) can be minimized with proximal gradient (Beck and Teboulle, 2009a) method. The difficulty is with s that is neither smooth nor has a known proximal operator.

4.2.3 Background

Although many solvers have already been proposed to minimize such function, their practical use in the context of high-dimensional neuroimaging data ($\geq 10^5$ features) remains an open issue (see Sec. Background Supplementary in (Hadj-Selem et al., 2018) and (Dohmatob et al., 2014; Varoquaux et al., 2015)). Next, we provide a short overview of the existing solvers' limitations:

1. *Primal-dual* proposed by Chambolle and Pock, 2011 (application to fMRI: Gramfort, Thirion, and Varoquaux, 2013) assumes to have access to the proximal operators of both the smooth part and the non-smooth part of the minimized function. This method requires the approximation (using inexact method described below) of proximal operators when they are not available. This is a major shortcoming for an efficient application to logistic regression.

2. *Inexact FISTA*: Inexact proximal gradient algorithm (Schmidt, Le Roux, and Bach, 2011), where the proximal operator of s is approximated numerically. While the main algorithm (FISTA) enjoys a convergence rate of $\mathcal{O}(1/k^2)$ (with k the number of iterations), the precision of the approximation of the proximal operator is required to decrease as $\mathcal{O}(1/k^{4+\delta})$ for any $\delta > 0$ (Schmidt, Le Roux, and Bach, 2011, Proposition 2). This results in prohibitive computations to reach a reasonable precision, especially with high-dimensional $\boldsymbol{w}$ vectors, as found with brain images that generally involve $\geq 10^4$ (functional MRI) to $\geq 10^5$ (structural MRI) features. This solver has been applied to fMRI by Michel et al., 2011.

3. *ADMM* The alternating direction method of multipliers (Boyd et al., 2011) is computationally expensive and/or difficult to compute for general structured penalties and which suffers from a difficulty of setting the regularization parameter, as mentioned in Dohmatob et al., 2014; Schmidt, Le Roux, and Bach, 2011.

4. *EGM* (Excessive Gap Method) (Nesterov, 2005b) does not allow true sparsity nor general loss functions.

5. Nesterov's smoothing technique (Nesterov, 2005a) provides an appealing and generic framework in which a broad range of non-smooth convex structured penalties can be minimized without computing their proximal operators. However, reasonable precision ($\approx 10^{-3}$ or higher) requires a very small smoothing parameter, which brings down the convergence rate to unacceptable levels.

We propose a continuation solver, called *CONESTA* (short for **CO**ntinuation with **NE**sterov smoothing in a **S**hrinkage-**T**hresholding **A**lgorithm), based on Nesterov's smoothing technique that automatically generates a decreasing sequence of smoothing parameters to maintain the optimal convergence speed towards any globally desired precision. Nesterov's smoothing technique makes the solver highly versatile: it can address a large class of complex penalties (the function s in Eq. 4.6) where the proximal operators are either not known or expensive to compute. The problem can be minimized using an accelerated proximal gradient method, possibly also utilizing (non-smoothed, e. g.ℓ_1) sparsity-inducing penalties. CONESTA can be understood as a smooth touchdown procedure that uses the duality gap to probe the distance to the ground (global optimum) and dynamically adapts its speed (the smoothing parameter) according to this distance.

4.2.4 Nesterov's smoothing of the structured penalty

We consider the convex non-smooth minimization of Eq. 4.6 with respect to $\boldsymbol{w}$. This problem includes a general structured penalty, s, that (for the purpose of this paper) covers the specific case of TV. The accelerated proximal gradient algorithm (FISTA, Beck and Teboulle, 2009a) can be used to solve the problem when applying only e. g.the ℓ_1 penalty. However, since the proximal operator of TV, together with the ℓ_1 penalty, has no closed-form expression, standard implementations of those algorithms are not suitable. In order to overcome this barrier we used Nesterov's smoothing technique (Nesterov, 2005a), which consists of approximating the non-smooth penalty for which the proximal operator is unknown (e. g., TV) with a smooth function (for which the gradient is known). Non-smooth penalties with known proximal operators (e. g., ℓ_1) are not affected by this smoothing. Hence, as described in Chen et al., 2012, this allowed us to use an exact accelerated proximal gradient algorithm.

Using the dual norm of the ℓ_2-norm (i. e.the ℓ_2-norm), Eq. 4.5 can be reformulated as

$$\begin{aligned}\mathrm{TV}(\boldsymbol{w}) &= \sum_{i,j,k} \|\boldsymbol{A}_{\phi(i,j,k)}\boldsymbol{w}\|_2 \\ &= \sum_{i,j,k} \max_{\|\boldsymbol{\alpha}_{\phi(i,j,k)}\|_2 \leq 1} \boldsymbol{\alpha}_{\phi(i,j,k)}^{\top} \boldsymbol{A}_{\phi(i,j,k)}\boldsymbol{w},\end{aligned} \tag{4.7}$$

where $\boldsymbol{\alpha}_{\phi(i,j,k)} \in \mathcal{K}_{\phi(i,j,k)} = \{\boldsymbol{\alpha}_{\phi(i,j,k)} \in \mathbb{R}^3 : \|\boldsymbol{\alpha}_{\phi(i,j,k)}\|_2 \leq 1\}$ is a vector of auxiliary variables in the ℓ_2 unit ball, associated with $\boldsymbol{A}_{\phi(i,j,k)}\boldsymbol{w}$. As with $\boldsymbol{A} \in \mathbb{R}^{3P \times P}$, which is the vertical concatenation of all the $\boldsymbol{A}_{\phi(i,j,k)}$, we concatenate all the $\boldsymbol{\alpha}_{\phi(i,j,k)}$ to form $\boldsymbol{\alpha} \in \mathcal{K} = \{[\boldsymbol{\alpha}_1^T, \ldots, \boldsymbol{\alpha}_P^T]^T : \boldsymbol{\alpha}_l \in \mathcal{K}_l, \forall\, l = \phi(i,j,k) \in \{1, \ldots, P\}\} \in \mathbb{R}^{3P}$. The set $\mathcal{K}$ is the Cartesian product of closed 3D unit balls in Euclidean space and, therefore, a compact convex set. Eq. 4.7 can now further be written as

$$TV(\boldsymbol{w}) = \max_{\boldsymbol{\alpha} \in \mathcal{K}} \boldsymbol{\alpha}^T \boldsymbol{A}\boldsymbol{w} = s(\boldsymbol{w}), \tag{4.8}$$

and with this formulation of s, we can apply Nesterov's smoothing technique. For a given smoothing parameter, $\mu > 0$, the function s is approximated by the smooth function

$$s_\mu(\boldsymbol{w}) = \max_{\boldsymbol{\alpha} \in \mathcal{K}} \left\{ \boldsymbol{\alpha}^T \boldsymbol{A}\boldsymbol{w} - \frac{\mu}{2}\|\boldsymbol{\alpha}\|_2^2 \right\}, \tag{4.9}$$

for which $\lim_{\mu \to 0} s_\mu(\boldsymbol{w}) = s(\boldsymbol{w})$. Nesterov (Nesterov, 2005a) demonstrates this convergence using the inequality in Eq. 4.13. The value of $\boldsymbol{\alpha}_\mu^*(\boldsymbol{w}) = [\boldsymbol{\alpha}_{\mu,1}^{*T}, \ldots, \boldsymbol{\alpha}_{\mu,\phi(i,j,k)}^{*T}, \ldots, \boldsymbol{\alpha}_{\mu,P}^{*T}]^T$ that maximizes Eq. 4.9 is the concatenation of projections of the vectors $\boldsymbol{A}_{\phi(i,j,k)}\boldsymbol{w} \in \mathbb{R}^3$ onto the ℓ_2 ball $\mathcal{K}_{\phi(i,j,k)}$, i. e.$\boldsymbol{\alpha}^*_{\mu,\phi(i,j,k)}(\boldsymbol{w}) = \text{proj}_{\mathcal{K}_{\phi(i,j,k)}}\left(\frac{\boldsymbol{A}_{\phi(i,j,k)}\boldsymbol{w}}{\mu}\right)$, where

$$\text{proj}_{\mathcal{K}_{\phi(i,j,k)}}(\boldsymbol{x}) = \begin{cases} \boldsymbol{x} & \text{if } \|\boldsymbol{x}\|_2 \leq 1 \\ \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_2} & \text{otherwise.} \end{cases} \tag{4.10}$$

The function s_μ, i. e.Nesterov's smooth transform of s, is convex and differentiable. Its gradient is given by Nesterov (Nesterov, 2005a) as

$$\nabla s_\mu(\boldsymbol{w}) = \boldsymbol{A}^T\boldsymbol{\alpha}^*_\mu(\boldsymbol{w}). \tag{4.11}$$

The gradient is Lipschitz-continuous, with constant

$$L(\nabla(s_\mu)) = \frac{\|\boldsymbol{A}\|_2^2}{\mu}, \tag{4.12}$$

in which $\|\boldsymbol{A}\|_2$ is the matrix spectral norm of $\boldsymbol{A}$. Moreover, Nesterov (Nesterov, 2005a) provides the following inequality, relating s_μ and s

$$s_\mu(\boldsymbol{w}) \leq s(\boldsymbol{w}) \leq s_\mu(\boldsymbol{w}) + \mu M, \quad \forall \boldsymbol{w} \in \mathbb{R}^P, \tag{4.13}$$

where $M = \max_{\boldsymbol{\alpha}\in\mathcal{K}} \frac{1}{2}\|\boldsymbol{\alpha}\|_2^2 = \frac{P}{2}$.

Thus, a new (smoothed) function, closely related to Eq. 4.6, arises as

$$f_\mu(\boldsymbol{w}) = \overbrace{\underbrace{L_\varepsilon(\boldsymbol{w}) + \lambda_2\|\boldsymbol{w}\|_2^2}_{g(\boldsymbol{w})} + \lambda_s \underbrace{\left\{\boldsymbol{\alpha}^*_\mu(\boldsymbol{w})^T\boldsymbol{A}\boldsymbol{w} - \frac{\mu}{2}\|\boldsymbol{\alpha}^*\|_2^2\right\}}_{s_\mu(\boldsymbol{w})}}^{\text{smooth}} + \lambda_1 \overbrace{\underbrace{\|\boldsymbol{w}\|_1}_{h(\boldsymbol{w})}}^{\text{non-smooth}}. \tag{4.14}$$

Hence, we can explicitly compute the gradient of the smooth part, $\nabla(g + \lambda_s s_\mu)$ using Eq. 4.11, its Lipschitz constant L (using Eq. 4.12) and also the proximal operator of the non-smooth part.

Linear regression loss

$$\begin{aligned} \nabla\left(g + \lambda_s s_\mu\right) &= \nabla(g) + \lambda_s\nabla(s_\mu) \\ &= \boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{w}^k - y) + \lambda_s\boldsymbol{A}^\top\boldsymbol{\alpha}^*_\mu(\boldsymbol{w}^k), \end{aligned} \tag{4.15}$$

$$L\left(\nabla\left(g + \lambda_s s_\mu\right)\right) = 2 + \lambda_s\frac{\|\boldsymbol{A}\|_2^2}{\mu}. \tag{4.16}$$

Logistic regression loss

$$\begin{aligned}\nabla\left(g+\lambda_s s_\mu\right) &= \nabla(g)+\lambda_s\nabla(s_\mu)\\ &= \boldsymbol{X}^T(y-\frac{1}{1+e^{-\boldsymbol{X}\boldsymbol{w}^k}})+\lambda_s\boldsymbol{A}^\top\boldsymbol{\alpha}^*_\mu(\boldsymbol{w}^k),\end{aligned} \tag{4.17}$$

$$L\left(\nabla\left(g+\lambda_s s_\mu\right)\right)=1/2\|\boldsymbol{X}\|_2^2+\lambda_s\frac{\|\boldsymbol{A}\|_2^2}{\mu}. \tag{4.18}$$

We thus have all the necessary ingredients to minimize the function using e. g.an accelerated proximal gradient method (Beck and Teboulle, 2009a). Given a starting point, $\boldsymbol{w}^0$, and a smoothing parameter, μ, FISTA (Algorithm 2) minimizes the smoothed function and reaches a given precision, ε_μ.

```
Algorithm 2 FISTA($\boldsymbol{w}^0, \varepsilon_\mu, \mu, \boldsymbol{A}, g, s_\mu, h, \lambda_s, \lambda_1$)
1: $\boldsymbol{w}^1 = \boldsymbol{w}^0; k = 2$
2: Step size $t_\mu = \left(L(\nabla(g)) + \lambda_s\frac{\|\boldsymbol{A}\|_2^2}{\mu}\right)^{-1}$
3: repeat
4:     $\mathbf{z} = \boldsymbol{w}^{k-1} + \frac{k-2}{k+1}\left(\boldsymbol{w}^{k-1} - \boldsymbol{w}^{k-2}\right)$
5:     $\boldsymbol{w}^k = \text{prox}_{\lambda_1 h}(\mathbf{z} - t_\mu\nabla(g+\lambda_s s_\mu)(\mathbf{z}))$
6: until $Gap_\mu(\boldsymbol{w}^k) \leq \varepsilon_\mu$ (see Sec. 4.2.6)
7: return $\boldsymbol{w}^k$
```

4.2.5 Principles of the CONESTA algorithm

The step size, t_μ, computed in Line 2 of Algorithm 2, must be smaller than or equal to the reciprocal of the Lipschitz constant of the gradient of the smooth part, i. e.of $g+\lambda_s s_\mu$ (Beck and Teboulle, 2009a). This relationship between t_μ and μ implies a trade-off between speed and precision: Indeed, the FISTA convergence rate, given in the Supplementary (Eq. SM 2.3), shows that a high precision (small μ and t_μ) will lead to slow convergence. Conversely, poor precision (large μ and t_μ) will lead to rapid convergence.

To optimize this trade-off, we propose a continuation approach (Algorithm 3) that decreases the smoothing parameter for the distance to the minimum. On the one hand, when we are far from $\boldsymbol{w}^*$ (the minimum of Eq. 4.6), we can use a large μ to decrease the objective function rapidly. On the other hand, when we are close to $\boldsymbol{w}^*$, we need a small μ to obtain an accurate approximation of the original objective function.

4.2.6 Duality gap

The distance to the unknown $f(\boldsymbol{w}^*)$ is estimated using a duality gap. Duality formulations are often used to control the achieved precision level when minimizing convex functions. The duality gap provides an upper bound of the error, $f(\boldsymbol{w}^k) - f(\boldsymbol{w}^*)$, for any $\boldsymbol{w}^k$, when the minimum is unknown. Moreover, it vanishes at the minimum:

$$\begin{aligned} \textsc{Gap}(\boldsymbol{w}^k) \geq f(\boldsymbol{w}^k) - f(\boldsymbol{w}^*) &\geq 0, \\ \textsc{Gap}(\boldsymbol{w}^*) &= 0. \end{aligned} \tag{4.19}$$

The duality gap is the cornerstone of the CONESTA algorithm. Indeed, it is used three times:

(i) As the stopping criterion in the inner FISTA loop (Line 6 in Algorithm 2). FISTA stops as soon as the current precision is achieved using the current smoothing parameter, μ, which prevents unnecessary iterations toward the approximated (smoothed) objective function.

(ii) In the ith CONESTA iteration, as a way to estimate the current error $f(\boldsymbol{w}^i) - f(\boldsymbol{w}^*)$ (Line 7 in Algorithm 3). The error is estimated using the gap of the smoothed problem, $\textsc{Gap}_{\mu=\mu^i}(\boldsymbol{w}^{i+1})$, which avoids unnecessary computation since it has already been computed during the last iteration of FISTA. The inequality in Eq. 4.13 is used to obtain the distance, ε^i, to the original non-smoothed problem. The next desired precision, ε^{i+1}, and the smoothing parameter, μ^{i+1} are derived from this value.

(iii) Finally, as the global stopping criterion in CONESTA (Line 10 in Algorithm 3). This guarantees that the obtained approximation of the minimum, $\boldsymbol{w}^i$, at convergence, satisfies $f(\boldsymbol{w}^i) - f(\boldsymbol{w}^*) < \varepsilon$.

Eq. 4.14 decomposes the smoothed objective function as a sum of a strongly convex loss, L_ε, and the penalties. Therefore, we can equivalently express the smoothed objective function as

$$\begin{aligned} f_\mu(\boldsymbol{w}) &= L_\varepsilon(\boldsymbol{w}) + \Omega_\mu(\boldsymbol{w}) \\ &= l(\boldsymbol{X}\boldsymbol{w}) + \Omega_\mu(\boldsymbol{w}), \end{aligned}$$

where Ω_μ represents all penalty terms of Eq. 4.14. We aim to compute the duality gap to obtain an upper bound estimation of the distance to the optimum. At any step k of the algorithm, given the current primal $\boldsymbol{w}^k$ and the dual $\boldsymbol{\sigma}(\boldsymbol{w}^k) \equiv \nabla L_\varepsilon(\boldsymbol{X}\boldsymbol{w}^k)$ variables (Borwein and Lewis, 2006), we can compute the duality gap using the Fenchel duality rules (**Mairal2010**). It requires computing the Fenchel conjugates, l^* and Ω_μ^*, of l and Ω_μ, respectively. While the expression of l^* is straightforward, to the best of our knowledge, there is no explicit expression for Ω_{μ}^{*} when using a complex penalty such as TV or group Lasso. Therefore, as a significant theoretical contribution of this paper, we provide the expression for Ω_{μ}^{*} to compute an approximation of the duality gap that maintains its properties (Eq. 4.19).

Theorem 1 (Duality gap for the smooth problem)**.** *The following estimation of the duality gap satisfies Eq. 4.19 , for any iterate* $\boldsymbol{w}^k$*:*

$$\textsc{Gap}_{\mu}(\boldsymbol{w}^k) \equiv f_{\mu}(\boldsymbol{w}^k) + l^{*}(\boldsymbol{\sigma}(\boldsymbol{w}^k)) + \Omega_{\mu,k}^{*}(-\boldsymbol{X}^T\boldsymbol{\sigma}(\boldsymbol{w}^k)), \qquad (4.20)$$

The proof of this theorem can be found in Supplementary of Hadj-Selem et al., 2018 (Sec. SM 3.1.3). Note that the Supplementary provides the expression and proof of the Fenchel conjugate for the non-smoothed problem, i. e., using Ω instead of Ω_{μ}.

Linear regression loss $L_{\varepsilon}(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{X}\boldsymbol{w} - y\|_2^2$, can be re-written as a function of $\boldsymbol{X}\boldsymbol{w}$ by $l(z) \equiv \frac{1}{2}\|z - y\|^2$, where $z = \boldsymbol{X}\boldsymbol{w}$. the dual variable is :

$$\sigma(\boldsymbol{w}^k) \equiv \nabla l(\boldsymbol{X}\boldsymbol{w}^k) = \boldsymbol{X}\boldsymbol{w}^k - y, \qquad (4.21)$$

and the Fenchel conjugates:

$$\begin{aligned} l^{*}(z) =& \frac{1}{2}\|z\|_2^2 + \langle z, y\rangle \\ \Omega_{\mu,k}^{*}(z) \equiv& \frac{1}{2\lambda_2}\sum_{j=1}^{P}\left(\left[\left|z_j - \lambda_s\left(\boldsymbol{A}^T\boldsymbol{\alpha}_{\mu}^{*}(\boldsymbol{w}^k)\right)_j\right| - \lambda_1\right]_+^2\right) \\ &+ \frac{\lambda_s\mu}{2}\|\boldsymbol{\alpha}_{\mu}^{*}(\boldsymbol{w}^k)\|_2^2, \qquad (4.22) \end{aligned}$$

where $[\,\cdot\,]_+ = \max(0, \cdot\,)$.

Logistic regression loss The dual variable is:

$$\sigma(\boldsymbol{w}^k) \equiv \nabla l(\boldsymbol{X}\boldsymbol{w}^k) = \frac{1}{1 + e^{-\boldsymbol{X}\boldsymbol{w}^k}} - y \qquad (4.23)$$

and the Fenchel conjugates

$$l^{*}(z) = \sum_{j=1}^{P}\left(z_j \log(z_j) + (1 - z_j)\log(1 - z_j)\right) \qquad (4.24)$$

with $z = \frac{1}{1+e^{-\boldsymbol{X}\boldsymbol{w}^k}}$

The expression in Eq. 4.20 of the duality gap of the smooth problem combined with the inequality in Eq. 4.13 provides an estimation of the distance to the minimum of the original non-smoothed problem. The sought distance is decreased geometrically by a factor $\tau \in (0, 1)$ at the end of each continuation, and the decreased value defines the precision that should be reached by the next iteration (Line 8 of Algorithm 3). Thus, the algorithm dynamically generates a sequence of decreasing precisions, ε^i. Such a scheme ensures the convergence towards a globally desired final precision, ε, which is the only parameter that the user needs to provide.

4.2.7 Determining the optimal smoothing parameter

Given the current precision, ε^i, we need to compute a smoothing parameter $\mu_{opt}(\varepsilon^i)$ (Line 9 in Algorithm 3) that minimizes the number of FISTA iterations required to achieve such a precision when minimizing Eq. 4.2 via Eq. 4.6 (i.e., such that $f(\boldsymbol{w}^k) - f(\boldsymbol{w}^*) < \varepsilon^i$). We have the following theorem giving the expression of the optimal smoothing parameter, for which a proof is provided in the Supplementary of Hadj-Selem et al., 2018 (Sec. SM 3.2).

Theorem 2 (Optimal smoothing parameter, μ). *For any given $\varepsilon > 0$, selecting the smoothing parameter as*

$$\mu_{opt}(\varepsilon) = \frac{-\lambda_s M \|\boldsymbol{A}\|_2^2 + \sqrt{(\lambda_s M \|\boldsymbol{A}\|_2^2)^2 + ML(\nabla(g))\|\boldsymbol{A}\|_2^2 \varepsilon}}{ML(\nabla(g))}, \tag{4.25}$$

minimizes the worst case bound on the number of iterations required to achieve the precision $f(\boldsymbol{w}^k) - f(\boldsymbol{w}^*) < \varepsilon$.

Note that $M = P/2$ (Eq. 4.13) and the Lipschitz constant of the gradient of g as defined in Eq. 4.14 is $L(\nabla(g)) = \lambda_{\max}(\boldsymbol{X}^T\boldsymbol{X}) + \lambda$, where $\lambda_{\max}(\boldsymbol{X}^T\boldsymbol{X})$ is the largest eigenvalue of $\boldsymbol{X}^T\boldsymbol{X}$.

4.2.8 CONESTA algorithm and convergence analysis

The user only has to provide the globally prescribed precision ε, which is guaranteed by the duality gap. Other parameters are related to the problem to be minimized (i.e.g, λ_s, s, λ_1, h) and the encoding of the data structure A. Finally, the value of τ was set to 0.5. Indeed, experiments shown in Supplementary of Hadj-Selem et al., 2018 (Sec. SM 4.2) have demonstrated that values of 0.5 or 0.2 led to similar and increased speeds compared to larger values, such as 0.8.

CONESTA acts as a smooth touchdown procedure that uses the duality gap to probe the distance to the ground (global optimum) to dynamically adapt its speed (the smoothing). Indeed, each continuation step of CONESTA (Algorithm 3) probes (Line 7) an upper bound ε^i of the current distance to the optimum ($f(\boldsymbol{w}^i) - f(\boldsymbol{w}^*)$) using the duality gap. Then, Line 8 computes the next precision to be reached, ε^{i+1}, decreasing ε^i by a factor $\tau \in (0, 1)$. Line 9 derives the optimal smoothing parameter, μ^{i+1}, required to reach this precision

Algorithm 3 CONESTA$(\varepsilon, \boldsymbol{A}, g, s, h, \lambda_s, \lambda_1, \tau = 0.5)$

```
1:  Initialize $\boldsymbol{w}^0 \in \mathbb{R}^P$
2:  $\varepsilon^0 = \tau \cdot \textsc{Gap}_{\mu=10^{-8}}(\boldsymbol{w}^0)$
3:  $\mu^0 = \mu_{opt}\left(\varepsilon^0\right)$
4:  repeat
5:      $\varepsilon^i_\mu = \varepsilon^i - \mu^i \lambda_s M$
6:      $\boldsymbol{w}^{i+1} = \textsc{Fista}(\boldsymbol{w}^i, \varepsilon^i_\mu, \mu^i, \boldsymbol{A}, g, s_{\mu^i}, h, \lambda_s, \lambda_1)$
7:      $\varepsilon^i = \textsc{Gap}_{\mu=\mu^i}(\boldsymbol{w}^{i+1}) + \mu^i \lambda_s M$
8:      $\varepsilon^{i+1} = \tau \cdot \varepsilon^i$
9:      $\mu^{i+1} = \mu_{opt}\left(\varepsilon^{i+1}\right)$
10: until $\varepsilon^i \leq \varepsilon$
11: return $\boldsymbol{w}^{i+1}$
```

as fast as possible. Finally, Line 5 transforms back the precision to the original problem into a precision for the smoothed problem, ε^i_μ, using the inequality in Eq. 4.13. Therefore, at the next iteration, FISTA (Line 6) will decrease f^i_μ until the error reaches ε^i_μ. Thanks to Line 5, this implies that the true error (toward the non-smoothed problem) is smaller than ε^i. The resulting weight vector, $\boldsymbol{w}^{i+1}$, is the initial value for the next continuation step using updated parameters. Note that we use the duality gap for the smoothed problem, $\textsc{Gap}_{\mu=\mu^i}$ (and ε^i_μ), and transform it back and forth using Eq. 4.13 to obtain the duality gap for the non-smooth problem, $\textsc{Gap}$ (and ε^i). We do this because the gap on Line 7 has already been computed at the last iteration of the FISTA loop (Line 6), since it was used in the stopping criterion. Moreover, $\textsc{Gap}_\mu$ converges to zero for any fixed μ unlike $\textsc{Gap}$.

The initialization (Line 2) is a particular case where we use $\textsc{Gap}_\mu$ with a negligible smoothing value of e. g.$\mu = 10^{-8}$. We then derive the initial smoothing parameter on Line 3. Therefore, if we start close to the solution, the algorithm will automatically pick a small smoothing parameter, which makes CONESTA an excellent candidate for warm-restart.

The following theorem ensures the convergence and convergence speed of CONESTA.

Theorem 3 (Convergence of CONESTA)**.** *Let $\left(\mu^i\right)_{i=0}^{\infty}$ and $\left(\varepsilon^i\right)_{i=0}^{\infty}$ be defined recursively by CONESTA (Algorithm 3). Then, we have that*

(i) $\lim_{i\to\infty} \varepsilon^i = 0,$ *and*

(ii) $f(\boldsymbol{w}^i) \xrightarrow{i\to\infty} f(\boldsymbol{w}^*)$.

(iii) Convergence rate of CONESTA with fixed smoothing (without continuation): For any given desired precision $\varepsilon > 0$, using a fixed smoothing

(line 6 of Algorithm 3) with an optimal value of μ, equal to $\mu_{opt}(\varepsilon)$, if the number of iterations k is larger than

$$\frac{\sqrt{8\|\boldsymbol{A}\|_2^2 M\lambda_s^2\|\boldsymbol{w}^0-\boldsymbol{w}^*\|_2^2}}{\varepsilon}+\frac{\sqrt{2L(\nabla(g))\|\boldsymbol{w}^0-\boldsymbol{w}^*\|_2^2}}{\sqrt{\varepsilon}}.$$

then the obtained $\boldsymbol{w}^k$ satisfies $f(\boldsymbol{w}^k)-f(\boldsymbol{w}^*)<\varepsilon$.

(iv) Convergence rate of CONESTA (with continuation), assuming uniqueness of the minimum (β^*): For any given desired precision $\varepsilon>0$, if the total sum of all the inner FISTA iterations is larger than

$$C/\varepsilon,$$

where $C>0$ is a constant, then the obtained solution (obtained from (iii)), i. e.$\boldsymbol{w}^i$, satisfies $f(\boldsymbol{w}^i)-f(\boldsymbol{w}^*)<\varepsilon$.

The proof is provided Sec. SM 3.3 in the Supplementary of Hadj-Selem et al., 2018. Claim (i): Sec. SM 3.3.1 demonstrates that the sequence of decreased precisions ε^i converges toward any prescribed precision. Claim (ii): Sec. SM 3.3.2 demonstrates, at each step of the sequence, the solutions of the smoothed problem converge toward the solution of the non-smoothed problem. Claim (iii): Sec. SM 3.3.3 demonstrates the number of iterations required to converge toward $\boldsymbol{w}^*$ using the auxiliary smoothed problem (without continuation) with a fixed and optimal smoothing value. Finally, claim (iv), Sec. SM 3.3.4 provides the convergence rate concerning the total number of iterations.

The continuation technique improves the convergence rate compared to the simple smoothing using a single value of μ. Indeed, it has been demonstrated in Beck and Teboulle, 2012 (see also Chen et al., 2012) that the convergence rate obtained with a single value of μ, even optimized, is $\mathcal{O}(1/\varepsilon)+\mathcal{O}(1/\sqrt{\varepsilon})$. However, the CONESTA algorithm achieves $\mathcal{O}(1/\varepsilon)$ for simply (non-strongly) convex functions.

4.2.9 Benchmarking solver convergence speed

In this section we compare CONESTA to the state-of-the-art algorithms mentioned above (see Supplementary of Hadj-Selem et al., 2018 for details), i. e., ADMM, EGM, Inexact FISTA and FISTA with fixed μ. We will use these algorithms to solve the problem on both simulated and high-dimensional structural neuroimaging data.

We used FISTA with fixed μ using two values of μ, chosen as follows: (i) Chen's μ where $\mu=\varepsilon/(2\lambda_s M)$ as was used in (Chen et al., 2012) and (ii) Large $\mu=(\text{Chen's }\mu)^{1/2}$. The first proposal for μ ensures that we reach the desired precision, although convergence may be slow. The second proposal has a value of μ that may not guarantee we reach the desired precision before convergence.

Benchmarking solvers on simulated data

Based on our contribution Löfstedt et al., 2018 we generated simulated data where we control the true minimizer, $\boldsymbol{w}^*$, and the associated regularization parameters λ_1, λ_2 and λ_s. The experimental setup for the simulated 1D data set was inspired by that of (Bach et al., 2011), including several sizes of the dataset, (N, P), the correlation between variables, the sparsity of the true $\boldsymbol{w}^*$, and the signal-to-noise ratio. For each dataset setting and each solver, we measured the number of iterations and the time (in seconds) required to reach a certain precision level, ε. For each precision level, ranging from 1 to 10^{-6}, for each solver and each dataset, we ranked solver according to the time they required to reach a given precision. Then, we averaged the ranks across data sets (see Tab. 4.1), and we tested (Friedman test (Friedman, 1940)) the significance of the difference in ranks between solvers.

Table 4.1: Average rank of the convergence speed of the algorithms to reach precisions $(f(\boldsymbol{w}^k) - f(\boldsymbol{w}^*))$ ranging from 1 to 10^{-6}. We have here reported whether the average rank of a given algorithm was significantly larger $>$ (slower) or significantly smaller $<$ (faster) than CONESTA (a missing '$>$' or '$<$' means non-significant). P-values were calculated with a post hoc analysis of the Friedman test corrected for multiple comparisons. Note that all reported significant differences had a corrected p-value of 10^{-3} or smaller. For a given dataset, all solver were evaluated with limited upper execution time. Thus, some high precisions (e. g., $10^{-5}, 10^{-6}$) were not always reached within a limited time. For FISTA with a large μ, the high precisions may, in fact, not be reachable at all. In those situations, the execution time was set to $+\infty$.

	Average rank of the time to reach a given precision						
Algorithm	1	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-6}
CONESTA	3.3 –	2.7 –	2.2 –	**1.6** –	**1.3** –	**1.0** –	**1.0** –
ADMM	2.9	**2.1**	**1.9**	1.8	1.7>	1.8>	1.5 >
EGM	**1.8**<	2.2	2.0	2.3>	2.3>	2.2>	1.7 >
FISTA large μ	4.7>	6.0>	4.6>	3.4>	2.7>	2.2>	1.7 >
FISTA Chen's μ	6.2>	5.0>	4.4>	3.4>	2.7>	2.2>	1.7 >
Inexact FISTA	6.2>	5.4>	4.3>	3.3>	2.7>	2.2>	1.7 >

Tab. 4.1 indicates that for precision higher (smaller) than 10^{-3}), CONESTA outperformed all other solvers. Its superiority was significant for all precisions, except for the comparison with ADMM at $\varepsilon = 10^{-3}$.

Benchmarking on structural MRI

We applied the solvers on a structural MRI data set of 199 subjects from the Alzheimer's Disease Neuroimaging Initiative (ADNI, http://adni.loni.usc.edu/) cohort. We included 119 controls and 80 patients with mild cognitive impairment (MCI) that converted to AD within 800 days, see Supplementary of Hadj-Selem et al., 2018 for details. Our goal was to predict (regression problem) the subjects' continuous ADAS (AD Assessment Scale-Cognitive Subscale) score ($\boldsymbol{y}$), measured 800 days after brain image acquisition. As input ($\boldsymbol{X}$), we retained 286 214 voxels of gray matters extracted with SPM8 using DARTEL normalization (Ashburner and Friston, 2005). We compared EGM, FISTA with the two different fixed μ and Inexact FISTA. Note that ADMM was excluded in this example. The version in (Wahlberg et al., 2012) is designed for 1D data was not adapted to 2D or 3D data. We fixed the desired precision, $\epsilon = 10^{-6}$ used as the stopping criterion. This precision was also used to derive the smoothing parameter for FISTA with fixed μ. We ran CONESTA and Inexact FISTA until they reached a precision of 10^{-7}, evaluated with the duality gap. The smallest value of $f(\boldsymbol{w}^k)$ was considered as the global minimum $f(\boldsymbol{w}^*)$ used to compute the errors $f(\boldsymbol{w}^k) - f(\boldsymbol{w}^*)$ in Tab. 4.4 and Tab. 4.2.

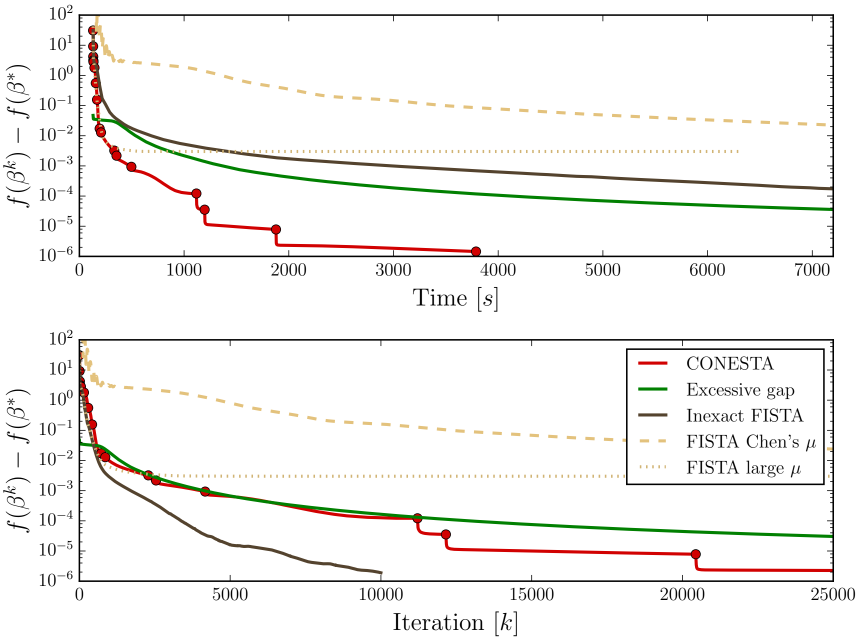


Figure 4.4: The error as a function of the computational time (top plot) and the number of iterations (bottom plot). The vertical axis is given in a logarithmic scale. Dots on the CONESTA curve indicate where the continuation steps take place, i. e.where the dynamic selection of a new smoothing parameter happened.

Fig. 4.4 and Tab. 4.2 shows that FISTA with fixed μ is either too slow (Chen's μ) or, as expected, does not reach the desired preci-

Table 4.2: Execution time ratios of each state-of-the-art algorithm over the time required by CONESTA to reach the same precision.

	Time ratio over CONESTA to reach a given precision						
Algorithm	10	1	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}
EGM	0.97	0.39	0.24	1.05	1.08	1.74	3.64
Inexact FISTA	1.13	1.92	2.64	3.97	4.38	5.45	6.07
FISTA Chen's μ	2.7	10.65	17.5	29.95	16.72	13.47	10.45
FISTA large μ	1.04	0.97	1.12	1.14	—	—	—

sion (as with the large μ). However, CONESTA competes with FISTA with large μ and EGM during the first iterations. This demonstrates two points: CONESTA dynamically picked an efficient (large enough) smoothing parameter and that the gap stopping criterion, used in the nested FISTA loop, allows stopping before reaching a plateau, and thus to quickly change to a smaller μ (illustrated with dots in Fig. 4.4 where the continuation steps occurred).

Fig. 4.4, bottom panel, shows the fast convergence of Inexact FISTA as a function of the number of iterations. However, the top panel of Fig. 4.4 shows that it is always considerably slower, in terms of the execution time, compared to the EGM or CONESTA. Tab. 4.2 reveals that Inexact FISTA is 4.38 times slower than CONESTA to reach an error of 10^{-3}, and this difference in speed increases with higher precisions. This demonstrates the hypothesis we stated in the introduction, that Inexact FISTA becomes slower after many iterations due to the necessity to decrease the precision faster than $1/k^4$ (k being the number of FISTA iterations), in the approximation.

As a conclusion, Fig. 4.4 and Tab. 4.2 demonstrate that on high-dimensional MRI data sets, CONESTA outperformed all other algorithms for precisions higher than $\varepsilon \leq 10^{-2}$.

Required precision and its gap estimate

The figure Fig. 4.5, top panel, shows that the similarity of coefficient maps to the true solution is reaching a plateau for precisions higher than 10^{-3}, using both the duality gap estimation and true precision.

An early stopping at $\varepsilon = 10^{-2}$ would provide a different solution than the expected one using either measures of precision: $\text{corr}(\boldsymbol{w}^k, \boldsymbol{w}^*) = 0.92$ with the gap estimate and $\text{corr}(\boldsymbol{w}^k, \boldsymbol{w}^*) = 0.45$ with the true precision. Moreover, the figure demonstrates the relevance of the duality gap as a stopping criterion: stopping the convergence at 10^{-3}, using the duality gap, provides a map with a 0.97 correlation with the true solution. The bottom panel shows that less than 10^4 iterations are sufficient to reach the target precision of 10^{-3}, which is less than

30 minutes of computation. It also shows that, for useful precisions ($\varepsilon \leq 10^{-1}$), the duality gap is an accurate upper-bound of the true error.

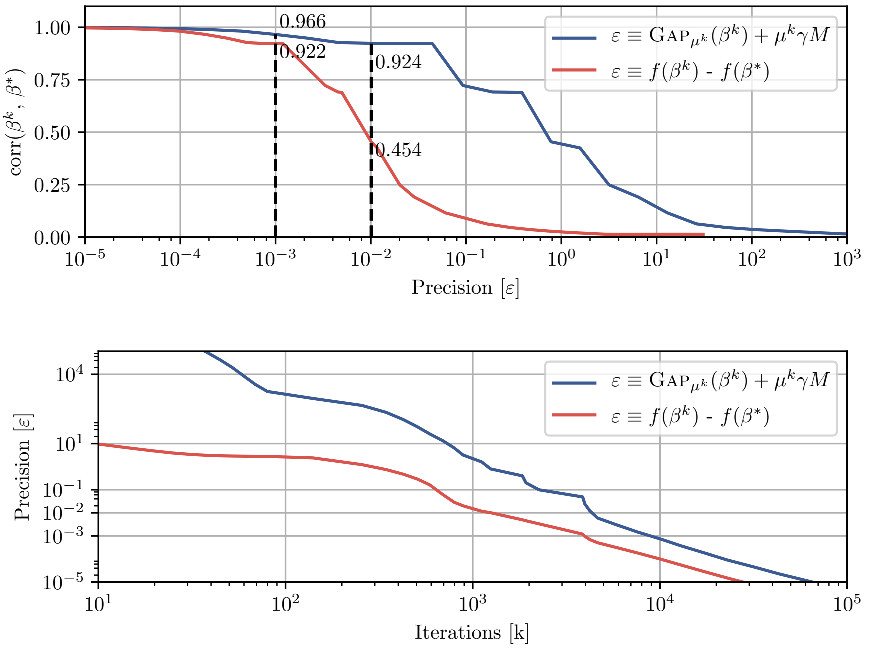


Figure 4.5: Top panel: Correlation between the coefficient maps $\boldsymbol{w}^k$ and the true solution $\boldsymbol{w}^*$ as a function of the true precision (red line) and precision estimated with the duality gap. The true solution has been estimated by running 10^6 iterations of CONESTA and Inexact FISTA. Bottom panel: True precision (red line) and precision estimated with the duality gap (blue line) as a function of the number of iteration. 10^4 iterations are sufficient to reach and outperform the required precision of 10^{-3}.

4.2.10 Conclusion

CONESTA minimizes any combination of the ℓ_1, ℓ_2 and TV penalties while preserving the exact ℓ_1 penalty. This solver uses Nesterov's technique to smooth the TV penalty such that objective function is minimized with an exact accelerated proximal gradient algorithm. The approximation of TV is controlled by a single smoothing parameter μ. This continuation algorithm uses successively smaller values of μ to reach a prescribed precision while achieving the best possible convergence rate.

Overall, the use of structured sparse supervised machine learning is highly relevant in providing a major breakthrough in terms of support recovery of the predictive brain regions. We will demonstrate the performance, interpretability and versatility of TV-Enet on two datasets of schizophrenia patients containing sMRI and fMRI, respectively in Chapters 5 and 6. In addition, we will see in Chapter 4 that the existence of structured and sparse regularization terms is not limited to supervised machine learning tools. Indeed, for some specific unsupervised machine learning analysis, the use of sparse and spatial constraint is also of great interest.

4.3 CONESTA applied to spatially regularized principal components analysis

The versatility of CONESTA has made possible the integration of structured sparsity in the popular unsupervised principal component analysis (PCA) applied to high-dimensional data.

4.3.1 Introduction

Principal components analysis (PCA) is an unsupervised statistical procedure whose aim is to capture dominant patterns of variability to provide an optimal representation of a data set in a lower-dimensional space defined by the principal components (PCs). Given a data set $\boldsymbol{X} \in \mathbb{R}^{N \times P}$ of N samples and P centered variables, PCA aims to find the most accurate rank-K approximation of the data:

$$\min_{\boldsymbol{U},\boldsymbol{D},\boldsymbol{V}} \left\| \boldsymbol{X} - \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T \right\|_F^2, \quad \text{s.t. } \boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}, \boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I}, d_1 \geq \cdots \geq d_K > 0 \tag{4.26}$$

where $\|.\|_F$ is the Frobenius norm of a matrix, $\boldsymbol{V} = [\boldsymbol{v}_1, \cdots, \boldsymbol{v}_K] \in \mathbb{R}^{P \times K}$ are the K loading vectors (right singular vectors) that define the new coordinate system where the original features are uncorrelated, $\boldsymbol{D}$ is the diagonal matrix of the K singular values, and $\boldsymbol{U} = [\boldsymbol{u}_1, \cdots, \boldsymbol{u}_K] \in \mathbb{R}^{N \times K}$ are the K projections of the original samples in the new coordinate system (called principal components (PCs) or left singular vector).

In a neuroimaging context, the goal is to discover the phenotypic markers accounting for the main variability in a population's brain images. For example, when considering structural images of patients that will convert to Alzheimer disease (AD), we are interested in revealing the brain patterns of atrophy, explaining the variability in this population. It provides indications of possible stratification of the cohort into homogeneous sub-groups that may be clinically similar but with a different pattern of atrophy. This could suggest different subtypes of patients with AD or some other etiologies such as dementia with Lewy bodies. Clustering methods might be natural approaches to address such situations; however, they can not reveal subtle differences that go beyond a global and trivial pattern of atrophy. Such patterns are usually captured by the first component of PCA, which, after being removed, offers the possibility to identify spatial patterns on the following components. However, PCA provides dense loading vectors (patterns), that cannot be used to identify brain markers without arbitrary thresholding.

Sparse PCA Recently, some alternatives propose to add sparsity in this matrix factorization problem (Li et al., 2015, Mairal et al., 2010, Ramezani et al., 2015). The sparse dictionary learning framework proposed by (Mairal et al., 2010) provides a sparse coding (rows of $\boldsymbol{U}$) of samples through a sparse linear combination of dense basis elements (columns of $\boldsymbol{V}$). However, the identification of biomarkers requires a sparse dictionary (columns of $\boldsymbol{V}$). This is precisely the objective of Sparse PCA (SPCA) proposed in (d'Aspremont et al., 2007; Jolliffe, Trendafilov, and Uddin, 2003; Journée et al., 2010; Witten, Tibshirani, and Hastie, 2009; Zou, Hastie, and Tibshirani, 2006) which adds a sparsity-inducing penalty on the columns of $\boldsymbol{V}$. Imposing such sparsity constraints on the loading coefficients is a procedure that has been used in fMRI to produce a sparse representation of the brain's functional networks (Eavani et al., 2015; Shen et al., 2017). However, sparse PCA ignores the inherent spatial correlation in the data, leading to scattered patterns that are difficult to interpret. Furthermore, constraining only the number of features included in the PCs might not always be fully relevant since most data sets are expected to have a spatial structure. For instance, MRI data is naturally encoded on a grid; some voxels are neighbors, while others are not.

Spatially regularized PCA We hypothesize that brain patterns are organized into distributed regions across the brain(Felleman and Essen, 1991; Korbinian Brodmann, 1909; Rudolf Nieuwenhuys, 2013). Recent studies tried to overcome this limitation by encoding prior information concerning the spatial structure of the data (see (Guo et al., 2015; Jenatton, Obozinski, and Bach, 2010; Wang and Huang, 2015)). However, they used methods that are difficult to plug into the optimization scheme (e. g., spline smoothing, wavelet smoothing) and incorporated prior information that sometimes may be difficult to define. One simple solution is the use of a GraphNet penalty (Dohmatob et al., 2015; Grosenick et al., 2013; Kandel et al., 2013; Mohr et al., 2015; Ng et al., 2012). It promotes local smoothness of the weight map by simply forcing adjacent voxels to have similar weights using an λ_2 penalty on the gradient of the weight map. Nonetheless, we hypothesized that, as with supersized problems, Graph-net should produce a smooth solution rather than clearly delineated regions.

For simplicity, rather than solving Eq. 4.27, we solve a slightly different criterion which results from using the Lagrange form, rather than the bound form, of the constraints on $\boldsymbol{V}$. Then, we extend the

Lagrangian form by adding penalties (ℓ_1, ℓ_2 and TV) to the minimization problem:

$$\begin{aligned} \min_{\boldsymbol{U},\boldsymbol{D},\boldsymbol{V}} & \frac{1}{N} \|\boldsymbol{X} - \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top\|_F^2 \\ & + \sum_{k=1}^{K} \left\{ \lambda_2 \|\boldsymbol{v}_k\|_2^2 + \lambda_1 \|\boldsymbol{v}_k\|_1 + \lambda_s \sum_{g \in \mathcal{G}} \|\boldsymbol{A}_g \boldsymbol{v}_k\|_2 \right\}, \\ \text{s. t. } & \|\boldsymbol{u}_k\|_2^2 = 1, \forall k = 1, \cdots, K, \end{aligned} \tag{4.27}$$

where λ_1, λ_2 and λ_s are hyper-parameters controlling the relative strength of each penalty. We further propose a generic optimization framework that can combine any differentiable convex (penalized) loss function with (i) penalties whose proximal operator is known (here $\|\cdot\|_1$), and (ii) a broad range of complex, non-smooth convex structured penalties that can be formulated as a $\|\cdot\|_{2,1}$-norm defined over a set of groups $\mathcal{G}$. Such group-penalties cover e. g., total variation, and overlapping group lasso.

This new problem aims at finding a linear combination of original variables that points in directions explaining as much variance as possible in data while enforcing sparsity and structure (piecewise smoothness for TV) of the loadings. To achieve this goal, it is necessary to sacrifice some of the explained variance as well as the orthogonality of both the loading and the principal components. Most existing SPCA algorithms (d'Aspremont et al., 2007; Journée et al., 2010; Witten, Tibshirani, and Hastie, 2009; Zou, Hastie, and Tibshirani, 2006), do not impose orthogonal loading directions either. While we forced the components to have unit norm for visualization purposes, we do not, in this formulation, enforce $\|\boldsymbol{v}_k\|_2 = 1$. Instead, the value of $\|\boldsymbol{v}\|_2$ is controlled by the hyper-parameter λ_2. This penalty on the loading, together with the unit norm constraint on the component, prevents us from obtaining trivial solutions. The optional $\frac{1}{N}$ factor acts on and conveniently normalizes the loss to account for the number of samples to simplify the settings of the hyper-parameters: $\lambda_1, \lambda_2, \lambda_s$.

de Pierrefeu et al., 2018c presents an extension of the popular PCA framework by adding structured sparsity-inducing penalties on the loading vectors to identify the few stable regions in the brain images accounting for most of the variability. The addition of a prior that reflects the data's structure within the learning process gives this contribution a scope that goes beyond Sparse PCA. To our knowledge, very few authors ((Abraham et al., 2013; Guo et al., 2015; Jenatton, Obozinski, and Bach, 2010; Wang and Huang, 2015)) addressed the use of structural constraints in PCA. The study (Jenatton, Obozinski, and Bach, 2010) proposes a norm that induces structured sparsity (called SSPCA) by restraining the support of the solution to be sparse with a particular set of group of variables. Possible supports include a set of variables forming rectangles when arranged on a grid. Only one study, recently used the total variation prior (Abraham et al., 2013), in a context of multi-subject dictionary learning, based on a different optimization scheme (Beck and Teboulle, 2009b).

4.3.2 Minimization of PCA with TV penalty

Iteration over components

A common approach to solve the PCA problem, (see d'Aspremont et al., 2007; Journée et al., 2010; Witten, Tibshirani, and Hastie, 2009), is to compute a rank-1 approximation of the data matrix, and then repeat this on the deflated matrix (Mackey, 2009), where the influence of the PCs are successively extracted and discarded.

Single component computation

Given a pair of loading/component vectors, $\boldsymbol{u} \in \mathbb{R}^N, \boldsymbol{v} \in \mathbb{R}^P$, the best rank-1 approximation of the problem given in Eq. 4.27 is equivalent (Witten, Tibshirani, and Hastie, 2009) to:

$$\begin{aligned} \min_{\boldsymbol{u},\boldsymbol{v}} f \equiv & \overbrace{\underbrace{-\frac{1}{N}\boldsymbol{u}^\top \boldsymbol{X}\boldsymbol{v} + \lambda_2\|\boldsymbol{v}\|_2^2}_{g(\boldsymbol{v})}}^{\text{smooth}} + \overbrace{\lambda_1 \underbrace{\|\boldsymbol{v}\|_1}_{h(\boldsymbol{v})} + \lambda_s \underbrace{\sum_{g\in\mathcal{G}} \|\boldsymbol{A}_g\boldsymbol{v}\|_2}_{s(\boldsymbol{v})}}^{\text{non-smooth}} \\ & \text{s. t. } \|\boldsymbol{u}\|_2^2 \leq 1, \end{aligned} \tag{4.28}$$

where $g(\boldsymbol{v})$ is the penalized smooth (i. e.differentiable) loss, $h(\boldsymbol{v})$ is a sparsity-inducing penalty whose proximal operator is known and $s(\boldsymbol{v})$ is a complex penalty on the structure of the input variables with an unknown proximal operator. This problem is convex in $\boldsymbol{u}$ and in $\boldsymbol{v}$ but not in $(\boldsymbol{u}, \boldsymbol{v})$.

Alternating minimization of the bi-convex problem

The objective function to minimize is bi-convex (Boyd and Vandenberghe, 2004). The most common approach to solve a bi-convex optimization problem (which does not guarantee global optimality of the solution) is to alternatively update $\boldsymbol{u}$ and $\boldsymbol{v}$ by fixing one of them at the time and solving the corresponding convex optimization problem on the other parameter vector.

On the one hand, when $\boldsymbol{v}$ is fixed, the problem to solve is

$$\begin{aligned} & \min_{\boldsymbol{u}\in\mathbb{R}^N} -\frac{1}{N}\boldsymbol{u}^\top \boldsymbol{X}\boldsymbol{v} \\ & \text{s. t. } \|\boldsymbol{u}\|_2^2 \leq 1,, \end{aligned} \tag{4.29}$$

with the associated explicit solution

$$\boldsymbol{u}^*(\boldsymbol{v}) = \frac{\boldsymbol{X}\boldsymbol{v}}{\|\boldsymbol{X}\boldsymbol{v}\|_2}. \tag{4.30}$$

On the other hand, solving the equation with respect to $\boldsymbol{v}$ with a fixed $\boldsymbol{u}$ presents a higher level of difficulty. It is solved with the CONESTA algorithm detailed in Sec. 4.2.

Minimization of the loading vectors with CONESTA

Using Nesterov's smoothing of the structured penalty, a new (smoothed) optimization problem, closely related to Eq. 4.28 (with fixed $\boldsymbol{u}$), arises from this regularization as

$$\min_{\boldsymbol{v}} \overbrace{\underbrace{-\frac{1}{n}\boldsymbol{u}^\top \boldsymbol{X}\boldsymbol{v} + \lambda_2\|\boldsymbol{v}\|_2^2}_{g(\boldsymbol{v})} + \lambda_s \underbrace{\left\{\boldsymbol{\alpha}_\mu^*(\boldsymbol{v})^\top \boldsymbol{A}\boldsymbol{v} - \frac{\mu}{2}\|\boldsymbol{\alpha}^*\|_2^2\right\}}_{s_\mu(\boldsymbol{v})}}^{\text{smooth}} + \lambda_1 \overbrace{\underbrace{\|\boldsymbol{v}\|_1}_{h(\boldsymbol{v})}}^{\text{non-smooth}} . \tag{4.31}$$

Since we are now able to explicitly compute the gradient of the smooth part $\nabla(g + \lambda_s s_\mu)$ (Eq. 4.33), its Lipschitz constant (Eq. 4.34) and also the proximal operator of the non-smooth part, we have all the ingredients necessary to solve this minimization function using the CONESTA algorithm.

However, to control the convergence of the algorithm (presented in Sec. 4.2.6), we introduce the Fenchel dual function and the corresponding dual gap of the objective function. The Fenchel duality requires the loss to be strongly convex, which is why we further reformulate Eq. 4.31 slightly: All penalty terms are divided by λ_2, and by using the following equivalent formulation for the loss, we obtain the minimization problem:

$$\min_{\boldsymbol{v}} f_\mu \equiv \overbrace{\underbrace{\frac{1}{2}\|\boldsymbol{v}-\boldsymbol{y}\|_2^2}_{L_\varepsilon(\boldsymbol{v})} + \frac{1}{2}\|\boldsymbol{v}\|_2^2}^{g(\boldsymbol{v})} + \underbrace{\frac{\lambda_s}{\lambda_2}\overbrace{\left\{\boldsymbol{\alpha}_\mu^*(\boldsymbol{v})^\top \boldsymbol{A}\boldsymbol{v} - \frac{\mu}{2}\|\boldsymbol{\alpha}^*\|_2^2\right\}}^{s_\mu(\boldsymbol{v})} + \frac{\lambda_1}{\lambda_2}\overbrace{\|\boldsymbol{v}\|_1}^{h(\boldsymbol{v})}}_{\psi_\mu(\boldsymbol{v})}. \tag{4.32}$$

This new formulation of the smoothed objective function (noted f_μ) preserves the decomposition of f_μ into a sum of a smooth term $g + \frac{\lambda_s}{\lambda_2}s_\mu$ and a non-smooth term h. Such decomposition is required for the application of CONESTA as detailed in Sec 4.2. Moreover, this formulation provides a decomposition of f_μ into a sum of a smooth loss L_ε and a penalty term ψ_μ required for the calculation of the gap presented in Sec. 4.2.6.

We provide all the required quantities to minimize Eq. 4.32. Using Eq. 4.11 we compute the gradient of the smooth part as

$$\begin{aligned}\nabla\left(g+\frac{\lambda_s}{\lambda_2}s_\mu\right) &= \nabla(g)+\frac{\lambda_s}{\lambda_2}\nabla(s_\mu)\\ &= (2\boldsymbol{v}-\boldsymbol{y})+\frac{\lambda_s}{\lambda_2}\boldsymbol{A}^\top\boldsymbol{\alpha}^*_\mu(\boldsymbol{v}^k), \qquad (4.33)\end{aligned}$$

and its Lipschitz constant (using Eq. 4.12)

$$L\left(\nabla\left(g+\frac{\lambda_s}{\lambda_2}s_\mu\right)\right)=2+\frac{\lambda_s}{\lambda_2}\frac{\|\boldsymbol{A}\|_2^2}{\mu}. \qquad (4.34)$$

Based on Eq. 4.32, which decomposes the smoothed objective function as a sum of a strongly convex loss and the penalty,

$$f_\mu(\boldsymbol{v})=L_\varepsilon(\boldsymbol{v})+\psi_\mu(\boldsymbol{v}),$$

we compute the duality gap that provides an upper bound estimation of the error to the optimum. At any step k of the algorithm, given the current primal $\boldsymbol{v}^k$ and the dual $\boldsymbol{\sigma}(\boldsymbol{v}^k)\equiv\nabla L_\varepsilon(\boldsymbol{v}^k)$ variables (Borwein and Lewis, 2006), we can compute the duality gap using the Fenchel duality rules (**Mairal2010**):

$$\textsc{Gap}(\boldsymbol{v}^k)\equiv f_\mu(\boldsymbol{v}^k)+L^*_\varepsilon\left(\boldsymbol{\sigma}(\boldsymbol{v}^k)\right)+\psi^*_\mu\left(-\boldsymbol{\sigma}(\boldsymbol{v}^k)\right), \qquad (4.35)$$

where L^*_ε and ψ^*_μ are respectively the Fenchel conjugates of L_ε and ψ_μ. Denoting by $\boldsymbol{v}^*$ the minimum of f_μ (solution of Eq. 4.32), the interest of the duality gap is that it provides an upper bound for the difference with the optimal value of the function. Moreover, it vanishes at the minimum Eq. 4.19. The dual variable is

$$\boldsymbol{\sigma}(\boldsymbol{v}^k)\equiv\nabla L_\varepsilon(\boldsymbol{v}^k)=\boldsymbol{v}-\frac{\boldsymbol{X}^\top\boldsymbol{u}}{n\lambda_2}, \qquad (4.36)$$

the Fenchel conjugate of the squared loss $L_\varepsilon(\boldsymbol{v}^k)$ is

$$L^*_\varepsilon(\boldsymbol{\sigma}(\boldsymbol{v}^k))=\frac{1}{2}\|\boldsymbol{\sigma}(\boldsymbol{v}^k)\|_2^2+\boldsymbol{\sigma}(\boldsymbol{v}^k)^\top\boldsymbol{y}. \qquad (4.37)$$

The algorithm for the SPCA-TV problem

The computation of a single component through SPCA-TV is achieved by combining CONESTA and Eq. 4.30 within an alternating minimization loop. Mackey (Mackey, 2009) demonstrated that further components can be efficiently obtained by incorporating this single-unit procedure in a deflation scheme as done in e. g. (d'Aspremont et al., 2007; Journée et al., 2010). The stopping criterion is defined as

$$\textsc{StoppingCriterion}=\frac{\left\|\boldsymbol{X}^k-\boldsymbol{u}^{i+1}{\boldsymbol{v}^{i+1}}^\top\right\|_F-\left\|\boldsymbol{X}^k-\boldsymbol{u}^i{\boldsymbol{v}^i}^\top\right\|_F}{\left\|\boldsymbol{X}^k-\boldsymbol{u}^{i+1}{\boldsymbol{v}^{i+1}}^\top\right\|_F}. \qquad (4.38)$$

All the presented building blocks were combined into Algorithm 4 to solve the SPCA-TV problem.

Algorithm 4 SPCA-TV(X, ε)

```
$X_0 = X$
for all $k = 0, \dots, K$ do                                   ▷ Components
    Initialize $u^0 \in \mathbb{R}^N$
    repeat                                                     ▷ Alternating minimization
        $v^{i+1} = \text{CONESTA}(X_k^\top u^i, \varepsilon)$
        $u^{i+1} = \frac{X_k v^{i+1}}{\|X_k v^{i+1}\|_2}$
    until StoppingCriterion $\leq \varepsilon$
    $v_{k+1} = v^{i+1}$
    $u_{k+1} = u^{i+1}$
    $X_{k+1} = X_k - u^{k+1} v^{k+1\top}$                    ▷ Deflation
end for
return $U = [u_1, \cdots, u_K], V = [v_1, \cdots, v_K]$
```

4.3.3 Experiments on synthetic data

We compare the performance of SPCA-TV with existing sparse PCA models: Sparse PCA, ElasticNet PCA, GraphNet PCA and SSPCA from Jenatton, Obozinski, and Bach, 2010.

Performances were evaluated through a 5-fold × 5-fold nested cross-validation. In the outer (external) loop, the sample is split into training and test sets. The test sets are exclusively used for model assessment, while the train sets are used in the inner (internal) loop for model fitting and selection. The inner folds select the set of parameters (over a grid given in de Pierrefeu et al., 2018c), minimizing the reconstruction error on the outer fold.

The reconstruction accuracy was evaluated with the average (across the folds) of the MSE (Mean Squared Error) or Frobenius norm of the error between test data and their reconstructed versions. However, the TV penalty has a more important purpose than to minimize the reconstruction error: the estimation of coherent and reproducible loadings. Therefore the stability of the loading vectors obtained across various training data sets (variation in the learning samples) was assessed through a similarity measure: the pairwise Dice index between loading vectors obtained with different folds/data sets (Dice, 1945).

We simulated 2D datasets with three latent variables whose spatial support is illustrated in Fig. 4.6 (top row) with carefully controlled signal-to-noise ratio (de Pierrefeu et al., 2018c).

Quantitative evaluation of reconstruction and stability

Tab. 4.3 shows that reconstruction error (MSE between reconstructed and true test data), is significantly lower with SPCA-TV than with all other methods: Sparse PCA ($T = 6.9$, $p = 8.0 \cdot 10^{-9}$), ElasticNet PCA ($T = 6.2$, $p = 1.1 \cdot 10^{-07}$), GraphNet-PCA ($T = 4.1$, $p = 1.4 \cdot 10^{-04}$) and SSPCA, from Jenatton, Obozinski, and Bach, 2010, ($T = 22.6$, $p < 10^{-16}$). Moreover, when evaluating the stability of the loading vectors across resampling, we found a higher statistically significant mean Dice index when using SPCA-TV compared to the other methods ($p < 0.001$).

Table 4.3: Scores are averaged across the 50 independent data sets. We tested whether the scores obtained with existing PCA methods are significantly different from scores obtained with SPCA-TV. Significance notations: ***: $p \leq 10^{-3}$

Method	MSE	Dice Index
Sparse PCA	0.91***	0.28***
ElasticNet PCA	0.83***	0.43***
GraphNet PCA	0.83***	0.30***
SSPCA	1.54***	0.07***
SPCA-TV	0.64	0.52

Qualitative evaluation

Fig. 4.6 represents the loading vectors extracted with different methods. Please note that the sign is arbitrary. Indeed, if we consider the loss of Eq. 4.28, $\boldsymbol{u}$ and $\boldsymbol{v}$ can both be multiplied by -1 without changing anything. We observe that the recovered support for the loading vectors of SPCA-TV are sparse, but also organized in clear regions. SPCA-TV provides loading vectors that closely match the ground truth.

Convergence of the algorithm

One of the issues linked to biconvex optimization is the risk of falling into locals minima. Conscious of this potential risk, we set up an experiment in which we ran 50 times the optimization of the same problem, with a different starting point at each run. We then compared the resulting loading vectors obtained at each run and computed a similarity measure, the Dice index. It quantifies the proximity between each independently-run solution with a different starting point. We obtained a Dice index of 0.99 on the 1st component, 0.99 on the 2nd component, and 0.72 on the 3rd component. Off the strength of this

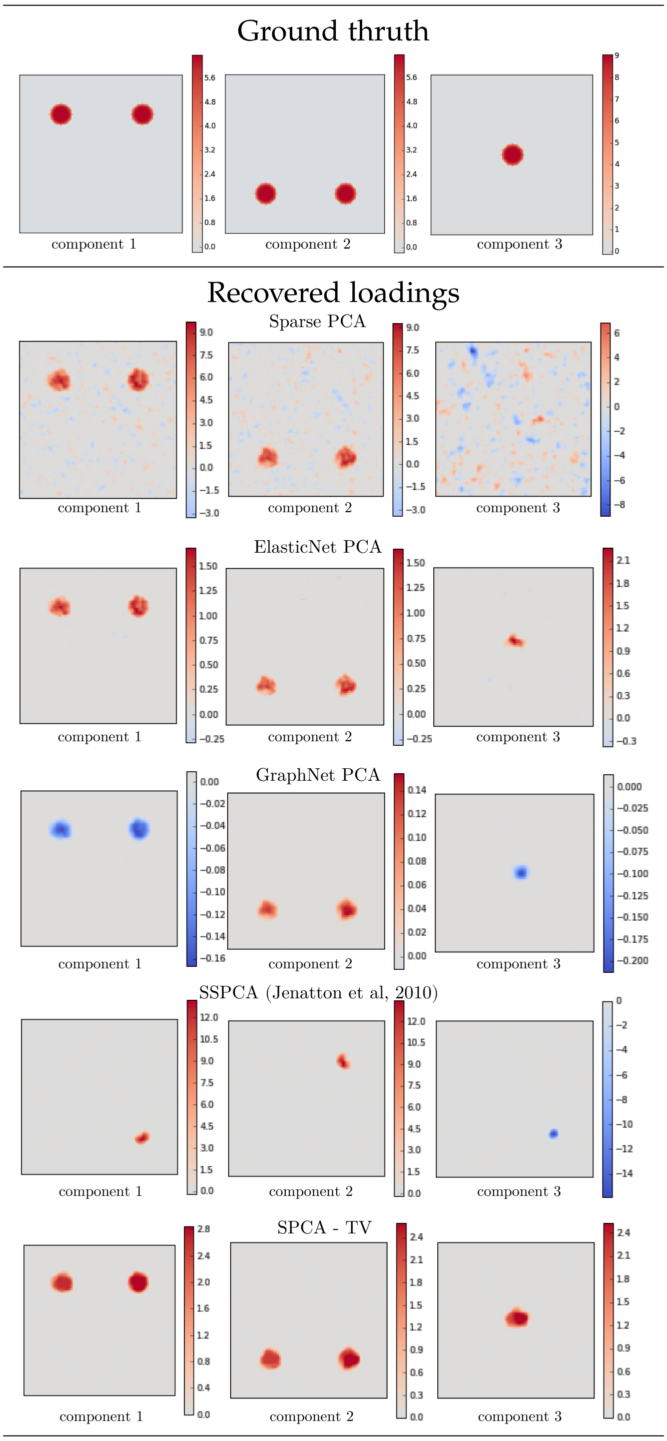


Figure 4.6: Synthetic data. Top row: the true loading vectors used to generate the images. Following rows present, the loading vectors recovered using different sparse models.

indices, we are confident of this algorithm's robustness and ability to converge toward the same stable solution independently from the choice of the starting point.

4.3.4 Experiments on meshes of cortical thickness in Alzheimer disease

SPCA-TV was applied to the whole brain cortical thickness (317 379 features) of 133 patients with a diagnosis of mild cognitive impairments (MCI) from the ADNI database. We did not use SSPCA, which is restricted to array images and does not support meshes of cortical surfaces.

Quantitative evaluation of reconstruction and stability

The reconstruction Tab. 4.4 error is significantly lower in SPCA-TV than in Sparse PCA ($T = 12.7$, $p = 2.1 \cdot 10^{-4}$), ElasticNet PCA ($T = 6.8$, $p = 2.3 \cdot 10^{-3}$) and GraphNet PCA ($T = 2.83$, $p = 4.7 \cdot 10^{-2}$). The results are presented in Table 4.4. Moreover, when assessing the stability of the loading vectors across the folds, the mean Dice index is significantly higher in SPCA-TV than in other methods.

Table 4.4: Scores are averaged across the 5 folds. We tested whether the averaged scores obtained with existing PCA methods are significantly lower from scores obtained with SPCA-TV. Significance notations: ***: $p \leq 10^{-3}$, **: $p \leq 10^{-2}$, *: $p \leq 10^{-1}$.

	Scores	
Methods	Test Data Reconstruction Error	Dice Index
Sparse PCA	2991.8***	0.44**
ElasticNet PCA	2832.6**	0.43**
GraphNet PCA	2813.6*	0.62*
SPCA-TV	2795.0	0.65

Clinical interpretation of coefficient maps

The loading vectors obtained from the data set with sparse PCA and SPCA-TV are presented in Fig. 4.7. As expected, Sparse PCA loadings are not easily interpretable because the patterns are irregular and dispersed throughout the brain surface. In contrast, SPCA-TV reveals structured and smooth clusters in relevant regions. The first loading vector, which maps the whole surface of the brain, can be interpreted as the variability between patients, resulting from global cortical atrophy, as often observed in AD patients. The second loading vector includes variability in the entorhinal cortex, hippocampus, and temporal regions. Last, the third loading vector might be related to the atrophy of the frontal lobe and captures variability in the precuneus too. Thus, SPCA-TV provides a smooth map that closely matches the well-known brain regions involved in Alzheimer's disease (Frisoni et al., 2010).

Indeed, it is well-documented that cortical atrophy progresses over three main stages in Alzheimer's disease (Braak and Braak, 1991; Delacourte et al., 1999) The cortical structures are sequentially being affected because of the accumulation of amyloid plaques. Cortical atrophy is first observed, in the mild stage of the disease, in regions surrounding the hippocampus (Jack et al., 2004, Ridha et al., 2008, and Thompson et al., 2004) and the enthorinal cortex (Cardenas et al., 2011), as seen in the second component. This finding is consistent with early memory deficits. Then, the disease progresses to a moderate stage, where atrophy gradually extends to the prefrontal association cortex, as revealed in the third component (McDonald et al., 2009). In the severe stage of the disease, the whole cortex is affected by atrophy (Delacourte et al., 1999) (as revealed in the first component).

Clinical interpretation of neuroimaging latent scores

To assess the clinical significance of these weight maps, we tested the correlation between the components' scores and the subjects' performance on the ADAS (The Alzheimer's Disease Assessment Scale-Cognitive subscale) clinical test. ADAS is scored in terms of errors, so a high score indicates poor performance. We obtained significant correlations between ADAS test performance and components 'scores in Figure 4.8. $r = -0.34, p = 4.2 \cdot 10^{-11}$ for the first component, $r = -0.26, p = 3.6 \cdot 10^{-7}$ for the second component and $r = -0.35, p = 4.0 \cdot 4.5^{-12}$ for the third component). The same behavior is observable for all three components: The ADAS score grows proportionately to the level to which a patient is affected and to the severity of atrophy he presents (in temporal pole, prefrontal region, and also globally). Conversely, controls subjects score low on the ADAS metric and present low levels of cortical atrophy. Therefore, SPCA-TV provides us with precise biomarkers, that are entirely relevant to the scope of Alzheimer's disease progression.

Sensitivity analysis

We conducted a sensitivity analysis on the real neuroimaging data set to increase the understanding of the relationships between input parameters and output weight maps.

- First, increase of ℓ_{TV} (Fig. 4.9) results in a more structured and smoother map. In addition, it tends to increase the extent of the support, even with a fixed ℓ_1.

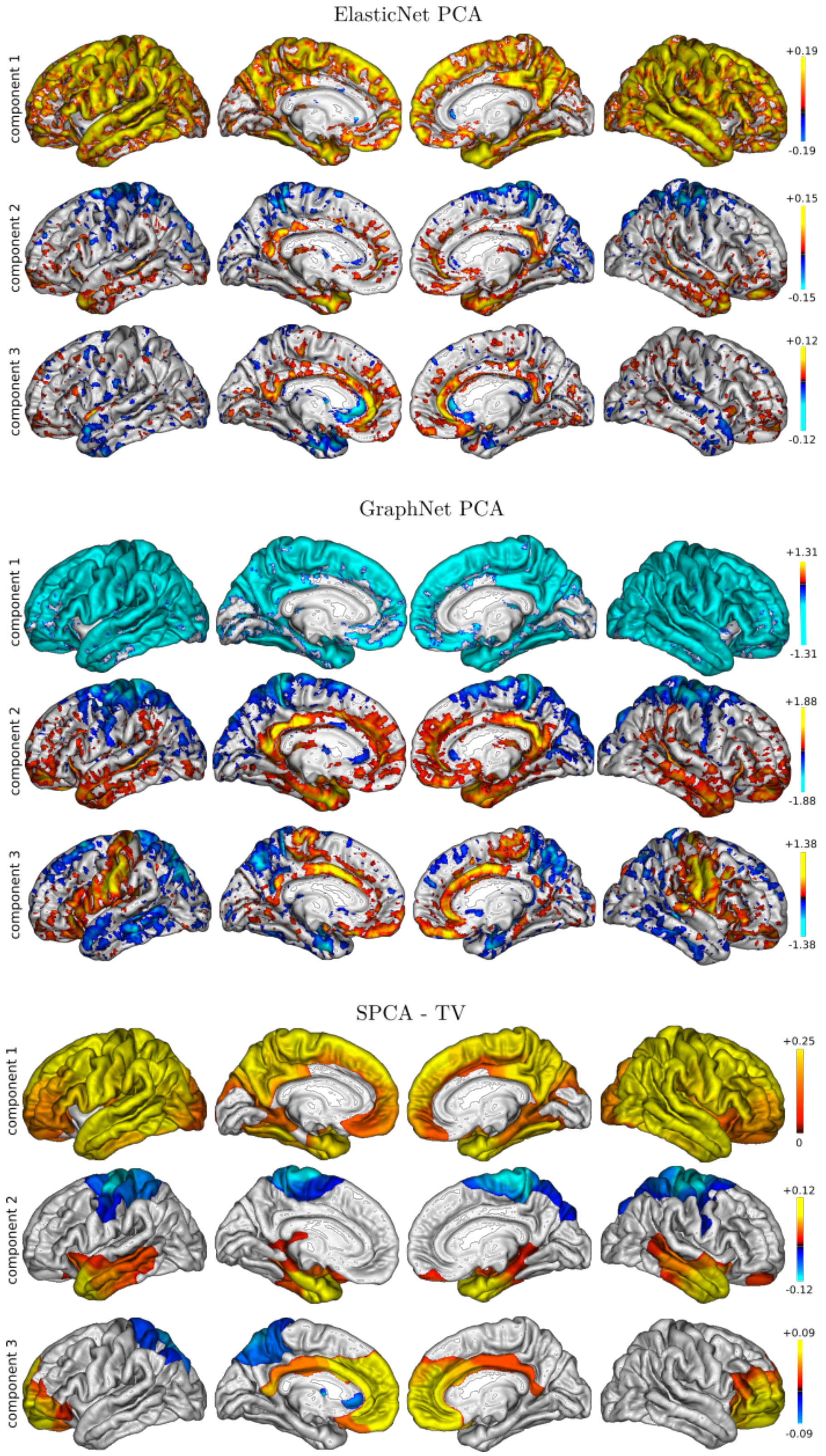


Figure 4.7: Loading vectors recovered from the 133 MCI patients using different methods

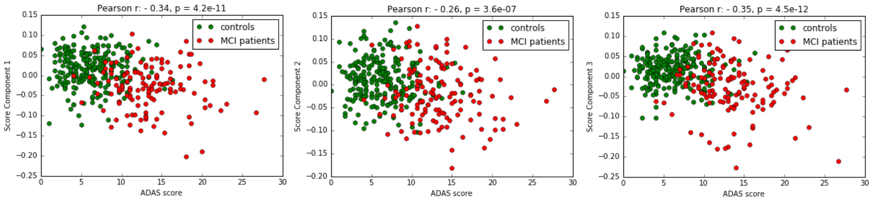


Figure 4.8: Correlation of components scores with ADAS test performance.

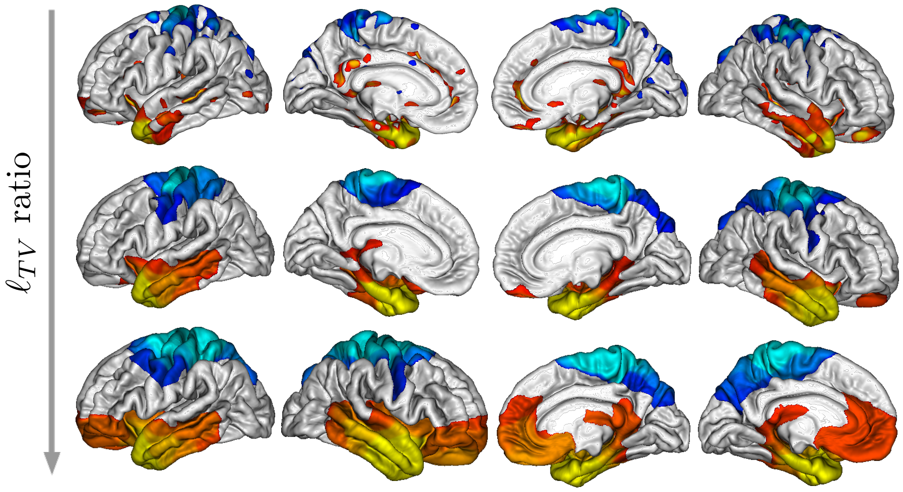


Figure 4.9: Sensitivity analysis : Effect of the variation of the ℓ_{TV} ratio parameter on the weight maps.

- Second, increase of ℓ_1 (Fig. 4.10) results, as expected, in a more parsimonious map.

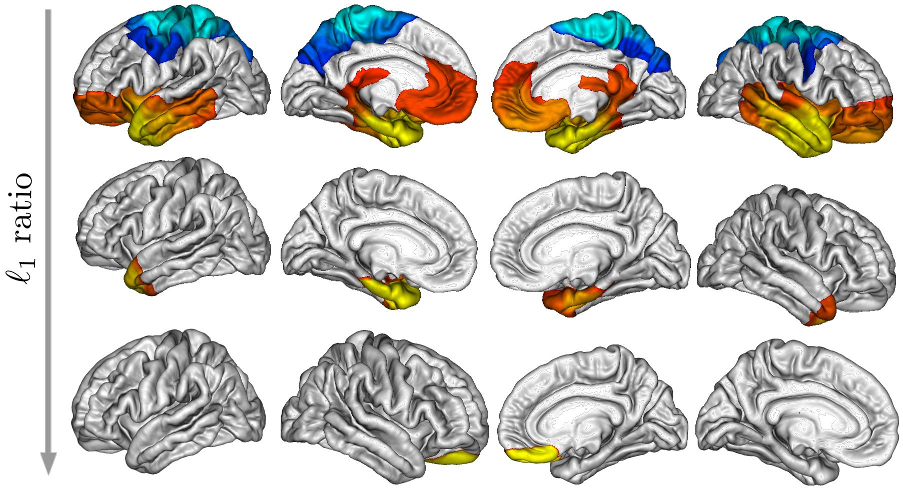


Figure 4.10: Sensitivity analysis : Effect of the variation of the ℓ_1 parameter on the weight maps.

- Last, increasing α Fig. 4.11 produces a solution that evolves from a dense a map (dominated by ℓ_2), to a structured map (dominated by TV) and, finally an extremely sparse map (dominated by ℓ_1).

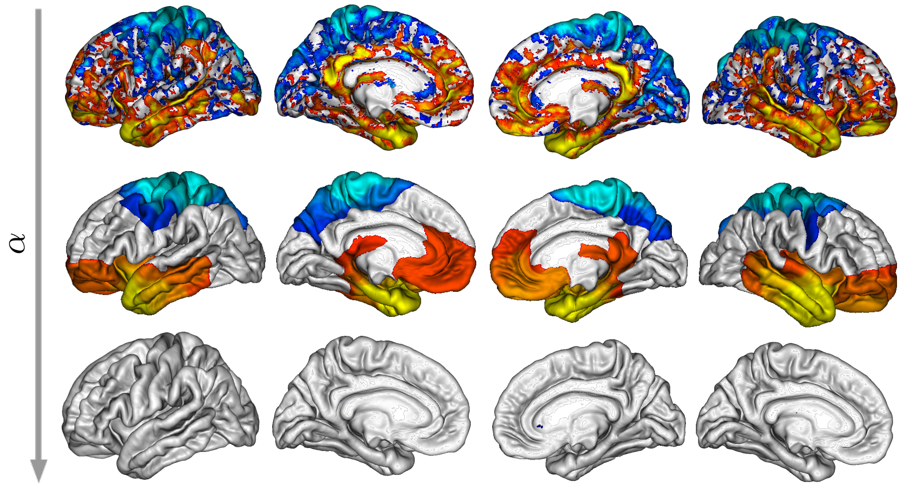


Figure 4.11: Sensitivity analysis : Effect of the variation of the α parameter on the weight maps.

It is also interesting to note the extreme effects of ℓ_{TV} and ℓ_1 parameters: Extremely high values of these two parameters tend to push the solution toward two opposite weight maps configuration. An extremely high value of ℓ_{TV} produces very extended support with constant coefficient values. On the other hand, an extremely high value of ℓ_1 tends to yield fully sparse weight maps where every coefficient have a zero coefficient.

4.3.5 Conclusion

We proposed an extension of Sparse PCA that takes into account the spatial structure of the data. We observe that SPCA-TV, in contrast to other existing sparse PCA methods, yields clinically interpretable results, and reveals significant sources of variability in data, by highlighting structured clusters of interest in the loading vectors. Furthermore, SPCA-TV 's loading vectors were more stable across the learning samples compared to other methods.

5 Identification of predictive signatures of brain disorders

This chapter presents clinical applications of our methodological contributions, in psychiatry:

- Sec. 5.1 summarizes de Pierrefeu et al., 2018a which demonstrates tha ML with spatial regularization (ElasticNet-TV) working at a voxel level can identify a reproducible neuroanatomical signature of Schizophrenia;
- Sec. 5.2 outlines de Pierrefeu et al., 2018b which shows that ElasticNet-TV can identify an interpretable functional predictive signature (clusters in speech-related brain regions) of the upcoming hallucinations in patients with schizophrenia;

and neurology:

- Sec. 5.3 presents an application of spatialy regularized PCA to identify white matter hyperintensities spatial patterns of variability in patient with CADASIL syndrome, which has been published in Duchesnay et al., 2018.

5.1 Identifying a neuroanatomical signature of schizophrenia

In de Pierrefeu et al., 2018a, we used anatomical MRI to learn a predictor of the schizophrenia that generalizes to the early stage of the disorder while providing insight into the neurobiological predictive signature.

5.1.1 Introduction

Schizophrenia is a disabling chronic mental disorder characterized by various symptoms such as hallucinations, delusions as well as impairments in high-order cognitive functions. The development of magnetic resonance imaging (MRI) provides an effective and noninvasive approach to investigate the neuroanatomy of the brain. Specifically, structural MRI (sMRI) allows the study of structural changes in the brain and their relationship with the clinical diagnosis. Over the

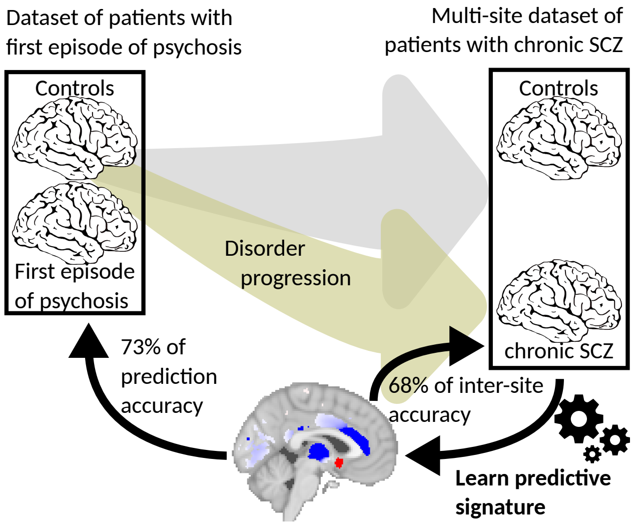


Figure 5.1: Learn a predictive signature on patients with chronic schizophrenia that generalizes to the early stage of the disorder, i.e., patients with first episode of psychosis.

years, sMRI has been increasingly used to gain insights on the structural abnormalities inherent to the disorder and to identify brain regions where schizophrenia patients differ significantly from healthy controls (van Erp et al., 2016). Unfortunately, group analyses do not offer the possibility to uncover individual subject deviations from normality. There is indeed a broad overlap between brain-imaging measurements in schizophrenia patients and the normal range (Sun et al., 2009). Thus, group analyses cannot be easily used to assist in the diagnosis process.

Recent progress in machine learning, together with the availability of large datasets, paves the way for automatic detection of brain disorders, solely based on MRI data (Kambeitz et al., 2015; Orrù et al., 2012). In the past, an extensive number of studies have focused on the prediction of schizophrenia based on neuroanatomical features (Lu et al., 2016; Rozycki et al., 2017; Sabuncu, Konukoglu, Initiative, et al., 2015). These studies uncovered relevant structural brain patterns that are different between controls and patients and that achieve a prediction at the individual level. Based on these structural discrepancies alone, classifiers reached various prediction performances ranging from 65% to 90% of accuracy. However, to date, despite initial promising results, these studies have barely impacted clinical practice. Significant challenges still need to be tackled for translational implementation of such findings in psychiatry.

Schizophrenia is a complex and very heterogeneous disorder. Small size cohorts, typically composed of highly-selected patients, suffer from a bias in the recruitment. They do not represent the full and broad cross-sectional spectrum of the disorder phenotype. Given this variability, a significant heterogeneity can be found in the effect-sizes and patterns of brain differences across studies (Adriano et al., 2010; Shepherd et al., 2012; Vita et al., 2006). To date, most studies recruited subjects scanned at a single acquisition site (i.e., the subjects were scanned at the same site, using similar scanner hardware and MRI protocols). Such results are difficult to generalize to large-scale clinical settings, i.e., with patients scanned at widely different locations (Orban et al., 2018). Validation on independent datasets is a more realistic approach to quantifying generalization accuracy. Consequently, multi-site populations can ensure consistency and reproducibility in the results. To our knowledge, only few studies have relied on a completely independent validation cohort to estimate prediction performances of a classifier (Kawasaki et al., 2007; Nieuwenhuis et al., 2012; Rozycki et al., 2017)

Leveraging those studies, we intend to further develop our findings along two different aspects. First, in the context of predictive signature discovery, it is crucial to understand the brain's structural patterns that underpin a prediction. Unfortunately, in most cases, despite accurate prediction performance achieved, classifiers still behave as a "black box" model, not providing objective neuroanatomical markers thus ruling out the prospect of clinical application. We will therefore focus on the interpretability of such predictive patterns. Second, we strive to filter-out chronic pharmaceutical treatments' impact on the brain. Given that the literature has consistently reported that some regions of the brain are affected by antipsychotic medication (Radua et al., 2012), our intention is to evaluate the generalization of the developed predictive models on subjects that are still in an early stage of the disease. Hence, we need to address the non-negligible probability that previous classifiers rely heavily on the medication impacts over the brain rather than as "true" markers of the disorder able to distinguish healthy individuals from those affected by schizophrenia.

Here, we validated automatic methods to classify schizophrenia using exclusively sMRI scans. We tested different sMRI-based features to assess inter-site performance replicability using data from 606 subjects scanned at four distinct sites with no prior coordination. In addition, we investigated the interpretability of the obtained neuroanatomical predictive signature and its independance regarding medication. Finally, we tested the ability of our classifiers to generalize to an independent set of patients with first-episode psychosis.

5.1.2 Methods

Brain imaging data from 4 independent studies with no prior coordination were gathered in the current analysis (`http://schizconnect.org`). The training dataset included 276 patients with strict schizophrenia, according to DSM-IV criteria, and 330 healthy controls. One additional independent set of healthy controls and patients with first-episode psychosis (FEP) was used for additional validation of the prediction performance. We compared the predictive power of 3 types of features extracted from anatomical MRI: *VBM*: Gray matter maps (125,959 features); *Vertex-based cortical thickness* (299,862 features) and Regions of interest features (66 structural measurements of regions of interest were extracted with FreeSurfer).

Classifiers We compared linear Support Vector Machine (SVM) and logistic regressions with respectively ElasticNet, GraphNet and TV-Enet penalties, implemented in the Parsimony package.

Evaluation on an independent set of first-episode patients The impact of antipsychotic treatments on brain anatomy has been previously reported in the literature (Radua et al., 2012; Roiz-Santiañez, Suarez-Pinilla, and Crespo-Facorro, 2015). It raises questions about the validity of the learned models and the predictive signature. Our concern was that patients and controls might be classified with regard to their medication status rather than their diagnosis. In order to discard the hypothesis of a confounding effect of medication on discriminative patterns, we conducted two additional analyses.

We trained the classifier by masking out the regions that are known to be affected by antipsychotic drugs, such as the striatum (Smieskova et al., 2009; Torres et al., 2013). We created a new predictive model using the remaining features and evaluated its performance. We used Leave-one-site-out (LOSO) procedure for model selection. As a test sample, we used an independent of 133 subjects: 90 healthy controls and 43 participants with first episode-psychosis. Some of those patients have taken antipsychotic medication. However, the duration of treatment is minimal (average: 2.56± 5.1 months). Thus, we assumed that the medication impacts on the brain are very limited in this cohort.

Stability of coefficient maps We assessed the stability of coefficient maps across re-sampling using the average correlation (denoted r_w) and Dice index between pairs of weights maps computed across the CV folds. This measure of stability was evaluated on the weight maps provided by the sparse classifiers: Enet, GraphNet and Enet-TV. Indeed, SVM yields dense weight maps, and thus comparing the region selected across fold is not relevant.

5.1.3 Results

Evaluation on independent set of first-episode patients

Prediction performances presented in Tab. 5.1 demonstrate that:

- VBM features outperformed +10% cortical thickness features.

- On VBM, all models provide similar prediction performances.
- TV regularization provides a significant breakthrough in terms of stability of coefficient maps. The average correlation between maps (r_w) is drastically increased. Note that similar finding is obtained with other measures of similarities between maps: Dice index and Fleiss-Kappa statistic.

Table 5.1: Prediction performances on an independent cohort of controls and patients with first-episode psychosis. The models were learned on a multi-sites cohort of controls and patients with chronic schizophrenia. Scores: Balanced accuracy (bAcc), AUC of ROC analysis, and the predictive maps' stability measured as the average correlation r_w between pairs of maps across the CV folds. All accuracies and AUCs were significant with $p \leq 10^{-2}$

Features	**Model**	**AUC**	**bAcc**	r_w
Gray Matter VBM	SVM	0.78	0.71	-
	Enet	0.78	0.73	0.34
	GraphNet	0.79	**0.76**	0.42
	TV-Enet	**0.80**	**0.76**	0.74
Vertex based cortical thickness	SVM	0.68	0.64	-
	Enet	0.65	0.62	0.09
	GraphNet	0.63	0.60	0.19
	TV-Enet	0.67	0.62	**0.76**
ROIs based volume	SVM	0.72	0.66	-

Neuroanatomical predictive signature

Predictive weight maps are presented in Figs. 5.2 and 5.3:

ℓ_2 regularization of SVM classifier produces a dense map with high-frequency changes and, more confusing, rapid sign flipping observable as with irregular alternate of red and blue values. Let us consider the scenario observed in both Figs. 5.2 and 5.3, of two nearby regions A and B with negative (blue) value in A and positive (red) values in B. In blue A, an increase of GM leads reduces the risk of being predicted as a patient. While in nearby red B, the opposite conclusion can be drawn. Such a situation rules out the interpretability of the solution provided by ℓ_2 regularization.

ElasticNet and GraphNet solutions are scattered and unstable (Tab.5.1)

TV-Enet provides a smooth map made of several clearly identifiable regions. The stability of maps is considerably increased com-

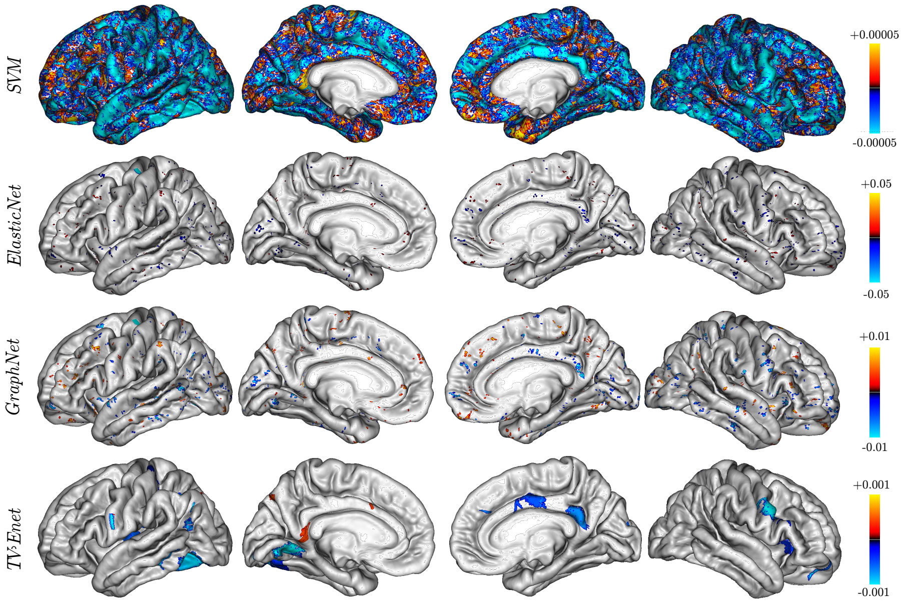


Figure 5.2: Freesurfer predictive signatures obtained with the classifiers-SVM, ElasticNet, GraphNet and Enet-TV

pared to Enet and GraphNet on both GM VBM (0.74 vs 0.34 and 0.42) and meshes of cortical thickness (0.76 vs 0.09 and 0.19).

Brain score and symptomatic level

The neuroanatomical predictive signature ($\boldsymbol{w}$) can be applied to a dataset $\boldsymbol{X}$ to produce the decision function ($\boldsymbol{X}\boldsymbol{w}$), i. e., an individual brain score of the disorder for each patient. In a post hoc analysis, we investigated to what extent this neural score can track the symptomatic level.

The VBM brain score, of 118 patients, was correlated (controlling for the effects of age and gender) with patients' cognitive functions (Crystallized intelligence, Working memory, Episodic memory, and Executive functions) and scales of symptoms dimensions (Scale for the Assessment of Positive Symptoms, SAPS and the Scale for the Assessment of Negative Symptoms, SANS)

We found significant positive correlations between the VBM predictive signature and both, the negative symptoms scores (r = 0.17, $p = 3.5e^{-2}$) and the positive symptoms scores (r = 0.18, $p = 2.2e^{-2}$). The predictive signature also correlated with the extent of cognitive deficits in all domains tested: Crystallized intelligence, working memory, episodic memory and executive functions. Fig. 5.4 illustrates one of those correlation between the brain score and the positive symptoms score (SAPS) of patients.

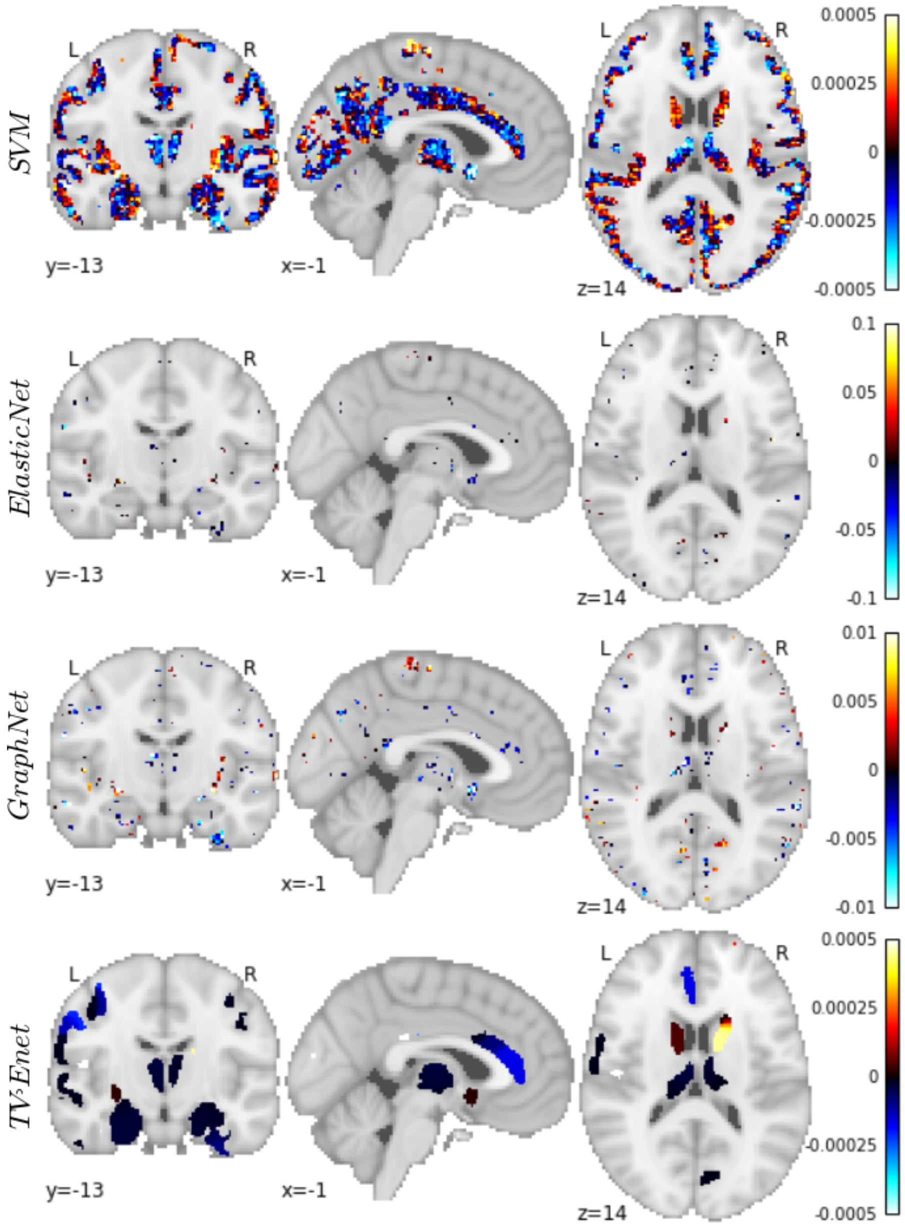


Figure 5.3: VBM predictive signatures obtained with the classifiers- SVM, ElasticNet, GraphNet and Enet-TV.

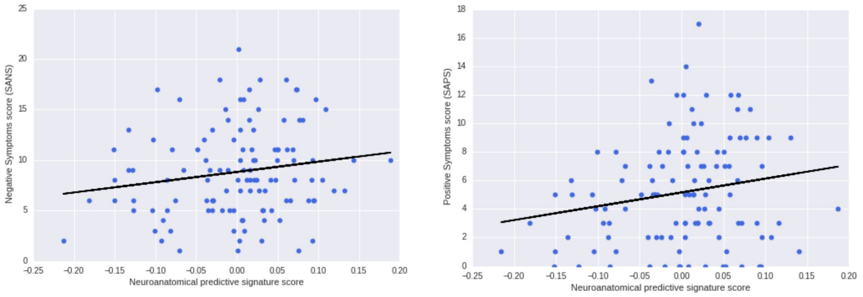


Figure 5.4: Correlation between the neuroanatomical signature score and the negative and positive symptoms scores (SANS and SAPS) of patients.

5.1.4 Discussion

Neuroanatomical predictive signature

The interpretation of the coefficient map is not straightforward. As raised by some papers (Haufe et al., 2014; Kia et al., 2017; Weichwald et al., 2015), we are facing a backward "decoding" problem where we intend to predict the causal clinical status given the brain phenotypes results. Some coefficients can capture a general variability associated to a latent variable (typically the age) that is not specific to the disease of interest. Conversely, some regions may be overlooked due to the sparsity constraint.

Only TV-penalization provides the opportunity to discuss identified regions (Figs. 5.5 and 5.6): Those regions appear largely consistent with available neural data in schizophrenia and may fill the criteria to become a biomarker of the disorder. We indeed found that classification of patients with schizophrenia relied on reduced gray matter compared to healthy controls in the cingulate gyrus, precentral and postcentral gyrus, temporal pole, hippocampus, amygdala, and thalamus. These regional deficits of gray matter in schizophrenia patients have been consistently reported in univariate studies (Fornito et al., 2009; Glahn et al., 2008; Honea et al., 2005; Kim, Kim, and Jeong, 2017; Torres et al., 2016). On the other hand, we found a regional increase of gray matter in schizophrenia patients compared to healthy controls in the putamen, caudate, and pallidum. These local increased GM in schizophrenia were also frequently reported in previous studies (Fornito et al., 2009; Glahn et al., 2008; Honea et al., 2005; Kim, Kim, and Jeong, 2017; Pol et al., 2001; Torres et al., 2016).

Furthermore, significant correlations were found between this predictive signature and both negative and positive symptom scores. Such a result is consistent with the literature where negative symptoms have already been reported to be associated with the extent of structural brain abnormalities in schizophrenia (Ren et al., 2013; Rozycki et al., 2017). Additionally, the neural score obtained from the predictive signature is also correlated with the extent of cognitive impairments in all the domains that are known to be impacted in schizophrenia. This result is promising since it paves the way towards the use of a neuroanatomical signature as an objective measure to monitor the evolution of the disorder.

To demonstrate the clinical relevance of predictive models, the next step would be to evaluate the specificity of the classifiers in differential diagnosis situations. There is now an urgent need for transdiagnostic studies able to compare the specificity of the identified neuroanatomical predictive signature in schizophrenia but also in bipolar disorder or autism spectrum disorder.

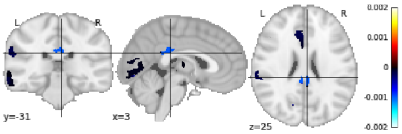
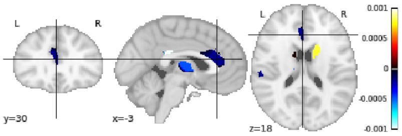
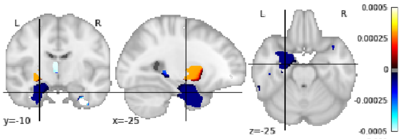
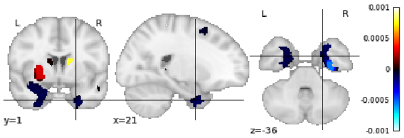
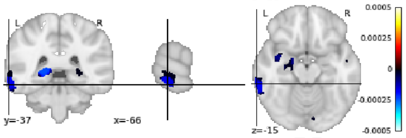
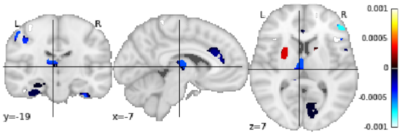
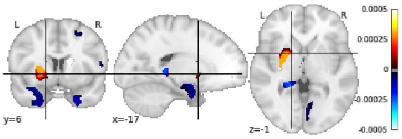
Cluster size (voxel)	Cluster mean weight	Regions involved	Visualisation
240	$-5.9 \cdot 10^{-4}$	Cingulate Gyrus	
1934	$-1.2 \cdot 10^{-4}$	Paracingulate Gyrus, Cingulate Gyrus	
4608	$-3.2 \cdot 10^{-5}$	Left Hippocampus, Left Amygdala	
4279	$-5 \cdot 10^{-5}$	Parahippocampal gyrus, Temporal Pole	
303	$-1.8 \cdot 10^{-5}$	Middle Temporal Gyrus	
864	$-2.9 \cdot 10^{-5}$	Left thalamus	
3173	$3.2 \cdot 10^{-5}$	Left Caudate, Left Putamen, Left Pallidum, Left Lateral Ventricle	

Figure 5.5: Main discriminative regions found in VBM.

5.1.5 Conclusion

These results highlight the existence of a neuroanatomical signature of schizophrenia, shared by a majority of patients across different sites and already present at the early stage of the disorder. Moreover, this signature is associated with the symptoms severity and the amount of cognitive deficit.

5.2 Functional patterns to predict hallucinations in Schizophrenia

In de Pierrefeu et al., 2018b, we demonstrated that supervised classification methods can accurately learn a functional MRI-based pattern that predicts the imminence of a hallucinatory episode. Thus, leveraging real-time pattern decoding capabilities and applying them in

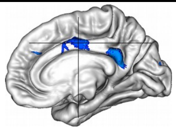
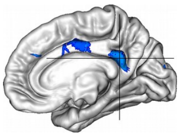
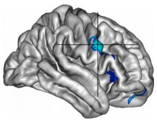
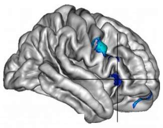
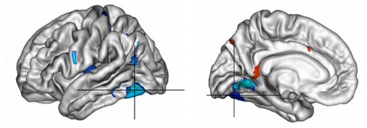
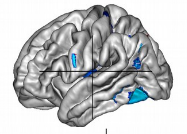
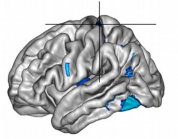
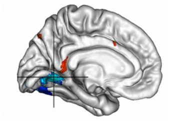
Cluster size (voxel)	Cluster mean weight	Regions involved	Visualisation
678	-0.10	Right Posterior Cingulate Gyrus	
653	-0.12	Isthmus of Right Cingulate Gyrus	
1085	-0.23	Right Precentral Gyrus	
675	-0.10	Right Insula	
2216	-0.46	Left Inferior Temporal Gyrus	
248	-0.05	Left Inferior Parietal Gyrus	
256	-0.03	Left Postcentral Gyrus	
591	-0.18	Left Lingual Gyrus	

Figure 5.6: Discriminative regions found with cortical thickness.

the case of hallucinations could lay the foundation for alternative solutions for affected patients in the near future, such as fMRI-based neurofeedback.

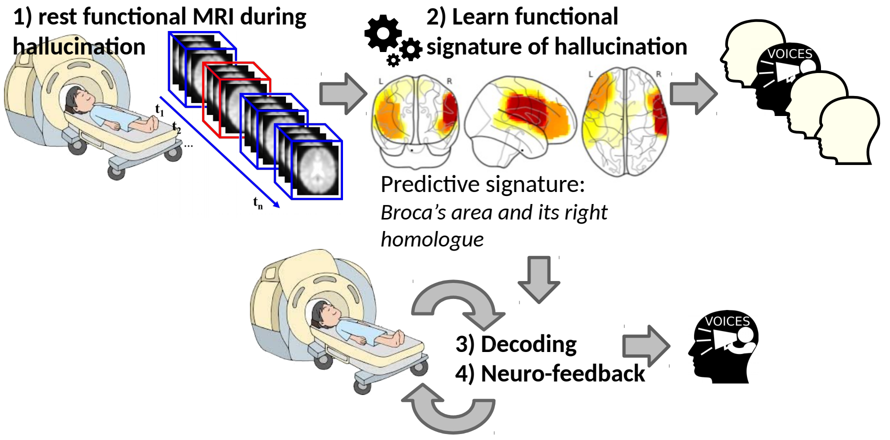

1) rest functional MRI during hallucination
t1
t2
...
tn
2) Learn functional signature of hallucination
L
R
VOICES
Predictive signature:
Broca's area and its right homologue
3) Decoding
4) Neuro-feedback
VOICES

Figure 5.7: Functional MRI activation patterns to predict hallucinations in Schizophrenia for neurofeedback.

5.2.1 Introduction

Hallucinations are defined as abnormal perceptions in the absence of causative stimuli. These experiences, especially auditory hallucinations, constitute fundamental features of psychosis (64-80% lifetime prevalence among schizophrenia-diagnosed patients) and can lead to functional disability and low quality of life (McCarthy-Jones et al., 2017).

Over the past years, auditory hallucinations have been studied in-depth with brain imaging methods, such as functional and structural magnetic resonance imaging (fMRI and sMRI), to decipher their underlying neural mechanisms. Numerous abnormalities have been found in patients suffering from auditory hallucinations (e.g., Allen et al., 2008; Bohlken, Hugdahl, and Sommer, 2017; Jardri et al., 2011; Sommer et al., 2008. Beyond location, the functional dynamics of the neural networks involved in auditory hallucinations have also been studied in several studies that focused on so-called intrinsic connectivity networks (ICN) and their potential role in the onset of hallucinations Alderson-Day et al., 2016; Northoff and Qin, 2011. ICNs typically reveal interactions among brain regions when the subject is not engaged in any particular task. Reported networks include the default mode network (DMN), the control executive network (CEN), the salience network (SAL) and the sensorimotor network (SMN) (Alderson-Day et al., 2016). Numerous studies have asserted that fluctuations in those ICNs are associated with the onset of hallucination periods. For instance, the emergence of hallucinations correlates with a disengagement of the DMN (Jardri et al., 2013). More recently, stochastic effective connectivity analyses revealed complex interactions among hallucination-related networks, DMN, SAL and CEN, during the ignition, active phase, and extinction of hallucinatory experiences (Lefebvre et al., 2016).

Despite significant progress in the field, “capturing” the neural correlates of subjective mental events (such as hallucinations) remains a time-consuming task with multiple post-processing steps and analyses. However, recent progress in machine learning has now paved the way for real-time automatic fMRI decoding of hallucination-related patterns. Such developments may have crucial impacts on the implementation of innovative fMRI-based therapy for drug-resistant hallucinations, such as fMRI-based neurofeedback (Arns et al., 2017; Fovet, Jardri, and Linden, 2015). During fMRI-based neurofeedback, brain activity is measured and fed back in real-time to the subject to help her/him progressively achieve voluntary control over her/his own neural activity. Precisely defining strong a priori strategies for choosing the appropriate target brain area/network(s) for fMRI-based protocols appears critical.

The feasibility of fMRI-based neurofeedback relies on robust and reliable classifying performances and on the ability to detect hallucinations sufficiently early to allow the patients the necessary time to modulate their cerebral activity (Fovet et al., 2016). Rather than detecting hallucinatory events per se, we aim to help patients become aware of the imminence of this experience based on online detection of fMRI signal changes in key networks involved in the ignition of hallucinations. Thus in this study, we specifically focused on the period preceding the occurrence of a hallucination, i.e., the few seconds corresponding to the brain’s transition from a resting-state to a full hallucinatory state. Interestingly, previous fMRI studies have noted the existence of specific fMRI changes prior to hallucinations (Diederen et al., 2010; Hoffman et al., 2008; Lefebvre et al., 2016; Lennox et al., 1999).

Among the current machine-learning approaches available for fMRI analysis, multi-voxel pattern analysis (MVPA), a supervised classification method, is gaining recognition for its potential to discriminate between complex cognitive states (Fovet et al., 2016; Haxby, Connolly, and Guntupalli, 2014). MVPA seeks to identify significantly reproducible spatial activity patterns differentiated according to mental states. Extending these methods to the prediction of the phenomena of transition towards hallucinations should provide better insight into the mechanisms of these subjective experiences. Thus, leveraging real-time pattern decoding capabilities and applying them in the case of hallucinations could lay the foundation for potential solutions for affected individuals.

Variations in transition-to-hallucination functional patterns from one patient to another (e.g., due to phenomenological differences) and from one occurrence to the next (e.g., depending on the modalities involved) appears to be the potential major shortcomings in developing an effective classifier. Indeed, such disparities may inexorably lead to a decrease in decoding performances. Therefore, characterizing the variability within the pre-hallucination patterns across subjects and occurrences is highly desired. Principal component analysis (PCA) is one such unsupervised method that has been successfully applied in the analysis of the variability of a given dataset. The principal components (PCs) and the associated basis patterns shed light on the intrinsic structures of the variability present in a dataset. This unsupervised approach is complementary to the supervised approach described above, as it can help with interpreting the classification performances.

Here, we applied both supervised and unsupervised machine-learning methods to an fMRI dataset collected during hallucinatory episodes. The goal of this paper was two-fold: i) to predict the activation patterns preceding hallucinations using a supervised analysis and ii) to uncover the variability in these activation patterns during the emergence of hallucinations using unsupervised analysis. The goals of these two analyses appear completely complementary in the context of future fMRI-based clinical and therapeutic applications.

5.2.2 Methods

Participants and experimental paradigms

The population was composed of 37 patients with schizophrenia (DSM-IV-TR criteria) who were suffering from very frequent multimodal hallucinations (i.e., more than 10 episodes/hour).

fMRI was acquired at rest. Participants were asked to lie in the scanner in a state of wakeful rest with their eyes closed. The subjects experienced an average of 4 hallucinatory episodes per session. The patients' states at different acquisition times were labelled using a semi-automatic difficult procedure, as described in (Jardri et al., 2013; Leroy et al., 2017) and were assigned to one of the following four categories: transition towards hallucinations (trans), on-going hallucinations (on), no hallucinations (off) and end of hallucinations (end).

This labelling task is a non-straightforward two-steps strategy; the first step is a data-driven analysis of the fMRI signal using an ICA in the spatial domain. The second step involves the selection of the ICA components associated with possible sensory experiences that occurred while scanning. This pipeline is said to be semi-automatic since it combined the following: (a) an automatic denoising part, and (b) a manual and time-consuming part, with the use of an immediate post-fMRI interview conducted with the patient, in which the sensory modalities, number of episodes, and phenomenological features of the experiences were specified. Usual fMRI pre-processing were performed leading to 67,665 voxels in the MNI referential.

Computation of samples

Before training classifiers, the first step involved computing samples from the fMRI signal. The intention was to convert the fMRI signal into a brain map that could capture reflecting an evolution toward hallucination. From each set of consecutive images within a pre-hallucination state ("trans" periods) or "off" state, we estimated a statistical map by regressing the fMRI time course on a linear ramp function. This choice is based on the hypothesis that activation in some regions presents a ramp-like increase during the time preceding the onset of hallucinations.

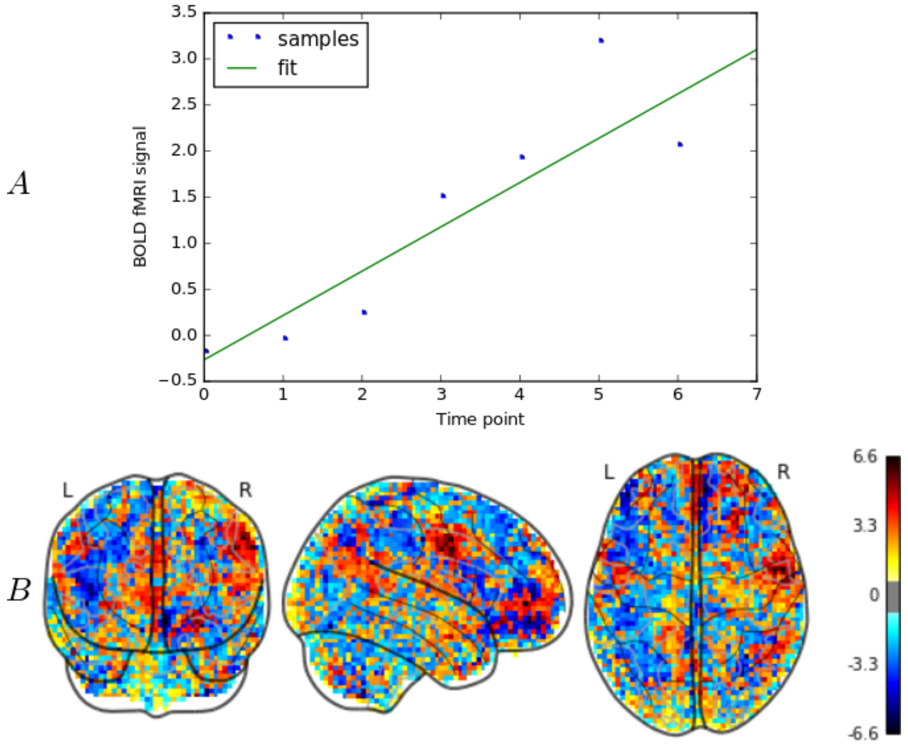


Figure 5.8: (a) Regression of the fMRI signal time course of a voxel on a linear ramp function (fit is represented in green). (b) Sample created from one set of consecutive pre-hallucinations scans. The features are the T-statistic values associated with the coefficients of the regression in each voxel

Fig. 5.8 (A) represents the evolution of the signal intensity in one single voxel over the 8 consecutive volumes of a pre-hallucination period of a subject. Fig. 5.8 (B) as an example of one sample containing 67,665 features.

Given that most of the patients hallucinated more than once during the scanning session, we had more samples than patients (376 samples created from 36 patients): 166 in the resting state (off periods) and 210 in the pre-hallucination state (trans periods).

Supervised analysis

We compared the prediction performance and interpretability of weight maps provided by two different classifiers: the linear Support Vector Machine and the TV-Enet classifier.

Performances were evaluated using a double cross-validation made of two nested cross-validation loops. In the outer (external) loop we employed a leave-one-subject-out pipeline where all subjects except one were referred to as the training data, and the remaining subject was used as test data. The test sets were exclusively used for model assessment, whereas the training sets were used in the inner 5-fold cross-validation loop for model fitting and model selection. Classifier performances were assessed by computing the balanced accuracy, sensitivity, and specificity with which the test samples were classified. Sensitivity was defined as the ability to identify the transition towards hallucination state (trans), whereas specificity evaluated the ability to identify the resting-state activity (off). The balanced accuracy score was defined as the average of the sensitivity and specificity.

Unsupervised Analysis

Subsequently, in addition to the supervised analysis, we conducted an extensive analysis of the data using unsupervised machine learning. The goal was to characterize the variability within the pre-hallucination scans. We used SPCA-TV model, described in Sec. 4.3, to produce intelligible patterns.

We hypothesized that the principal components extracted with SPCA-TV could uncover significant variability trends within the pre-hallucination samples. Thus, the principal components might reveal the existence of subgroups of hallucinations, notably according to the sensory modality involved (e.g., vision, audition). From the 376 samples, we retained the 210 elements corresponding to the pre-hallucinations samples. We applied SPCA-TV to these 210 samples and interpreted the resulting principal components.

Additionally, we computed the explained variance of each component yielded by SPCA-TV and investigated whether these components were capturing a signature of the cognitive process involved in the onset of hallucinations. To do so, we projected each activation map, “off” and “trans” samples, in the basis formed by the principal components and used the subsequent scores to decode the mental state of each subject. We used an SVM using the same cross-validation pipeline described in the supervised analysis method section.

5.2.3 Results

Supervised analysis

Classification performances Classification (Tab. 5.2) of resting state (i.e., non-hallucination) patterns (off) versus transition towards hallucinations patterns (trans) achieved above chance level decoding performances with both methods. The TV-Enet yielded a significantly increased AUC compared to SVM ($T = 2.87, p = 0.006$).

Table 5.2: The performance of the classifiers. Prediction accuracies: sensitivity (recall rate of "trans" samples), specificity (recall rate of "off" samples) and balanced accuracy (bAcc): (Sen+Spe)/2; AUC indicates area under the curve. We tested whether the scores obtained with SVM were significantly different from the scores obtained with TV-Enet. Significance notations: *: $p \leq 10^{-2}$

Model	**AUC**	**bAcc**	**Spe**	**Sen**
SVM	0.73*	0.73	0.78	0.67
TV-Enet	0.79*	0.74	0.76	0.71

Predictive weight maps When using the regular SVM classifier, the relevance of the obtained discriminating weight maps was limited (Fig. 5.9 (A): The whole-brain contributes to the prediction, and more puzzling, the complex mixing of positive (red) and negative (blue) values rule out the opportunity to understand the brain regions involved in the detection. Conversely, The TV-Enet classifier yields a more coherent weight map with two defined stable predictive clusters (Fig. 5.9.B). The regions cover Broca's area, and its the right homolog.

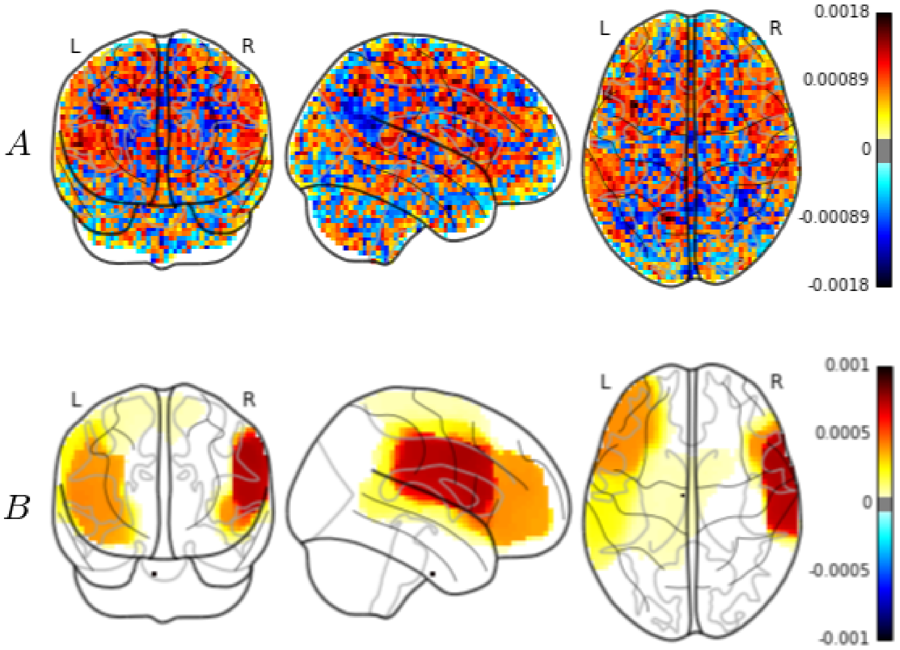


Figure 5.9: (a) Linear support vector machine (SVM) and (b) TV-Enet predictive weight map.

Unsupervised analysis

The explained variances of the four first components were: 2.5%, 1,4%, 0.09% and 0.05%. The coeficient map are provided in Fig. 5.10. The prediction of mental states based on the scores associated with each component yielded a significant decoding performance: the classifier was able to distinguish the "trans" samples from "off" samples, with an AUC of 0.65, a balanced accuracy of 65% with a sensitivity of 68% and a specificity of 64%.

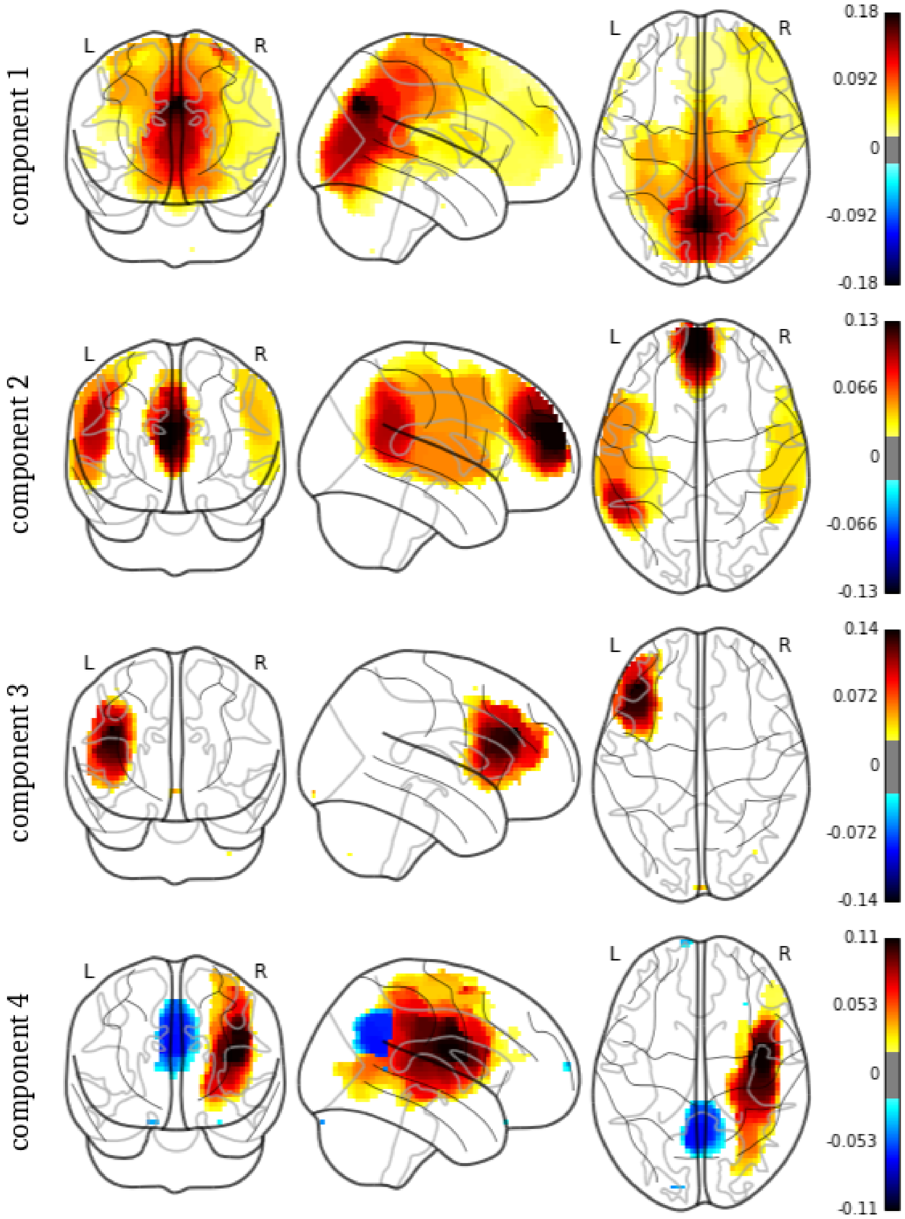


Figure 5.10: SPCA-TV principal components. Note that the sign is arbitrary

5.2.4 Discussion

Supervised analysis predictive signature

Spatial regularization provides two large, stable predictive fronto-temporal clusters that are consistent with what we currently know of the networks involved in auditory hallucinations. Indeed, numerous studies have highlighted abnormal resting-state functional connectivity among some temporo-parietal, frontal and subcortical regions in patients with auditory hallucinations (Alderson-Day, McCarthy-Jones, and Fernyhough, 2015; Allen et al., 2008). Otherwise, patients experiencing auditory hallucinations while in the MRI scanner (in so-called fMRI "capture" studies) demonstrated significantly increased activation in Broca's area, the insula, left middle and superior temporal gyrus, left inferior parietal lobule and left hippocampal region (Jardri et al., 2013).

The right cluster identified in our study also emphasized the role of the right-sided homologues of the classical speech-related areas (i.e., the right inferior frontal gyrus, right superior temporal and supramarginal gyrus) in auditory hallucinations, as previously described in the literature. It has been hypothesized that activity in these regions, especially the insula and the right homologue of Broca's area, is associated with the occurrence of auditory hallucinations (Jardri et al., 2011; Sommer et al., 2008), whereas language production in a natural context predominantly activates left-lateralized frontal and temporal language areas. The role of right-sided speech-related areas in the pathophysiology of auditory hallucinations was also mentioned by Mondino et al., 2016. By neuromodulating a speech-related fronto-parietal network, these authors demonstrated that a reduction in the resting-state functional connectivity between the left temporoparietal junction and right inferior frontal areas could be measured, and this reduction was associated with a significant reduction in the severity of the hallucinations.

Taken together, these results confirm that adding a penalty to account for the spatial structure of the brain seems relevant in fMRI captures, given that it significantly improves the classifier performance and results in clinically interpretable weight maps.

Unsupervised analysis

Relevance of components The total amount of explained variance was surprisingly low. However, when predicting the mental state of subjects based on the SPCA-TV scores, the decoding accuracy was significant, demonstrating that the component captured a functional variability specific to the cognitive processes involved in the onset of hallucinations.

Weight map interpretation The variability in the pre-hallucination patterns across occurrences and subjects were represented in the form of intelligible components shown in Fig. 5.10.

1. *The first PC* mainly included the weights in the precuneus cortex and the posterior cingulate cortex. The posterior cingulate cortex, which is part of the DMN, is associated with auditory hallucinations (Rotarska-Jagiela et al., 2010). We believe that this component may have captured the visual pathways typically involved in the occurrence of visual hallucinations.

2. *The second PC* was composed of one activation cluster in the paracingulate gyrus and the anterior cingulate gyrus and two

symmetric bilateral activation clusters in the temporal cortex. This fronto-temporal component appeared compatible with the processes at the roots of the auditory hallucinations. Interestingly, some processes involved in the occurrence of hallucinations, such as the monitoring of inner speech processes and error detection, are classical functions of the anterior cingulate cortex included in this component (Allen et al., 2008; Mechelli et al., 2007). This second PC yielded regions classically involved in inhibition (paracingulate gyrus, anterior cingulate gyrus) (Allen et al., 2008; Mechelli et al., 2007). The severity of auditory hallucinations has been found inversely related to the strength of the functional connectivity between the temporal-parietal junction, the anterior cingulate cortex (ACC), and the amygdala (Vercammen et al., 2010). This ACC dysconnectivity supposedly drove the external misattribution observed during auditory hallucinations (Allen et al., 2007; Mechelli et al., 2007), and might explain global inhibition impairments in the pathophysiology of hallucinations (Jardri et al., 2016), which may account for this feature beyond the schizophrenia-spectrum, for instance in LSD-induced hallucinations Schmidt et al., 2017.

3. *The third PC* revealed a cluster in the frontal gyrus and the anterior insula. These regions are important for speech production, encompassing the well-known Broca's area and are involved in auditory hallucinations (Jardri et al., 2011; Sommer et al., 2008).

4. *The fourth PC* included two clusters of opposing signs. In the right hemisphere, there was a large activation cluster that involved the temporo-parietal junction and a deactivation cluster that involved the precuneus cortex and the posterior cingulate gyrus. Interestingly, this PC revealed activation of the brain regions involved in auditory hallucination-related processes and in self-other distinction, such as the right temporo-parietal junction (Decety and Lamm, 2007; Jardri et al., 2013; Plaze et al., 2015), together with a deactivation of key nodes of the DMN, including the posterior cingulate cortex, medial prefrontal cortex, medial temporal cortex, and lateral parietal cortex (Buckner, Andrews-Hanna, and Schacter, 2008). Our results appeared fully compatible with recent fMRI-capture findings demonstrating that aberrant activations of speech-related areas concomitant with hallucinatory experiences follow complex interactions between ICNs, such as the DMN and the CEN (Lefebvre et al., 2016). Disengagement of the DMN during goal-directed behaviors has been seminally evidenced in the resting-state literature (Lefebvre et al., 2016; "From The Cover: The Human Brain Is Intrinsically Organized into Dynamic, Anticorrelated"), and similar mechanisms might be involved in hallucinatory occurrences

(Jardri et al., 2013; Leroy et al., 2017). Such fluctuations in the ICNs are, thus, thought to be highly involved in the transition from a resting state to an active hallucinatory state.

Perspectives

Real-time recognition of the "trans" period using the TV-Enet classifier could enable the delivery of visual information (i.e., visual feedback) regarding the imminent onset of hallucinations to the participant during a fMRI session. Such a procedure could help the subject learn effective coping strategies to prevent the occurrence of hallucinations.

One of the major limits of such fMRI-based therapies remains the accessibility and cost of the equipment. It appears fundamental to develop less complex devices as potential second-line treatments for hallucinations, such as near-infrared spectroscopy (NIRS). From this technological transfer perspective, the discriminative maps obtained using the TV-Enet classifier also appear advantageous, given that the identified clusters are cortical regions with activity that are easily measured with NIRS.

5.2.5 Conclusion

Because the hallucinations were frequently multimodal in the sample of patients recruited for this study, we expected more disparities in the functional patterns associated with their complex hallucinations and the transition towards this state compared with pure auditory experiences. In this context, the significant inter-subject decoding performances obtained appeared satisfactory and are promising for future fMRI-based therapy for drug-resistant hallucinations.

We have successfully demonstrated the interest of using structured sparse machine learning tools on a clinical dataset of fMRI-recorded pre-hallucination patterns in a population of schizophrenia patients.

5.3 Spatial Patterns of White Matter Hyperintensities in CADASIL

CADASIL (Cerebral Autosomal Dominant Arteriopathy with Subcortical Infarcts and Leukoencephalopathy) is a neurological, vascular disorder associated with white matter hyperintensities (WMH) that are considered to result from hypoperfusion. We hypothesized that the burden of WMH results from the combination of several spatial patterns of WMH associated with different mechanisms and clinical evolution. In Duchesnay et al., 2018 (summarized in Fig.5.11) the aim was to identify spatial pattern WMH variability using spatial Principle Component Analysis (PCA-TV) . PCA-TV identified two patterns explaining part of the WMH variability. The first pattern includes subcortical WMH was found to be associated with better clinical outcomes than the second pattern of WMH in pyramidal tracts that is associated with worse outcomes. Those two brain patterns of WMH associated with two different clinical evolution suggest two different pathological mechanisms.

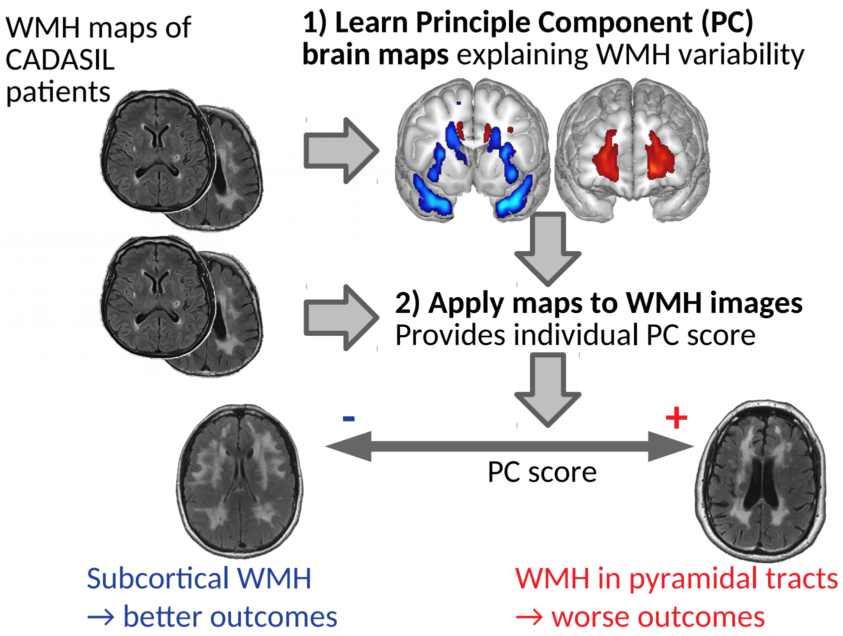

WMH maps of CADASIL patients
1) Learn Principle Component (PC) brain maps explaining WMH variability
2) Apply maps to WMH images
Provides individual PC score
-
+
PC score
Subcortical WMH
→ better outcomes
WMH in pyramidal tracts
→ worse outcomes

Figure 5.11: PCA-TV found two spatial patterns of WMH variability in CADASIL.

5.3.1 Introduction

White matter hyperintensities (WMH) are a hallmark of cerebral small vessel disease (SVD). While it is still widely considered that they result from chronic hypoperfusion, other mechanisms are likely involved (Joutel and Chabriat, 2017; Pantoni, 2010). In CADASIL (Cerebral Autosomal Dominant Arteriopathy with Subcortical Infarcts and Leukoencephalopathy), the most frequent monogenic form of SVD, WMH are commonly seen in anterior temporal poles and superior frontal gyri, which are generally spared by WMH in age- and hypertension-related SVD Auer et al., 2001. We recently showed that the WMH observed in these areas in CADASIL are characterized by far longer T1 and T2* relaxometry values than WMH observed in the remaining white matter. This large difference, tightly linked to the local water content De Guio et al., 2018, suggests that WMH in anterior temporal poles and superior frontal gyri might result from different mechanisms than WMH observed in other brain regions. In the present work, we hypothesized that the whole burden of WMH observed on conventional MRI in CADASIL results from the combination of different regional populations of WMH. We tested this hypothesis using a priori free, exploratory spatial Principle Component Analysis (PCA-TV).

5.3.2 Methods

Patients

Three hundred one patients with CADASIL Chabriat et al., 2009 with four clinical scores of global the cognitive performances that were assessed by the Mattis dementia rating scale (MDRS) and mini mental state examination (MMSE); executive functions that were assessed by the time to complete part B of Trail Making Test (TMTB); and disability that was assessed by the modified Rankin's scale (mRS).

Image processing and analysis

Masks of WMH and lacunes were semi-automatically determined. The number of microbleeds (MB_N) recorded in all subjects from FLAIR, 3D-T1, and T2* sequences, respectively, providing the volume of WMH and lacunes (WMH_V and LL_V respectively). The brain parenchymal fraction (BPF) was defined as the ratio of brain tissue volume to that of intracranial cavity volume to take into account inter-subject variability in head size. All masks of WMH were registered to the Montreal Neurological Institute (MNI) template, first with a linear registration between FLAIR and T1 images (FLIRT) and then with a nonlinear registration between T1 images and the MNI template (FNIRT) (http://www.fmrib.ox.ac.uk/fsl).

Spatially regularized PCA applied to aligned masks of WMH produced a sequence of Principal Components (PCs) explaining a decreasing proportion of the WMH variability. We chose to stop the addition of new PC as soon as the relative improvement of the total explained variance by the new PC was inferior to 5%.

Each principal component k is made of a *PC map*, i. e.the brain pattern that is multiplied with individual WMH mask to produce a *PC score*. The visualization *PC map* highlight the brain pattern that best explains the spatial variability of WMH at step k. Individuals *PC score* are used to explore associations with clinical or other phenotypes.

Thereafter, we tested the relationships between the different PC scores and: (1) MRI markers of CADASIL, namely BPF, LL_V , WMH_V, and MB_N , with systematic adjustment for age and sex; (2) cognitive scores (MDRS, MMSE, TMTB) and disability scale (mRS), with systematic adjustment for age, sex, level of education, and MRI markers in agreement with the literature in CADASIL (Chabriat et al., 2016; Jouvent et al., 2016).

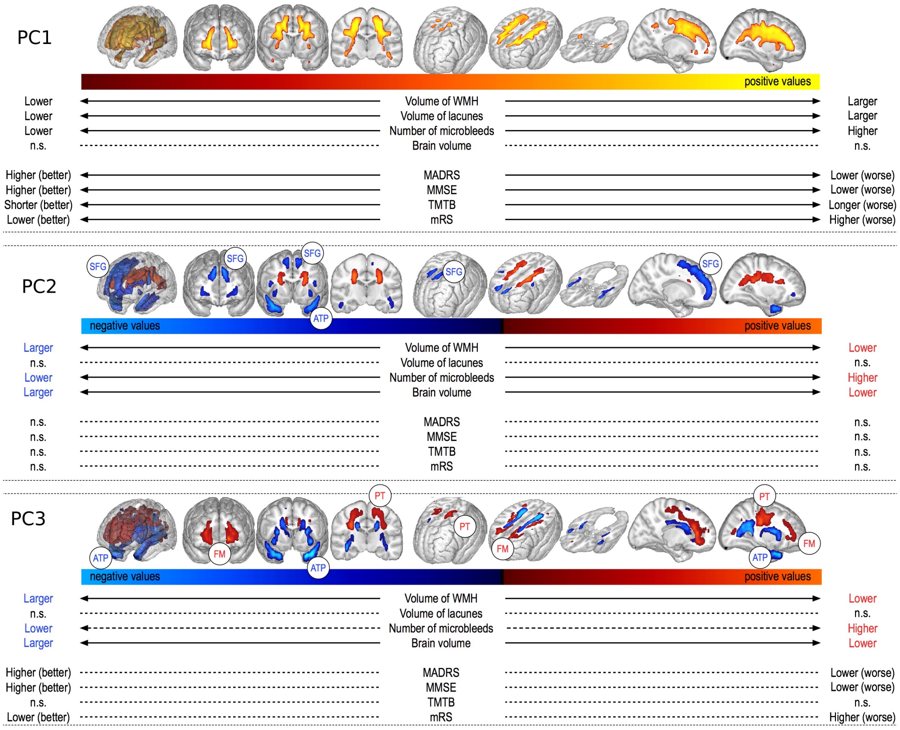

PC1
positive values
Lower Volume of WMH Larger
Lower Volume of lacunes Larger
Lower Number of microbleeds Higher
n.s. Brain volume n.s.
Higher (better) MADRS Lower (worse)
Higher (better) MMSE Lower (worse)
Shorter (better) TMTB Longer (worse)
Lower (better) mRS Higher (worse)
PC2
SFG
ATP
negative values
positive values
Larger Volume of WMH Lower
n.s. Volume of lacunes n.s.
Lower Number of microbleeds Higher
Larger Brain volume Lower
n.s. MADRS n.s.
n.s. MMSE n.s.
n.s. TMTB n.s.
n.s. mRS n.s.
PC3
PT
FM
ATP
negative values
positive values
Larger Volume of WMH Lower
n.s. Volume of lacunes n.s.
Lower Number of microbleeds Higher
Larger Brain volume Lower
Higher (better) MADRS Lower (worse)
Higher (better) MMSE Lower (worse)
n.s. TMTB n.s.
Lower (better) mRS Higher (worse)

Figure 5.12: Principal component maps and their relationships with other MRI markers and clinical scores. Each box depicts one principal component. The Top of the box shows the component map (the combination of voxels that explains a part of the variability of WMH shape) and the Bottom the corresponding component score.

5.3.3 Results

Main Sources of Variation of the Spatial Pattern of WMH

- PC1 explained 19.9% of the variability of WMH. PC1 map (Fig. 5.12) has positive coefficients that spread all over WM capturing the global extends of WMH (95% of correlation between PC1 score and WMH_v).

- PC2 explained 15% of the variability of WMH. PC2 map (Fig. 5.12) has positive coefficients in pyramidal tract (PT) is and negative coefficients in anterior temporal pole (ATP) and superior frontal gyrus (SFG).

- PC3 explained 6% of the variability of WMH, PC2 map (Fig. 5.12) has positive coefficients in PT and forceps minor (FM), and negative coefficients in ATP. Therefore, patients with WMH in PT or/and FM will have PC3 a large positive, while patients with WMH in ATP will result in a low negative PC3 score.

Relationships between PC scores and other MRI markers and clinical scores

- In addition to their strong correlation with WMH_V, PC1 score was positively and significantly associated with WMH_V, LL_V, MB_N but not to BPF. PC1 score was positively and significantly associated with clinical worsening of all scores MADRS, MMSe, TMTB, mRS.

- PC2 score was positively and significantly associated with WMH_V, MB_N ad BPF. PC2 score was not found to significantly associated with any clinical score.

- PC3 score was positively and significantly associated with WMH_V, MB_N ad BPF. PC2 score was positively and significantly associated with clinical worsening of MADRS, MMSe, and mRS.

5.3.4 Conclusion

The results of the present study support the hypothesis that the whole burden of WMH in CADASIL is, in fact, the combination of different regional populations of WMH, with different mechanisms and clinical consequences.

This analysis confirmed different and sometimes inverse relationships between the local extent of WMH and clinical severity. PC3 shows that subcortical WMH (anterior temporal poles and superior frontal gyri) are associated with milder forms of the disease, while larger volumes of WMH in pyramidal tracts or in the forceps minor are associated with more severe forms.

6 Conclusion and perspectives

6.1 Feedback on a fifteen years journey of designing machine learning models for neuroimaging

High-dimensional ("large P small N") data, such as neuroimaging or OMICs, disrupts theory and our common understanding of data analysis. Indeed our representation of clusters of points in space is misleading since all points lie on the hull containing them. Thus, whereas no one would have foreseen it 15 years ago, a consensus emerged in the community that linear approaches were the most effective.

We found that simple ℓ_2-regularization has a broad stability range and provides satisfying baseline performance on most datasets. Some improvements may be obtained with univariate feature selection or sparse ℓ_1-regularized models. However, those models' outcome is limited to understand the brain's patterns that underpin the prediction. We demonstrated, with several clinical applications, that TV spatial regularization provides a qualitative breakthrough in terms of support recovery of the predictive brain regions.

However, besides the ℓ_2 regularization, the model selection issue undermines the potentiality of more sophisticated models. With a large sample size (a few hundred), thanks to low variability, cross-validated grid-search can efficiently be used. On a smaller sample size (< 100), the model selection remains an open problem. We found that in-sample bounds were promising axis to explore. Nevertheless, to be practically useful sophisticated models should be provided with a "ready-to-use" reduced set possible parameters' values for a given data type. This practical consideration will reduce the variability of possible solutions with a likely improvement of generalization. More importantly, this will promote "good science" with sophisticated models, preventing the overfitting risk of cherry-picking the best model among many tested ones. We have to do this work concerning our TV-based models. When it comes to application to real data, providing instructions on how to use a sophisticated model is as essential as the theoretical contribution behind it.

6.2 Perspectives: leveraging psychiatric neuroimaging biomarkers discovery

6.2.1 Toward precision psychiatry: datasets and strategies

Unlike many other medical specialties, psychiatry lacks objective quantitative measures (such as blood dosage) to guide clinicians in choosing a therapeutic strategy. Brain anatomy is an imprint of the individual's genetic and environmental background. The identification of prognostic brain signatures of clinical course or response to treatment would pave the way for personalized medicine in psychiatry.

Disorder-specific retrospective and heterogeneous cohorts

Many international initiatives (schizconnect, abide, enigma) opened the doors to such perspective by retrospectively aggregating existing cohorts for each specific disorders. However, the considerable heterogeneity and cross-sectional designs of such datasets limit the scope of clinical relevance to basic case/control prediction (Arbabshirani et al., 2016; Rashid and Calhoun, 2020).

Large transnosographic heterogeneous cohorts

Another strategy has recently emerged on the constitution of very large ($N \geq 10,000$) transnosographic cohorts (UK Biobank, HBN). Those initiatives opened the way to new strategies: (i) investigation of general variability; (ii) dimensional exploration across clinical categories according to the RDoC (Cuthbert and Insel, 2013) paradigm. However, it is becoming clear that these cohorts have limited potential to investigate critical clinical problems. For example, HBN does not contain a sufficient number of subjects nor the required clinical assessments to learn a predictor of psychotic transition in at-risk subjects.

Disorder-specific longitudinal and homogeneous cohorts

It is becoming strategic to gather new cohorts with improved clinical homogeneity, including longitudinal follow-ups, to assess response to treatment and transition to disease in patients at risk. In this aim, I developed a network of clinicians bringing NeuroSPin to join two of the main European projects:

1. PsyCARE, RHU, 2019-2024. Preventing psychosis through personalized care. PI: MO Krebs, WP leader: E Duchesnay, Team budget: 715k€.

2. R-LiNK, H2020, 2018-2023. Optimizing response to Li treatment through personalized evaluation of individuals with bipolar I

disorder: the R-LiNK initiative. PI: F. Bellivier, WP leader: E Duchesnay and leader for the CEA, Team budget: 800k€.

Note that the CATI neuroimaging infrastructure (funded by NeuroSPin and ICM) is a key asset to build such large multi-site cohorts. Indeed, the CATI provides support for multi-center neuroimaging studies (i.e., harmonization of MRI data acquisition, monitoring and quality control, pre-processing, etc.). Such studies typically investigate from a few hundred too few thousands deeply phenotyped participants.

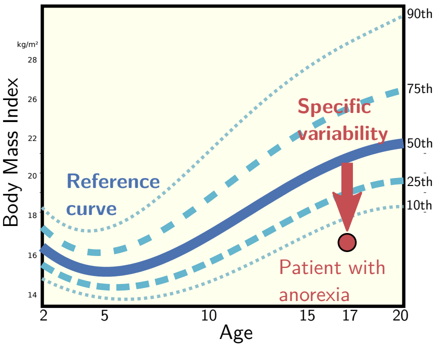
The high cost per patient (>10K€), however, limits the feasibility to scale such cohorts up to the sample size ($\approx 10,000$) necessary to build predictive models that are sufficiently reproducible for regular clinical application.

Indeed, the anatomo-functional variability is dominated by a natural interindividual variability that stems from demographical or possibly other unobserved (latent) variables such as age, sex, education level, toxic/alcohol consumption, etc. The information specific to the disorder is mixed up or hidden with non-specific information. For example, response to an antidepressant (of patients with unipolar depression) is associated with strategic atrophies related to the history of depression of a patient (Chen et al., 2007), which could be used to identify a prognostic brain signature to antidepressants. Unfortunately, this signature is overlooked by numerous other factors (age, alcohol) that are known to modify the brain but with a different spatial pattern.

6.2.2 Perspectives: Transfer Learning from Big data to Small Data

Therefore, a future methodological challenge is to bridge the gap between large heterogeneous cohorts and more relevant, but smaller, longitudinal, and homogeneous cohorts. This goal can be addressed using transfer learning strategies described in Fig. 6.1.

The idea, related to normative models proposed by Marquand et al., 2016, is illustrated in the enclosed figure: Let us suppose that we want to identify individuals with anorexia nervosa. The reference curve of Body Mass Index (BMI) as a function of the age will leverage a supervised algorithm to "dig" in a pathology-specific direction (red arrow in the figure) by overcoming the general variability. The reference curve can learned by an "universal"



Encoder E_u (Fig. 6.1, top) that maps subjects to a reduced latent space z_u.

We propose to pre-train the "universal" Encoder E_u of the general brain variability on *large transnosographic heterogeneous cohorts*, see top of Fig. 6.1.

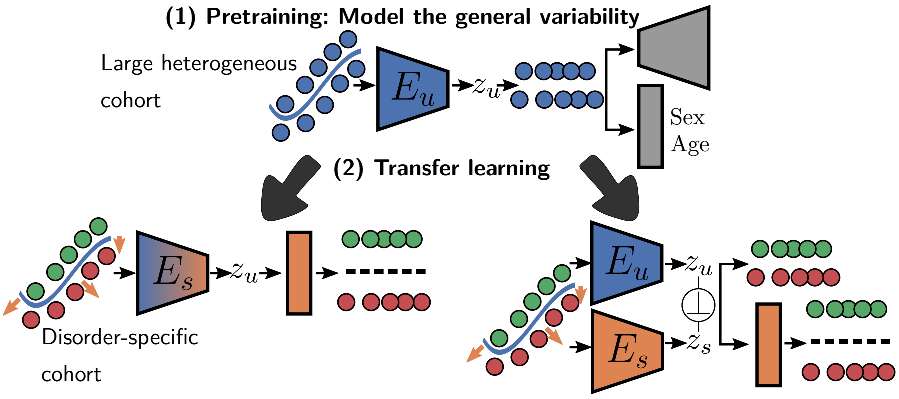


Figure 6.1: Transfer Learning strategies: from large trans-diagnostic heterogeneous cohorts to small longitudinal cohorts.

E_u (typically convolutional neural network) will be trained using a combination of a supervised multitask learning (of orthogonal targets, such as sex and age) and an additional decoder that will minimize a reconstruction loss. E_u will be transfered to *disorder-specific cohorts* to learn a second specific encoder E_s to re-focus on the disorder-related variability. A first solution (Fig. 6.1, bottom left), uses the ImageNet-like pre-training paradigm, to fine-tune E_u to get E_s. In a second approach (Fig. 6.1, bottom right), a "frozen" E_u will force a second specific encoder E_s to focus its latent space z_s toward disorder-specific variability (e.g., Zheng and Sun, 2019).

Bibliography

Abraham, Alexandre, Elvis Dohmatob, Bertrand Thirion, Dimitri Samaras, and Gael Varoquaux (2013). “Extracting Brain Regions from Rest fMRI with Total-Variation Constrained Dictionary Learning.” In: *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2013 - 16th International Conference, Nagoya, Japan, 2013, Proceedings, Part II*, pp. 607–615.

Adriano, Fulvia, Ilaria Spoletini, Carlo Caltagirone, and Gianfranco Spalletta (2010). “Updated Meta-Analyses Reveal Thalamus Volume Reduction in Patients with First-Episode and Chronic Schizophrenia.” In: *Schizophrenia research* 123.1, pp. 1–14.

Akaike, H. (Dec. 1974). “A New Look at the Statistical Model Identification.” In: *IEEE Transactions on Automatic Control* 19.6, pp. 716–723.

Alderson-Day, Ben, Kelly Diederen, Charles Fernyhough, Judith M Ford, Guillermo Horga, Daniel S Margulies, Simon McCarthy-Jones, Georg Northoff, James M Shine, Jessica Turner, et al. (2016). “Auditory Hallucinations and the Brain’s Resting-State Networks: Findings and Methodological Observations.” In: *Schizophrenia bulletin* 42.5, pp. 1110–1123.

Alderson-Day, Ben, Simon McCarthy-Jones, and Charles Fernyhough (2015). “Hearing Voices in the Resting Brain: A Review of Intrinsic Functional Connectivity Research on Auditory Verbal Hallucinations.” In: *Neuroscience & Biobehavioral Reviews* 55, pp. 78–87.

Allen, Paul, Edson Amaro, Cynthia HY Fu, Steven CR Williams, Michael J Brammer, Louise C Johns, and PHILIP K McGUIRE (2007). “Neural Correlates of the Misattribution of Speech in Schizophrenia.” In: *The British Journal of Psychiatry* 190.2, pp. 162–169.

Allen, Paul, Frank Larøi, Philip K McGuire, and Andrè Aleman (2008). “The Hallucinating Brain: A Review of Structural and Functional Neuroimaging Studies of Hallucinations.” In: *Neuroscience & Biobehavioral Reviews* 32.1, pp. 175–191.

Arbabshirani, Mohammad R., Sergey Plis, Jing Sui, and Vince D. Calhoun (Mar. 21, 2016). “Single Subject Prediction of Brain Disorders in Neuroimaging: Promises and Pitfalls.” In: *NeuroImage*.

Arns, Martijn, J-M Batail, Stéphanie Bioulac, Marco Congedo, Christophe Daudet, Dominique Drapier, Thomas Fovet, Renaud Jardri, M Le-Van-Quyen, Fabien Lotte, et al. (2017). “Neurofeedback: One of Today’s Techniques in Psychiatry?” In: *L’Encéphale* 43.2, pp. 135–145.

Ashburner, John (Oct. 15, 2007). “A Fast Diffeomorphic Image Registration Algorithm.” In: *NeuroImage* 38.1, pp. 95–113.

Ashburner, John and Karl J. Friston (July 1, 2005). "Unified Segmentation." In: *NeuroImage* 26.3, pp. 839–851.

Auer, D. P., B. Pütz, C. Gössl, G. Elbel, T. Gasser, and M. Dichgans (Feb. 2001). "Differential Lesion Patterns in CADASIL and Sporadic Subcortical Arteriosclerotic Encephalopathy: MR Imaging Study with Statistical Parametric Group Comparison." In: *Radiology* 218.2, pp. 443–451.

Bach, Francis, Rodolphe Jenatton, Julien Mairal, and Guillaume Obozinski (2011). "Convex Optimization with Sparsity-Inducing Norms." In: *Optimization for Machine Learning*. Ed. by S. Sra, S. Nowozin, and S. J. Wright. MIT Press.

Beck, A. and M. Teboulle (Jan. 1, 2009a). "A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems." In: *SIAM Journal on Imaging Sciences* 2.1, pp. 183–202.

Beck, Amir and Marc Teboulle (2009b). "Fast Gradient-Based Algorithms for Constrained Total Variation Image Denoising and Deblurring Problems." In: *IEEE Transactions on Image Processing* 18.11, pp. 2419–2434.

— (2012). "Smoothing and First Order Methods: A Unified Framework." In: *SIAM Journal on Optimization* 22.2, pp. 557–580.

Benjamini, Yoav and Yosef Hochberg (1995). "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." In: *Journal of the Royal Statistical Society. Series B (Methodological)* 57.1, pp. 289–300.

Birgé, Lucien and Pascal Massart (May 1, 2007). "Minimal Penalties for Gaussian Model Selection." In: *Probability Theory and Related Fields* 138.1, pp. 33–73.

Bohlken, Marc, K Hugdahl, and Iris Sommer (2017). "Auditory Verbal Hallucinations: Neuroimaging and Treatment." In: *Psychological medicine* 47.2, pp. 199–208.

Borwein, Jonathan. and Adrian Lewis (2006). *Convex Analysis and Nonlinear Optimization: Theory and Examples*. CMS Books in Mathematics. Springer.

Boyd, Stephen, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein (Jan. 2011). "Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers." In: *Foundations and Trends in Machine Learning* 3.1, pp. 1–122.

Boyd, Stephen and Lieven Vandenberghe (2004). *Convex Optimization*. New York, NY, USA: Cambridge University Press.

Braak, Heiko and Eva Braak (1991). "Neuropathological Stageing of Alzheimer-Related Changes." In: *Acta Neuropathologica* 82.4, pp. 239–259.

Breiman, Leo (Oct. 1, 2001). "Random Forests." In: *Machine Learning* 45.1, pp. 5–32.

Buckner, Randy L, Jessica R Andrews-Hanna, and Daniel L Schacter (2008). "The Brain's Default Network." In: *Annals of the New York Academy of Sciences* 1124.1, pp. 1–38.

Burnham, Kenneth P. and David R. Anderson (1998). *Model Selection and Inference: A Practical Information-Theoretic Approach*. New York: Springer-Verlag.

Calhoun, V.D., J. Liu, and T. Adali (2009). "A Review of Group ICA for fMRI Data and ICA for Joint Inference of Imaging, Genetic, and ERP Data." In: *NeuroImage* 45 (1(Suppl 1)), S163–S172.

Cardenas, Valerie, LL Chao, Colin Studholme, Kristin Yaffe, Bruce Miller, Cindee Madison, Shannon Buckley, Dan Mungas, Norbert Schuff, and Michael Weiner (2011). "Brain Atrophy Associated with Baseline and Longitudinal Measures of Cognition." In: *Neurobiology of aging* 32.4, pp. 572–580.

Chabriat, Hugues, Anne Joutel, Martin Dichgans, Elizabeth Tournier-Lasserve, and Marie-Germaine Bousser (July 2009). "Cadasil." In: *The Lancet. Neurology* 8.7, pp. 643–653.

Chabriat, Hugues et al. (Jan. 2016). "Predictors of Clinical Worsening in Cerebral Autosomal Dominant Arteriopathy With Subcortical Infarcts and Leukoencephalopathy: Prospective Cohort Study." In: *Stroke* 47.1, pp. 4–11.

Chambolle, A. and T. Pock (2011). "A First-Order Primal-Dual Algorithm for Convex Problems with Applications to Imaging." In: *Journal of Mathematical Imaging and Vision* 40.1, pp. 120–145.

Chen, Chi-Hua, Khanum Ridler, John Suckling, Steve Williams, Cynthia H. Y. Fu, Emilio Merlo-Pich, and Ed Bullmore (Sept. 1, 2007). "Brain Imaging Correlates of Depressive Symptom Severity and Predictors of Symptom Improvement after Antidepressant Treatment." In: *Biological Psychiatry* 62.5, pp. 407–414.

Chen, Xi, Qihang Lin, Seyoung Kim, Jaime G. Carbonell, and Eric P. Xing (June 2012). "Smoothing Proximal Gradient Method for General Structured Sparse Regression." In: *The Annals of Applied Statistics* 6.2, pp. 719–752.

Chu, Carlton, Ai-Ling Hsu, Kun-Hsien Chou, Peter Bandettini, and ChingPo Lin (Mar. 2012). "Does Feature Selection Improve Classification Accuracy? Impact of Sample Size and Feature Selection on Classification Using Anatomical Magnetic Resonance Images." In: *NeuroImage* 60.1, pp. 59–70.

Chun, H. and S. Keleş (2010). "Sparse Partial Least Squares Regression for Simultaneous Dimension Reduction and Variable Selection." In: *Journal of the Royal Statistical Society - Serie B* 72 (1), pp. 3–25.

Cuingnet, R., J. A. Glaunès, M. Chupin, H. Benali, and O. Colliot (Mar. 2013). "Spatial and Anatomical Regularization of SVM: A General Framework for Neuroimaging Data." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.3, pp. 682–696.

Cuingnet, Remi, Emilie Gerardin, Jerôme Tessieras, Guillaume Auzias, Stephane Lehericy, Marie-Odile Habert, Marie Chupin, Habib Benali, Olivier Colliot, and Alzheimer's Disease Neuroimaging Initiative (May 2011). "Automatic Classification of Patients with Alzheimer's Disease from Structural MRI: A Comparison of Ten Methods Using the ADNI Database." In: *NeuroImage* 56.2, pp. 766–781.

Cuthbert, Bruce N. and Thomas R. Insel (May 14, 2013). "Toward the Future of Psychiatric Diagnosis: The Seven Pillars of RDoC." In: *BMC medicine* 11, p. 126.

D'Aspremont, Alexandre, Laurent El Ghaoui, Michael I. Jordan, and Gert R. G. Lanckriet (Jan. 2007). "A Direct Formulation for Sparse PCA Using Semidefinite Programming." In: *SIAM Review* 49.3, pp. 434–448.

De Guio, François, Alexandre Vignaud, Hugues Chabriat, and Eric Jouvent (Sept. 2018). "Different Types of White Matter Hyperintensities in CADASIL: Insights from 7-Tesla MRI." In: *Journal of Cerebral Blood Flow and Metabolism: Official Journal of the International Society of Cerebral Blood Flow and Metabolism* 38.9, pp. 1654–1663.

De Pierrefeu, A. et al. (2018a). "Identifying a Neuroanatomical Signature of Schizophrenia, Reproducible across Sites and Stages, Using Machine Learning with Structured Sparsity." In: *Acta Psychiatrica Scandinavica* 0.0.

De Pierrefeu, Amicie, Thomas Fovet, Fouad Hadj-Selem, Tommy Löfstedt, Philippe Ciuciu, Stephanie Lefebvre, Pierre Thomas, Renaud Lopes, Renaud Jardri, and Edouard Duchesnay (Apr. 1, 2018b). "Prediction of Activation Patterns Preceding Hallucinations in Patients with Schizophrenia Using Machine Learning with Structured Sparsity." In: *Human Brain Mapping* 39.4, pp. 1777–1788.

De Pierrefeu, Amicie, Tommy Lofstedt, Fouad Hadj-Selem, Mathieu Dubois, Renaud Jardri, Thomas Fovet, Philippe Ciuciu, Vincent Frouin, and Edouard Duchesnay (Feb. 2018c). "Structured Sparse Principal Components Analysis With the TV-Elastic Net Penalty." In: *IEEE Transactions on Medical Imaging* 37.2, pp. 396–407.

Decety, Jean and Claus Lamm (2007). "The Role of the Right Temporoparietal Junction in Social Interaction: How Low-Level Computational Processes Contribute to Meta-Cognition." In: *The Neuroscientist* 13.6, pp. 580–593.

Delacourte, Andre, Jean-Philippe David, Nicolas Sergeant, L Buee, A Wattez, P Vermersch, F Ghozali, C Fallet-Bianco, F Pasquier, F Lebert, et al. (1999). "The Biochemical Pathway of Neurofibrillary Degeneration in Aging and Alzheimer's Disease." In: *Neurology* 52.6, pp. 1158–1158.

Desikan, Rahul S. et al. (July 1, 2006). “An Automated Labeling System for Subdividing the Human Cerebral Cortex on MRI Scans into Gyral Based Regions of Interest.” In: *NeuroImage* 31.3, pp. 968–980.

Dice, Lee (1945). “Measures of the Amount of Ecologic Association between Species.” In: *Ecology* 26, pp. 297–302.

Diederen, Kelly MJ, Sebastiaan FW Neggers, Kirstin Daalman, Jan Dirk Blom, Rutger Goekoop, René S Kahn, and Iris EC Sommer (2010). “Deactivation of the Parahippocampal Gyrus Preceding Auditory Hallucinations in Schizophrenia.” In: *American Journal of Psychiatry* 167.4, pp. 427–435.

Dohmatob, Elvis, Michael Eickenberg, Bertrand Thirion, and Gaël Varoquaux (June 10, 2015). “Speeding-up Model-Selection in Graph-Net via Early-Stopping and Univariate Feature-Screening.” In: PRNI.

Dohmatob, Elvis, Alexandre Gramfort, Bertrand Thirion, and Gaël Varoquaux (June 3, 2014). “Benchmarking Solvers for TV-L1 Least-Squares and Logistic Regression in Brain Imaging.” In: *Pattern Recognition in Neuroimaging (PRNI).*

Draganski, Bogdan, Christian Gaser, Volker Busch, Gerhard Schuierer, Ulrich Bogdahn, and Arne May (Jan. 22, 2004). “Neuroplasticity: Changes in Grey Matter Induced by Training.” In: *Nature* 427.6972, pp. 311–312.

Dubois, M., F. Hadj-Selem, T. Löfstedt, M. Perrot, C. Fischer, V. Frouin, and E. Duchesnay (June 2014). “Predictive Support Recovery with TV-Elastic Net Penalty and Logistic Regression: An Application to Structural MRI.” In: *2014 International Workshop on Pattern Recognition in Neuroimaging*. 2014 International Workshop on Pattern Recognition in Neuroimaging. IEEE, pp. 1–4.

Duchesnay, Edouard, Arnaud Cachia, Nathalie Boddaert, Nadia Chabane, Jean-Franois Mangin, Jean-Luc Martinot, Francis Brunelle, and Monica Zilbovicius (Aug. 1, 2011). “Feature Selection and Classification of Imbalanced Datasets: Application to PET Images of Children with Autistic Spectrum Disorders.” In: *NeuroImage*. Special Issue: Educational Neuroscience 57.3, pp. 1003–1014.

Duchesnay, Edouard, Arnaud Cachia, Alexis Roche, Denis Rivière, Yann Cointepas, Dimitri Papadopoulos-Orfanos, Monica Zilbovicius, Jean-Luc Martinot, Jean Régis, and Jean-François Mangin (Apr. 2007). “Classification Based on Cortical Folding Patterns.” In: *IEEE transactions on medical imaging* 26.4, pp. 553–565.

Duchesnay, Edouard et al. (2018). “Different Types of White Matter Hyperintensities in CADASIL.” In: *Frontiers in Neurology* 9.

Dudoit, Sandrine, Juliet Popper Shaffer, and Jennifer C. Boldrick (Feb. 2003). “Multiple Hypothesis Testing in Microarray Experiments.” In: *Statistical Science* 18.1, pp. 71–103.

Eavani, Harini, Theodore Satterthwaite, Roman Filipovych, Raquel Gur, and Christos Davatzikos (2015). "Identifying Sparse Connectivity Patterns in the Brain Using Resting-State fMRI." In: *Neuroimage* 105, pp. 286–299.

Ecker, Christine, Vanessa Rocha-Rego, Patrick Johnston, Janaina Mourao-Miranda, Andre Marquand, Eileen M. Daly, Michael J. Brammer, Clodagh Murphy, Declan G. Murphy, and MRC AIMS Consortium (Jan. 1, 2010). "Investigating the Predictive Value of Whole-Brain Structural MR Scans in Autism: A Pattern Classification Approach." In: *NeuroImage* 49.1, pp. 44–56.

Farokhian, Farnaz, Iman Beheshti, Daichi Sone, and Hiroshi Matsuda (Aug. 24, 2017). "Comparing CAT12 and VBM8 for Detecting Brain Morphological Abnormalities in Temporal Lobe Epilepsy." In: *Frontiers in Neurology* 8.

Felleman, Daniel and David Van Essen (1991). "Distributed Hierarchical Processing in the Primate Cerebral Cortex." In: *Cerebral cortex (New York, NY: 1991)* 1.1, pp. 1–47.

Fellhauer, Iven, Frank G. Zöllner, Johannes Schröder, Christina Degen, Li Kong, Marco Essig, Philipp A. Thomann, and Lothar R. Schad (Sept. 30, 2015). "Comparison of Automated Brain Segmentation Using a Brain Phantom and Patients with Early Alzheimer's Dementia or Mild Cognitive Impairment." In: *Psychiatry Research* 233.3, pp. 299–305.

Fornito, Alex, Murat Yücel, Jatinder Patti, Stephen Wood, and Christos Pantelis (2009). "Mapping Grey Matter Reductions in Schizophrenia: An Anatomical Likelihood Estimation Analysis of Voxel-Based Morphometry Studies." In: *Schizophrenia research* 108.1, pp. 104–113.

Fovet, Thomas, Renaud Jardri, and David Linden (2015). "Current Issues in the Use of fMRI-Based Neurofeedback to Relieve Psychiatric Symptoms." In: *Current pharmaceutical design* 21.23, pp. 3384–3394.

Fovet, Thomas, Natasza Orlov, Miriam Dyck, Paul Allen, Klaus Mathiak, and Renaud Jardri (2016). "Translating Neurocognitive Models of Auditory-Verbal Hallucinations into Therapy: Using Real-Time fMRI-Neurofeedback to Treat Voices." In: *Frontiers in psychiatry* 7, p. 103.

Friedman, Milton (Mar. 1940). "A Comparison of Alternative Tests of Significance for the Problem of m Rankings." In: *The Annals of Mathematical Statistics* 11.1, pp. 86–92.

Frisoni, Giovanni B., Nick C. Fox, Clifford R. Jack, Philip Scheltens, and Paul M. Thompson (Feb. 2010). "The Clinical Use of Structural MRI in Alzheimer Disease." In: *Nature reviews. Neurology* 6.2, pp. 67–77.

Friston, K. J., A. P. Holmes, K. J. Worsley, J.-P. Poline, C. D. Frith, and R. S. J. Frackowiak (1994). "Statistical Parametric Maps in Functional Imaging: A General Linear Approach." In: *Human Brain Mapping* 2.4, pp. 189–210.

Furlanello, C., M. Serafini, S. Merler, and G. Jurman (2003). "An Accelerated Procedure for Recursive Feature Ranking on Microarray Data." In: *Neural Networks* 16, pp. 641–648.

Gaonkar, Bilwaj and Christos Davatzikos (Sept. 1, 2013). "Analytic Estimation of Statistical Significance Maps for Support Vector Machine Based Multi-Variate Image Analysis and Classification." In: *NeuroImage* 78, pp. 270–283.

Giraud, Christophe (Dec. 17, 2014). *Introduction to High-Dimensional Statistics*. 1 edition. Boca Raton: Chapman and Hall/CRC. 270 pp.

Glahn, D. C., P. M. Thompson, and J. Blangero (2007). "Neuroimaging Endophenotypes : Strategies for Finding Genes Influencing Brain Structure and Function." In: *Human Brain Mapping* 28, pp. 488–501.

Glahn, David C, Angela R Laird, Ian Ellison-Wright, Sarah M Thelen, Jennifer L Robinson, Jack L Lancaster, Edward Bullmore, and Peter T Fox (2008). "Meta-Analysis of Gray Matter Anomalies in Schizophrenia: Application of Anatomic Likelihood Estimation and Network Analysis." In: *Biological psychiatry* 64.9, pp. 774–781.

Goodkind, Madeleine et al. (Apr. 2015). "Identification of a Common Neurobiological Substrate for Mental Illness." In: *JAMA psychiatry* 72.4, pp. 305–315.

Gramfort, A., B. Thirion, and G. Varoquaux (June 2013). "Identifying Predictive Regions from fMRI with TV-L1 Prior." In: *2013 International Workshop on Pattern Recognition in Neuroimaging (PRNI)*. 2013 International Workshop on Pattern Recognition in Neuroimaging (PRNI). IEEE, pp. 17–20.

Grosenick, Logan, Brad Klingenberg, Kiefer Katovich, Brian Knutson, and Jonathan E. Taylor (May 15, 2013). "Interpretable Whole-Brain Prediction Analysis with GraphNet." In: *NeuroImage* 72, pp. 304–321.

Guo, Ruixin, Mihye Ahn, Hongtu Zhu, and the Alzheimer's Disease Neuroimaging Initiative (2015). "Spatially Weighted Principal Component Analysis for Imaging Classification." In: *Journal of Computational and Graphical Statistics* 24, pp. 274–296.

Guyon, Isabelle and André Elisseeff (2003). "An Introduction to Variable and Feature Selection." In: *Journal of Machine Learning Research* 3 (Mar), pp. 1157–1182.

Guyon, Isabelle, Steve Gunn, Asa Ben-Hur, and Gideon Dror (2005). "Result Analysis of the NIPS 2003 Feature Selection Challenge." In: *Advances in Neural Information Processing Systems 17*. Ed. by L. K. Saul, Y. Weiss, and L. Bottou. MIT Press, pp. 545–552.

Guyon, Isabelle, Steve Gunn, Masoud Nikravesh, and Lofti A. Zadeh (July 20, 2006). *Feature Extraction: Foundations And Applications.*

Har/Cdr. Berlin ; New York: Springer-Verlag Berlin and Heidelberg GmbH & Co. K. 778 pp.

Guyon, Isabelle, Jason Weston, Stephen Barnhill, and Vladimir Vapnik (Jan. 2002). “Gene Selection for Cancer Classification Using Support Vector Machines.” In: *Machine Learning* 46.1-3, pp. 389–422.

Hadj-Selem, Fouad, Tommy Löfstedt, Elvis Dohmatob, Vincent Frouin, Mathieu Dubois, Vincent Guillemot, and Edouard Duchesnay (Nov. 2018). “Continuation of Nesterov’s Smoothing for Regression With Structured Sparsity in High-Dimensional Neuroimaging.” In: *IEEE Transactions on Medical Imaging* 37.11, pp. 2403–2413.

Hand, David J. and C. C. Taylor (May 1, 1987). *Multivariate Analysis of Variance and Repeated Measures: A Practical Approach for Behavioural Scientists*. London ; New York: Chapman and Hall/CRC. 304 pp.

Haufe, Stefan, Frank Meinecke, Kai Görgen, Sven Dähne, John-Dylan Haynes, Benjamin Blankertz, and Felix Bießmann (2014). “On the Interpretation of Weight Vectors of Linear Models in Multivariate Neuroimaging.” In: *Neuroimage* 87, pp. 96–110.

Haxby, James V, Andrew C Connolly, and J Swaroop Guntupalli (2014). “Decoding Neural Representational Spaces Using Multivariate Pattern Analysis.” In: *Annual review of neuroscience* 37, pp. 435–456.

Hibar, D.P. et al. (2011). “Voxelwise Gene-Wide Association Study (vGeneWAS): Multivariate Gene-Based Association Testing in 731 Elderly Subjects.” In: *NeuroImage* 56, pp. 1875–1891.

Hoffman, Ralph E, Adam W Anderson, Maxine Varanko, John C Gore, and Michelle Hampson (2008). “Time Course of Regional Brain Activation Associated with Onset of Auditory/Verbal Hallucinations.” In: *The British Journal of Psychiatry* 193.5, pp. 424–425.

Honea, Robyn, Tim J. Crow, Dick Passingham, and Clare E. Mackay (Dec. 2005). “Regional Deficits in Brain Volume in Schizophrenia: A Meta-Analysis of Voxel-Based Morphometry Studies.” In: *The American Journal of Psychiatry* 162.12, pp. 2233–2245.

Hotelling, H. (1936). “Relations between Two Sets of Variates.” In: *Biometrika* 28, pp. 321–377.

Jack, Clifford, Maria Shiung, Jeffrey Gunter, PC O’brien, SD Weigand, David S Knopman, Bradley F Boeve, Robert J Ivnik, Glenn E Smith, RH Cha, et al. (2004). “Comparison of Different MRI Brain Atrophy Rate Measures with Clinical Disease Progression in AD.” In: *Neurology* 62.4, pp. 591–600.

Jardri, Renaud, Kenneth Hugdahl, Matthew Hughes, Jérôme Brunelin, Flavie Waters, Ben Alderson-Day, Dave Smailes, Philipp Sterzer, Philip R Corlett, Pantelis Leptourgos, et al. (2016). “Are Hallucinations Due to an Imbalance between Excitatory and Inhibitory

Influences on the Brain?" In: *Schizophrenia bulletin* 42.5, pp. 1124–1134.

Jardri, Renaud, Alexandre Pouchet, Delphine Pins, and Pierre Thomas (2011). "Cortical Activations during Auditory Verbal Hallucinations in Schizophrenia: A Coordinate-Based Meta-Analysis." In: *American Journal of Psychiatry* 168.1, pp. 73–81.

Jardri, Renaud, Pierre Thomas, Christine Delmaire, Pierre Delion, and Delphine Pins (2013). "The Neurodynamic Organization of Modality-Dependent Hallucinations." In: *Cerebral Cortex* 23.5, pp. 1108–1117.

Jenatton, Rodolphe, Guillaume Obozinski, and Francis Bach (2010). "Structured Sparse Principal Component Analysis." In: *International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Jolliffe, Ian, Nickolay Trendafilov, and Mudassir Uddin (2003). "A Modified Principal Component Technique Based on the LASSO." In: *Journal of Computational and Graphical Statistics* 12.3, pp. 531–547.

Journée, Michel, Yurii. Nesterov, Peter Richtárik, and Rodolphe Sepulchre (2010). "Generalized Power Method for Sparse Principal Component Analysis." In: *Journal of Machine Learning Research* 11, pp. 517–553.

Joutel, Anne and Hugues Chabriat (Apr. 25, 2017). "Pathogenesis of White Matter Changes in Cerebral Small Vessel Diseases: Beyond Vessel-Intrinsic Mechanisms." In: *Clinical Science (London, England: 1979)* 131.8, pp. 635–651.

Jouvent, Eric, Edouard Duchesnay, Foued Hadj-Selem, François De Guio, Jean-François Mangin, Dominique Hervé, Marco Duering, Stefan Ropele, Reinhold Schmidt, Martin Dichgans, et al. (2016). "Prediction of 3-Year Clinical Course in CADASIL." In: *Neurology* 87.17, pp. 1787–1795.

Kambeitz, Joseph, Lana Kambeitz-Ilankovic, Stefan Leucht, Stephen Wood, Christos Davatzikos, Berend Malchow, Peter Falkai, and Nikolaos Koutsouleris (June 2015). "Detecting Neuroimaging Biomarkers for Schizophrenia: A Meta-Analysis of Multivariate Pattern Recognition Studies." In: *Neuropsychopharmacology: Official Publication of the American College of Neuropsychopharmacology* 40.7, pp. 1742–1751.

Kandel, Benjamin M., David A. Wolk, James C. Gee, and Brian Avants (2013). "Predicting Cognitive Data from Medical Images Using Sparse Linear Regression." In: *Information processing in medical imaging : proceedings of the ... conference* 23, pp. 86–97.

Kawasaki, Yasuhiro, Michio Suzuki, Ferath Kherif, Tsutomu Takahashi, Shi-Yu Zhou, Kazue Nakamura, Mie Matsui, Tomiki Sumiyoshi, Hikaru Seto, and Masayoshi Kurachi (Jan. 1, 2007). "Multivariate Voxel-Based Morphometry Successfully Differentiates Schizophrenia Patients from Healthy Controls." In: *NeuroImage* 34.1, pp. 235–242.

Kerr, Wesley T., Pamela K. Douglas, Ariana Anderson, and Mark S. Cohen (Jan. 1, 2014). "The Utility of Data-Driven Feature Selection: Re: Chu et al. 2012." In: *NeuroImage* 84, pp. 1107–1110.

Kia, Seyed Mostafa, Sandro Vega Pons, Nathan Weisz, and Andrea Passerini (2017). "Interpretability of Multivariate Brain Maps in Linear Brain Decoding: Definition, and Heuristic Quantification in Multivariate Analysis of Meg Time-Locked Effects." In: *Frontiers in neuroscience* 10, p. 619.

Kim, Gwang-Won, Yun-Hyeon Kim, and Gwang-Woo Jeong (2017). "Whole Brain Volume Changes and Its Correlation with Clinical Symptom Severity in Patients with Schizophrenia: A DARTEL-Based VBM Study." In: *PloS one* 12.5, e0177251.

Klein, Arno et al. (July 1, 2009). "Evaluation of 14 Nonlinear Deformation Algorithms Applied to Human Brain MRI Registration." In: *NeuroImage* 46.3, pp. 786–802.

Korbinian Brodmann (1909). "Vergleichende Lokalisationslehre Der Grosshirnrinde in Ihren Prinzipien Dargestellt Auf Grund Des Zellenbaues." In:

Le Floch, Edith, Christophe Lalanne, Philippe Pinel, Antonio Moreno, Laura Trinchera, Arthur Tenenhaus, Bertrand Thirion, Jean-Baptiste Poline, Vincent Frouin, and Edouard Duchesnay (2011). "Bridging the Gap between Imaging and Genetics : A Multivariate Statistical Investigation." In: *Human Brain Mapping*.

Le Floch, Edith et al. (Oct. 15, 2012). "Significant Correlation between a Set of Genetic Polymorphisms and a Functional Brain Network Revealed by Feature Selection and Sparse Partial Least Squares." In: *NeuroImage* 63.1, pp. 11–24.

Lefebvre, Stéphanie, Morgane Demeulemeester, Arnaud Leroy, Christine Delmaire, Renaud Lopes, Delphine Pins, Pierre Thomas, and Renaud Jardri (2016). "Network Dynamics during the Different Stages of Hallucinations in Schizophrenia." In: *Human brain mapping* 37.7, pp. 2571–2586.

Lennox, Belinda R, S Bert, G Park, Peter B Jones, and Peter G Morris (1999). "Spatial and Temporal Mapping of Neural Activity Associated with Auditory Hallucinations." In: *The Lancet* 353.9153, p. 644.

Leonard, C. M., M. A. Eckert, and J. M. Kuldau (2006). "Exploiting Human Anatomical Variability as a Link between Genome and Cognome." In: *Genes, Brain, and Behavior* 5 Suppl 1, pp. 64–77.

Leroy, Arnaud, Jack R Foucher, Delphine Pins, Christine Delmaire, Pierre Thomas, Mathilde M Roser, Stéphanie Lefebvre, Ali Amad, Thomas Fovet, Nemat Jaafari, et al. (2017). "Fmri Capture of Auditory Hallucinations: Validation of the Two-Steps Method." In: *Human brain mapping* 38.10, pp. 4966–4979.

Li, Ming, Yadong Liu, Fanglin Chen, and Dewen Hu (2015). "Including Signal Intensity Increases the Performance of Blind Source Separation on Brain Imaging Data." In: *IEEE transactions on medical imaging* 34.2, pp. 551–563.

Lu, Xiaobing, Yongzhe Yang, Fengchun Wu, Minjian Gao, Yong Xu, Yue Zhang, Yongcheng Yao, Xin Du, Chengwei Li, Lei Wu, et al. (2016). "Discriminative Analysis of Schizophrenia Using Support Vector Machine and Recursive Feature Elimination on Structural MRI Images." In: *Medicine* 95.30.

Lê Cao, K.-A., P. G. Martin, C. Robert-Granié, and P. Besse (2009). "Sparse Canonical Methods for Biological Data Integration: Application to a Cross-Platform Study." In: *BMC Bioinformatics* 10.34.

Lê Cao, K.-A., D. Rossouw, C. Robert-Granié, and P. Besse (2008). "A Sparse PLS for Variable Selection When Integrating Omics Data." In: *Statistical Applications in Genetics and Molecular Biology* 7.1.

Löfstedt, Tommy, Vincent Guillemot, Vincent Frouin, Edouard Duchesnay, and Hadj-Selem (2018). "Simulated Data for Linear Regression with Structured and Sparse Penalties: Introducing Pylearn-Simulate | Löfstedt | Journal of Statistical Software." In: *Journal of Statistical Software* 87.3.

MacKay, David J. C. and Cavendish Laboratory (1994). "Bayesian Non-Linear Modelling for the Prediction Competition." In: *In ASHRAE Transactions, V.100, Pt.2*. ASHRAE, pp. 1053–1062.

Mackey, Lester W. (2009). "Deflation Methods for Sparse PCA." In: *Advances in Neural Information Processing Systems 21*. Ed. by D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou. Curran Associates, Inc., pp. 1017–1024.

Maguire, E. A., D. G. Gadian, I. S. Johnsrude, C. D. Good, J. Ashburner, R. S. Frackowiak, and C. D. Frith (Apr. 11, 2000). "Navigation-Related Structural Change in the Hippocampi of Taxi Drivers." In: *Proceedings of the National Academy of Sciences of the United States of America* 97.8, pp. 4398–4403.

Mairal, Julien., Francis. Bach, Jean J. Ponce, and G. Sapiro (2010). "Online Learning for Matrix Factorization and Sparse Coding." In: *Journal of Machine Learning Research* 11, pp. 19–60.

Mangin, J.-F., D. Rivière, A. Cachia, E. Duchesnay, Y. Cointepas, D. Papadopoulos-Orfanos, P. Scifo, T. Ochiai, F. Brunelle, and J. Régis (2004). "A Framework to Study the Cortical Folding Patterns." In: *NeuroImage* 23 Suppl 1, S129–138.

Marquand, Andre F., Iead Rezek, Jan Buitelaar, and Christian F. Beckmann (Oct. 1, 2016). "Understanding Heterogeneity in Clinical Cohorts Using Normative Models: Beyond Case-Control Studies." In: *Biological Psychiatry* 80.7, pp. 552–561.

McAllister, T. W., L. A. Flashman, B. C. McDonald, and A. J. Saykin (2006). "Mechanisms of Cognitive Dysfunction after Mild and

Moderate TBI (MTBI): Evidence from Functional MRI and Neurogenetics." In: *Journal of Neurotrauma* 23.10, pp. 1450–1467.

McCarthy-Jones, Simon, David Smailes, Aiden Corvin, Michael Gill, Derek W Morris, Timothy G Dinan, Kieran C Murphy, John L Waddington, Gary Donohoe, Robert Dudley, et al. (2017). "Occurrence and Co-Occurrence of Hallucinations by Modality in Schizophrenia-Spectrum Disorders." In: *Psychiatry research* 252, pp. 154–160.

McDonald, Carrie, Linda McEvoy, Lusineh Gharapetian, Christine Fennema-Notestine, Donald Hagler, Dominic Holland, Akihide Koyama, James Brewer, Anders Dale, Alzheimer's Disease Neuroimaging Initiative, et al. (2009). "Regional Rates of Neocortical Atrophy from Normal Aging to Early Alzheimer Disease." In: *Neurology* 73.6, pp. 457–465.

Mechelli, Andrea, Paul Allen, Edson Amaro, Cynthia HY Fu, Steven CR Williams, Michael J Brammer, Louise C Johns, and Philip K McGuire (2007). "Misattribution of Speech and Impaired Connectivity in Patients with Auditory Verbal Hallucinations." In: *Human brain mapping* 28.11, pp. 1213–1222.

Michel, Vincent, Alexandre Gramfort, Gaël Varoquaux, Evelyn Eger, and Bertrand Thirion (July 2011). "Total Variation Regularization for fMRI-Based Prediction of Behavior." In: *IEEE Transactions on Medical Imaging* 30.7, pp. 1328–1340.

Mohr, Holger, Uta Wolfensteller, Steffi Frimmel, and Hannes Ruge (Jan. 1, 2015). "Sparse Regularization Techniques Provide Novel Insights into Outcome Integration Processes." In: *NeuroImage* 104, pp. 163–176.

Mondino, Marine, Emmanuel Poulet, Marie-Françoise Suaud-Chagny, and Jerome Brunelin (2016). "Anodal tDCS Targeting the Left Temporo-Parietal Junction Disrupts Verbal Reality-Monitoring." In: *Neuropsychologia* 89, pp. 478–484.

Nesterov, Yu (May 2005a). "Smooth Minimization of Non-Smooth Functions." In: *Math. Program.* 103.1, pp. 127–152.

Nesterov, Yurii (Jan. 2005b). "Excessive Gap Technique in Nonsmooth Convex Minimization." In: *SIAM Journal on Optimization* 16.1, pp. 235–249.

Ng, Bernard, Arash Vahdat, Ghassan Hamarneh, and Rafeef Abugharbieh (Sept. 2012). "Generalized Sparse Classifiers for Decoding Cognitive States in fMRI." In: *SpringerLink*. MICCAI. Beijing, China: Springer Berlin Heidelberg, pp. 108–115.

Nieuwenhuis, Mireille, Neeltje EM van Haren, Hilleke E Hulshoff Pol, Wiepke Cahn, René S Kahn, and Hugo G Schnack (2012). "Classification of Schizophrenia Patients and Healthy Controls from Structural MRI Scans in Two Large Independent Samples." In: *Neuroimage* 61.3, pp. 606–612.

Northoff, Georg and Pengmin Qin (2011). "How Can the Brain's Resting State Activity Generate Hallucinations? A 'Resting State Hypothesis' of Auditory Verbal Hallucinations." In: *Schizophrenia research* 127.1, pp. 202–214.

Orban, Pierre, Christian Dansereau, Laurence Desbois, Violaine Mongeau-Pérusse, Charles-Édouard Giguère, Hien Nguyen, Adrianna Mendrek, Emmanuel Stip, and Pierre Bellec (2018). "Multisite Generalizability of Schizophrenia Diagnosis Classification Based on Functional Brain Connectivity." In: *Schizophrenia research* 192, pp. 167–171.

Orrù, Graziella, William Pettersson-Yeo, Andre F. Marquand, Giuseppe Sartori, and Andrea Mechelli (Apr. 2012). "Using Support Vector Machine to Identify Imaging Biomarkers of Neurological and Psychiatric Disease: A Critical Review." In: *Neuroscience & Biobehavioral Reviews* 36.4, pp. 1140–1152.

Paillère Martinot, Marie-Laure, Jean-Luc Martinot, Damien Ringuenet, André Galinowski, Thierry Gallarda, Frank Bellivier, Jean-Pascal Lefaucheur, Hervé Lemaitre, and Eric Artiges (Dec. 2011). "Baseline Brain Metabolism in Resistant Depression and Response to Transcranial Magnetic Stimulation." In: *Neuropsychopharmacology: Official Publication of the American College of Neuropsychopharmacology* 36.13, pp. 2710–2719.

Pantoni, Leonardo (July 1, 2010). "Cerebral Small Vessel Disease: From Pathogenesis and Clinical Characteristics to Therapeutic Challenges." In: *The Lancet Neurology* 9.7, pp. 689–701.

Parkhomenko, E., D. Tritchler, and J. Beyene (2007). "Genome-Wide Sparse Canonical Correlation of Gene Expression with Genotypes." In: *BMC Proceedings* 1(Suppl 1), S119.

— (2009). "Sparse Canonical Correlation Analysis with Application to Genomic Data Integration." In: *Statistical Applications in Genetics and Molecular Biology* 8(1), Article 1.

Pinel, P., B. Thirion, S. Meriaux, A. Jobert, J. Serres, D. Le Bihan, J.-B. Poline, and S. Dehaene (2007). "Fast Reproducible Identification and Large-Scale Databasing of Individual Functional Cognitive Networks." In: *BMC Neuroscience* 8.91.

Plaze, Marion, Jean-François Mangin, Marie-Laure Paillère-Martinot, Eric Artiges, Jean-Pierre Olié, Marie-Odile Krebs, Raphaël Gaillard, Jean-Luc Martinot, and Arnaud Cachia (2015). ""Who Is Talking to Me?"—Self–Other Attribution of Auditory Hallucinations and Sulcation of the Right Temporoparietal Junction." In: *Schizophrenia research* 169.1, pp. 95–100.

Pol, Hilleke E Hulshoff, Hugo G Schnack, René CW Mandl, Neeltje EM van Haren, Hilde Koning, D Louis Collins, Alan C Evans, and René S Kahn (2001). "Focal Gray Matter Density Changes in Schizophrenia." In: *Archives of General Psychiatry* 58.12, pp. 1118–1125.

Poldrack, Russell A, Jeanette A Mumford, and Thomas E Nichols (2011). *Handbook of Functional MRI Data Analysis*. Cambridge University Press.

Pudil, P., J. Novovi*v*cová, and J. Kittler (Nov. 1, 1994). "Floating Search Methods in Feature Selection." In: *Pattern Recognition Letters* 15.11, pp. 1119–1125.

Radua, Joaquim, S Borgwardt, A Crescini, D Mataix-Cols, A Meyer-Lindenberg, PK McGuire, and P Fusar-Poli (2012). "Multimodal Meta-Analysis of Structural and Functional Brain Changes in First Episode Psychosis and the Effects of Antipsychotic Medication." In: *Neuroscience & Biobehavioral Reviews* 36.10, pp. 2325–2333.

Radua, Joaquim, Erick Jorge Canales-Rodríguez, Edith Pomarol-Clotet, and Raymond Salvador (Feb. 1, 2014). "Validity of Modulation and Optimal Settings for Advanced Voxel-Based Morphometry." In: *NeuroImage* 86, pp. 81–90.

Raichle, Marcus E, Michael D Fox, Abraham Z Snyder, Justin L Vincent, Maurizio Corbetta, and David C Van Essen. "From The Cover: The Human Brain Is Intrinsically Organized into Dynamic, Anticorrelated." In: ().

Rakotomamonjy, Alain (2003). "Variable Selection Using SVM-Based Criteria." In: *Journal of Machine Learning Research* 3, pp. 1357–1370.

Ramezani, Mahdi, Kristopher Marble, Heather Trang, and Purang Abolmaesumi Ingrid Johnsrude (2015). "Joint Sparse Representation of Brain Activity Patterns in Multi-Task fMRI Data." In: *IEEE transactions on medical imaging* 34.1, pp. 2–12.

Rashid, Barnaly and Vince Calhoun (May 6, 2020). "Towards a Brain-Based Predictome of Mental Illness." In: *Human Brain Mapping*.

Ren, Wenting, Su Lui, Wei Deng, Fei Li, Mingli Li, Xiaoqi Huang, Yuqing Wang, Tao Li, John A Sweeney, and Qiyong Gong (2013). "Anatomical and Functional Brain Abnormalities in Drug-Naive First-Episode Schizophrenia." In: *American Journal of Psychiatry* 170.11, pp. 1308–1316.

Ridha, Basil et al. (2008). "Volumetric MRI and Cognitive Measures in Alzheimer Disease." In: *Journal of neurology* 255.4, pp. 567–574.

Roffman, J. L., A. P. Weiss, D. C. Goff, S. L. Rauch, and D. R. Weinberger (2006). "Neuroimaging-Genetic Paradigms: A New Approach to Investigate the Pathophysiology and Treatment of Cognitive Deficits in Schizophrenia." In: *Harvard Review of Psychiatry* 14.2, pp. 78–91.

Roiz-Santiañez, Roberto, Paula Suarez-Pinilla, and Benedicto Crespo-Facorro (2015). "Brain Structural Effects of Antipsychotic Treatment in Schizophrenia: A Systematic Review." In: *Current neuropharmacology* 13.4, pp. 422–434.

Rotarska-Jagiela, Anna, Vincent van de Ven, Viola Oertel-Knöchel, Peter J Uhlhaas, Kai Vogeley, and David EJ Linden (2010). "Resting-

State Functional Network Correlates of Psychotic Symptoms in Schizophrenia." In: *Schizophrenia research* 117.1, pp. 21–30.

Rozycki, Martin, Theodore D Satterthwaite, Nikolaos Koutsouleris, Guray Erus, Jimit Doshi, Daniel H Wolf, Yong Fan, Raquel E Gur, Ruben C Gur, Eva M Meisenzahl, et al. (2017). "Multisite Machine Learning Analysis Provides a Robust Structural Imaging Signature of Schizophrenia Detectable across Diverse Patient Populations and within Individuals." In: *Schizophrenia bulletin*.

Rudolf Nieuwenhuys (2013). "The Myeloarchitectonic Studies on the Human Cerebral Cortex of the Vogt–Vogt School, and Their Significance for the Interpretation of Functional Neuroimaging Data." In: *Brain Structure and Function* 218.2, pp. 303–352.

Sabuncu, Mert R, Ender Konukoglu, Alzheimer's Disease Neuroimaging Initiative, et al. (2015). "Clinical Prediction from Structural Brain MRI Scans: A Large-Scale Empirical Study." In: *Neuroinformatics* 13.1, pp. 31–46.

Sapolsky, Robert M. (Oct. 23, 2001). "Depression, Antidepressants, and the Shrinking Hippocampus." In: *Proceedings of the National Academy of Sciences of the United States of America* 98.22, pp. 12320–12322.

Schmidt, Andre, Felix Müller, Claudia Lenz, Patrick Dolder, Yasmin Schmid, Davide Zanchi, Undine Lang, Matthias Liechti, and Stefan Borgwardt (2017). "Acute LSD Effects on Response Inhibition Neural Networks." In: *Psychological medicine*, pp. 1–13.

Schmidt, Mark, Nicolas Le Roux, and Francis Bach (Dec. 2011). "Convergence Rates of Inexact Proximal-Gradient Methods for Convex Optimization." In: *NIPS'11 - 25 Th Annual Conference on Neural Information Processing Systems*.

Schwarz, G. (1978). "Estimating the Dimension of a Model." In: *The Annals of Statistics* 6, pp. 461–464.

Schölkopf, Bernhard and Alexander J. Smola (Jan. 2002). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press. 658 pp.

Sereno, M. I., A. M. Dale, J. B. Reppas, K. K. Kwong, J. W. Belliveau, T. J. Brady, B. R. Rosen, and R. B. Tootell (May 12, 1995). "Borders of Multiple Visual Areas in Humans Revealed by Functional Magnetic Resonance Imaging." In: *Science (New York, N.Y.)* 268.5212, pp. 889–893.

Shen, Hui et al. (2017). "Making Group Inferences Using Sparse Representation of Resting-State Functional mRI Data with Application to Sleep Deprivation." In: *Human Brain Mapping* 38.9, pp. 4671–4689.

Shepherd, Alana M, Kristin R Laurens, Sandra L Matheson, Vaughan J Carr, and Melissa J Green (2012). "Systematic Meta-Review and Quality Assessment of the Structural Brain Alterations in Schizophrenia." In: *Neuroscience & Biobehavioral Reviews* 36.4, pp. 1342–1356.

Smieskova, Renata, Paolo Fusar-Poli, Paul Allen, Kerstin Bendfeldt, Rolf-Dieter Stieglitz, Juergen Drewe, Ernst Radue, Philip McGuire, Anita Riecher-Rossler, and Stefan Borgwardt (2009). "The Effects of Antipsychotics on the Brain: What Have We Learnt from Structural Imaging of Schizophrenia? A Systematic Review." In: *Current pharmaceutical design* 15.22, pp. 2535–2549.

Sommer, Iris EC, Kelly MJ Diederen, Jan-Dirk Blom, Anne Willems, Leila Kushan, Karin Slotema, Marco PM Boks, Kirstin Daalman, Hans W Hoek, Sebastiaan FW Neggers, et al. (2008). "Auditory Verbal Hallucinations Predominantly Activate the Right Inferior Frontal Area." In: *Brain* 131.12, pp. 3169–3177.

Soneson, C., H. Lilljebjörn, T. Fioretos, and M. Fontes (2010). "Integrative Analysis of Gene Expression and Copy Number Alterations Using Canonical Correlation Analysis." In: *BMC Bioinformatics* 11.191.

Stein, J. et al. (2010). "Voxelwise Genome-Wide Association Study (vGWAS)." In: *NeuroImage* 53, pp. 1160–1174.

Sun, Daqiang, Theo G. M. van Erp, Paul M. Thompson, Carrie E. Bearden, Melita Daley, Leila Kushan, Molly E. Hardt, Keith H. Nuechterlein, Arthur W. Toga, and Tyrone D. Cannon (Dec. 1, 2009). "Elucidating a Magnetic Resonance Imaging-Based Neuroanatomic Biomarker for Psychosis: Classification Analysis Using Probabilistic Brain Atlas and Machine Learning Algorithms." In: *Biological Psychiatry* 66.11, pp. 1055–1060.

Teicher, Martin H., Carl M. Anderson, Kyoko Ohashi, and Ann Polcari (Aug. 15, 2014). "Childhood Maltreatment: Altered Network Centrality of Cingulate, Precuneus, Temporal Pole and Insula." In: *Biological Psychiatry* 76.4, pp. 297–305.

Thirion, Bertrand, Edouard Duchesnay, Edward Hubbard, Jessica Dubois, Jean-Baptiste Poline, Denis Lebihan, and Stanislas Dehaene (Dec. 2006). "Inverse Retinotopy: Inferring the Visual Content of Images from Brain Activation Patterns." In: *NeuroImage* 33.4, pp. 1104–1116.

Thompson, Paul et al. (2004). "Mapping Hippocampal and Ventricular Change in Alzheimer Disease." In: *NeuroImage* 22.4, pp. 1754–1766.

Tibshirani, Robert (1996). "Regression Shrinkage and Selection via the Lasso." In: *Journal of the Royal Statistical Society. Series B (Methodological)* 58.1, pp. 267–288.

Torres, Ulysses S, Fabio LS Duran, Maristela S Schaufelberger, José AS Crippa, Mario R Louzã, Paulo C Sallet, Caroline YO Kanegusuku, Helio Elkis, Wagner F Gattaz, Débora P Bassitt, et al. (2016). "Patterns of Regional Gray Matter Loss at Different Stages of Schizophrenia: A Multisite, Cross-Sectional VBM Study in First-Episode and Chronic Illness." In: *NeuroImage: Clinical* 12, pp. 1–15.

Torres, Ulysses S, Eduardo Portela-Oliveira, Stefan Borgwardt, and Geraldo F Busatto (2013). "Structural Brain Changes Associated with Antipsychotic Treatment in Schizophrenia as Revealed by Voxel-Based Morphometric MRI: An Activation Likelihood Estimation Meta-Analysis." In: *BMC psychiatry* 13.1, p. 342.

Van Leemput, Koen, Akram Bakkour, Thomas Benner, Graham Wiggins, Lawrence L. Wald, Jean Augustinack, Bradford C. Dickerson, Polina Golland, and Bruce Fischl (2008). "Model-Based Segmentation of Hippocampal Subfields in Ultra-High Resolution in Vivo MRI." In: *Medical image computing and computer-assisted intervention: MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention* 11 (Pt 1), pp. 235–243.

Van Erp, Theo GM, Derrek P Hibar, Jerod M Rasmussen, David C Glahn, Godfrey D Pearlson, Ole A Andreassen, Ingrid Agartz, Lars T Westlye, Unn K Haukvik, Anders M Dale, et al. (2016). "Subcortical Brain Volume Abnormalities in 2028 Individuals with Schizophrenia and 2540 Healthy Controls via the ENIGMA Consortium." In: *Molecular psychiatry* 21.4, p. 547.

Vapnik, Vladimir (Nov. 19, 1999). *The Nature of Statistical Learning Theory*. 2nd edition. New York: Springer. 314 pp.

Varoquaux, Gaël (Oct. 15, 2018). "Cross-Validation Failure: Small Sample Sizes Lead to Large Error Bars." In: *NeuroImage*. New Advances in Encoding and Decoding of Brain Signals 180, pp. 68–77.

Varoquaux, Gaël, Michael Eickenberg, Elvis Dohmatob, and Bertand Thirion (Sept. 2015). "FAASTA: A Fast Solver for Total-Variation Regularization of Ill-Conditioned Problems with Application to Brain Imaging." In: *Colloque GRETSI*. P. Gonçalves, P. Abry. Lyon, France.

Vercammen, Ans, Henderikus Knegtering, Johann A den Boer, Edith J Liemburg, and André Aleman (2010). "Auditory Hallucinations in Schizophrenia Are Associated with Reduced Functional Connectivity of the Temporo-Parietal Area." In: *Biological psychiatry* 67.10, pp. 912–918.

Vita, A, Luca De Peri, Claudio Silenzi, and Massimiliano Dieci (2006). "Brain Morphology in First-Episode Schizophrenia: A Meta-Analysis of Quantitative Magnetic Resonance Imaging Studies." In: *Schizophrenia research* 82.1, pp. 75–88.

Vounou, M., T.E. Nichols, and G. Montana (2010). "Discovering Genetic Associations with High-Dimensional Neuroimaging Phenotypes: A Sparse Reduced-Rank Approach." In: *NeuroImage* 53, pp. 1147–1159.

Waaijenborg, S., P. Verselewel de Witt Hamer, and A. Zwinderman (2008). "Quantifying the Association between Gene Expressions and DNA-Markers by Penalized Canonical Correlation Analysis." In: *Statistical Applications in Genetics and Molecular Biology* 7(1), Article 3.

Wahlberg, Bo, Stephen Boyd, Mariette Annergren, and Yang Wang (Mar. 2012). "An ADMM Algorithm for a Class of Total Variation Regularized Estimation Problems." In: *ArXiv e-prints*.

Wang, Lei, Alex Kogan, Derin Cobia, Kathryn Alpert, Anthony Kolasny, Michael I. Miller, and Daniel Marcus (2013). "Northwestern University Schizophrenia Data and Software Tool (NUSDAST)." In: *Frontiers in Neuroinformatics* 7, p. 25.

Wang, Wen-Ting. and Hsin-Cheng. Huang (2015). "Regularized Principal Component Analysis for Spatial Data." In: *ArXiv e-prints*.

Wang, Ze, Anna R. Childress, Jiongjiong Wang, and John A. Detre (July 15, 2007). "Support Vector Machine Learning-Based fMRI Data Group Analysis." In: *NeuroImage* 36.4, pp. 1139–1151.

Weichwald, Sebastian, Timm Meyer, Ozan Özdenizci, Bernhard Schölkopf, Tonio Ball, and Moritz Grosse-Wentrup (2015). "Causal Interpretation Rules for Encoding and Decoding Models in Neuroimaging." In: *NeuroImage* 110, pp. 48–59.

Wisse, Laura E. M., Geert Jan Biessels, and Mirjam I. Geerlings (Sept. 25, 2014). "A Critical Appraisal of the Hippocampal Subfield Segmentation Package in FreeSurfer." In: *Frontiers in Aging Neuroscience* 6.

Witten, D.M. and R. Tibshirani (2009). "Extensions of Sparse Canonical Correlation Analysis, with Applications to Genomic Data." In: *Statistical Applications in Genetics and Molecular Biology* 8(1), Article 28.

Witten, Daniela, Robert Tibshirani, and Trevor Hastie (2009). "A Penalized Matrix Decomposition, with Applications to Sparse Principal Components and Canonical Correlation Analysis." In: *Biostatistics* 10.3, pp. 515–534.

Wold, H. (1966). "Multivariate Analysis." In: Academic Press, New York. Chap. Estimation of principal components and related models by iterative least squares, pp. 391–420.

Wold, S., H. Martens, and H. Wold (1983). "The Multivariate Calibration Problem in Chemistry Solved by the PLS Method." In: *Proceedings Conference Matrix Pencils*. Ed. by A. Ruhe and B. Kastrøm. Vol. Lecture Notes in Mathematics. Springer-Verlag, pp. 286–293.

Zatorre, Robert J., R. Douglas Fields, and Heidi Johansen-Berg (Apr. 2012). "Plasticity in Gray and White: Neuroimaging Changes in Brain Structure during Learning." In: *Nature Neuroscience* 15.4 (4), pp. 528–536.

Zheng, Zhilin and Li Sun (Mar. 15, 2019). "Disentangling Latent Space for VAE by Label Relevant/Irrelevant Dimensions." In:

Zou, Hui and Trevor Hastie (2005). "Regularization and Variable Selection via the Elastic Net." In: *Journal of the Royal Statistical Society, Series B* 67, pp. 301–320.

Zou, Hui, Trevor Hastie, and Robert Tibshirani (2006). "Sparse Principal Component Analysis." In: *Journal of Computational and Graphical Statistics* 15.2, pp. 265–286.