Multi-subject dictionary learning to segment an atlas of brain spontaneous activity

Gaël Varoquaux, Alexandre Gramfort, Fabian Pedregosa, Vincent Michel, Bertrand Thirion

Multi-subject dictionary learning to segment an atlas of brain spontaneous activity

G. Varoquaux[123], A. Gramfort[23], F. Pedregosa[23], V. Michel[23], B. Thirion[23]

[1] INSERM U992 Cognitive Neuroimaging unit,
[2] INRIA, Parietal team, Saclay, France
[3] LNAO/NeuroSpin, CEA Saclay, Bat. 145, 91191 Gif-sur-Yvette, cedex France

Abstract. Fluctuations in brain on-going activity can be used to reveal its intrinsic functional organization. To mine this information, we give a new hierarchical probabilistic model for brain activity patterns that does not require an experimental design to be specified. We estimate this model in the dictionary learning framework, learning simultaneously latent spatial maps and the corresponding brain activity time-series. Unlike previous dictionary learning frameworks, we introduce an explicit difference between subject-level spatial maps and their corresponding population-level maps, forming an atlas. We give a novel algorithm using convex optimization techniques to solve efficiently this problem with non-smooth penalties well-suited to image denoising. We show on simulated data that it can recover population-level maps as well as subject specificities. On resting-state fMRI data, we extract the first atlas of spontaneous brain activity and show how it defines a subject-specific functional parcellation of the brain in localized regions.

1 Introduction

The study of intrinsic brain functional organization via distant correlations in the fluctuations of brain signals measured by functional Magnetic Resonance Imaging (fMRI) is receiving increasing interest. In particular, the 1000 Functional Connectomes project aims at parceling the brain in functional regions and then at studying the correlation structure of brain function across these nodes [5]. Independent Component Analysis (ICA) is the most popular data-driven approach to analyze spontaneous activity, as it has been shown to extract interpretable spatial patterns [3] that are reproducible across subjects [9]. They form networks of functional regions that are also found in task-driven studies [23].

From a medical point of view, the development of statistically-controlled analysis of brain spontaneous is interesting as it can lead to new diagnostic or prognostic tools applicable on impaired patients. In particular, correlations in the functional signal between predefined regions have been shown to contain markers of post-stroke functional reorganization [25]. However, inferences drawn from these regions depends on the targeted regions, and on their precise delineation. In addition, subject-to-subject local variability, for instance in functional topography, may confound the changes in long-distance interactions.

We address the segmentation of functional regions directly from the fMRI signal. The challenge stems from the lack of salient features in the original signal, as well as the lack of a controlled experimental design to perform model fitting as in task-driven fMRI experiments. In particular, it is difficult to optimize the parameters (dimension and regularization) of the models, hence to obtain an arguably faithful and meaningful representation of this data. ICA tackles these difficulties by estimating a mixing matrix to minimize the mutual information between the resulting spatial components. Departing from ICA, [12] performs segmentation by clustering the time series through a mixture model. However, these approaches lack an explicit noise model and do not take into account the subject-to-subject variability nor the spatial structure of the signal. In this paper, we formulate the problem in the dictionary learning framework and reject observation noise based on the assumption that the relevant patterns are spatially sparse [10, 26], and we focus on the choice of the involved parameters. The paper is organized as follows: we give in section 2 a two-level probabilistic model that involves subject-specific spatial maps as well as population-level latent maps, and in section 3 an associated efficient learning algorithm. In section 4 we describe how to set the model parameters from the data. In section 5, we study different learning schemes on synthetic data with simulated inter-individual variability. Finally, in section 6 we apply the method to learning a detailed population-level atlas of regions describing spontaneous activity as recorded in fMRI.

2 Multi-subject decomposition model for brain activity

Problem statement We consider a dataset of brain signal time series of length n for S subjects, measured on p voxels: $\{\mathbf{Y}^s \in \mathbb{R}^{n\times p}, s = 1 \dots S\}$. We stipulate that the corresponding 3D images are the observation of k spatial latent factors $\mathbf{V}^s \in \mathbb{R}^{p\times k}$, that characterize functional processes or structured measurement artifacts, and associated time series $\mathbf{U}^s \in \mathbb{R}^{n\times k}$: $\mathbf{Y}^s \approx \mathbf{U}^s{\mathbf{V}^s}^T$. We are interested in the study of resting state, or on-going activity, for which no experimental design can be used to model time-courses, thus we propose to learn $\mathbf{U}^s$ and $\mathbf{V}^s$ simultaneously, a problem known as *dictionary learning*, or linear signal decomposition [7, 17].

Generative model In the case of a multi-subject dataset, we give a hierarchical probabilistic model for dictionary learning. Following the standard dictionary learning model, the data observed for each subject is written as the linear combination of subject-specific dictionary elements, that are spatial maps $\mathbf{V}_s$. For resting-state brain activity, we do not model the loadings $\mathbf{U}_s$ themselves, but their covariance.

$$\forall s \in \{1 \dots S\}, \mathbf{Y}^s = \mathbf{U}^s{\mathbf{V}^s}^T + \mathbf{E}^s, \qquad \mathbf{E}^s \sim \mathcal{N}(0, \sigma\mathbf{I}), \qquad \mathbf{U}^s \sim \mathcal{N}(0, \mathbf{\Sigma_U}) \quad (1)$$

In addition, the subject-specific maps $\mathbf{V}_s$ are generated from population-level latent factors, the spatial patterns written as brain maps $\mathbf{V}$:

$$\forall s \in \{1 \dots S\}, \mathbf{V}^s = \mathbf{V} + \mathbf{F}^s, \qquad \mathbf{F}^s \sim \mathcal{N}(0, \zeta\mathbf{I}) \quad (2)$$

Finally, we specify the prior distribution on $\mathbf{V}$: $\mathcal{P}(\mathbf{V}) \propto \exp\left(-\xi\, \Omega(\mathbf{V})\right)$, where Ω is typically a norm or a quasi-norm.

Relation to existing models With $\zeta = 0$, the model identifies $\mathbf{V}^s$ with $\mathbf{V}$: all latent factors are the same across subjects. In this context, if the prior on $\mathbf{V}$ is un-informative, the model boils down to a principal component analysis (PCA) on the concatenated $\mathbf{Y}^s$. For a Laplace prior, we recover probabilistic formulation of a standard sparse ℓ_1-penalized PCA [22]. More generally, in this framework, sparsity-inducing priors give rise to a family of probabilistic projection models [1]. Our multi-subject model however differs from generalized canonical correlation analysis [16], and its sparse variants [1], as these approaches do not model subject-specific latent factors and thus do not allow for two levels of variance. Note that, for multi-subject studies, non-hierarchical models based on PCA and ICA impose orthogonality constraints on the loadings at the group level, and thus introduce a unnatural constraint on the $\mathbf{U}^s$ across the different subjects.

ICA can be formulated in a maximum likelihood approach [4] and thus falls in the same general class of non-hierarchical dictionary learning models [17]. However, as ICA disregards explained variance, it leads to improper priors on $\mathbf{V}$ and requires the use of a PCA pre-processing step to estimate the noise[4] [3]. In neuroimaging, multi-subject dictionary learning using a fixed group model ($\zeta = 0$) in combination with ICA is popular, and called *concatenated ICA* [6]. In the experimental section of this paper, we will focus on the use of proper priors on $\mathbf{V}$ based on sparsity-inducing norms Ω, such as the ℓ_1 norm. They are known to be efficient in terms of separating signal from noise, in the supervised settings [27], and lead to tractable optimizations that are convex, though non-smooth.

3 Optimization strategy for efficient learning

We now present a new algorithm to efficiently estimate from the data at hand the model specified by Eq. (1) and (2). In the following, we call this problem Multi-Subject Dictionary Learning (MSDL). In the maximum a posteriori (MAP) estimation framework, we learn the parameters from the data by maximizing the sum of the log-likelihood of the data given the model, and penalization terms that express our hierarchical priors. In addition, as the variance of the group-level residuals in Eq. (2) could be arbitrarily shrunk by shrinking the norm of $\mathbf{V}$, we impose an upper bound on the norm of the columns of $\mathbf{U}^s$:

$$(\mathbf{U}^s, \mathbf{V}^s)_{s\in\{1\dots S\},}, \mathbf{V} = \underset{\mathbf{U}^s,\mathbf{V}^s,\mathbf{V}}{\operatorname{argmin}} \mathcal{E}(\mathbf{U}^s, \mathbf{V}^s, \mathbf{V}), \qquad \text{s.t. } \|\mathbf{u}_l^s\|_2^2 \leq 1 \tag{3}$$

$$\text{with} \quad \mathcal{E}(\mathbf{U}^s, \mathbf{V}^s, \mathbf{V}) = \sum_{s=1}^{S} \frac{1}{2}\left(\|\mathbf{Y}^s - \mathbf{U}^s \mathbf{V}^{sT}\|_{\text{Fro}}^2 + \mu \|\mathbf{V}^s - \mathbf{V}\|_{\text{Fro}}^2 \right) + \lambda\, \Omega(\mathbf{V}),$$

where $\mu = \frac{\sigma}{\zeta}$ and $\lambda = \frac{\sigma}{\xi}$. The optimization problem given by Eq. (3) is not jointly convex in $\mathbf{U}^s$, $\mathbf{V}^s$, and $\mathbf{V}$, however it is separately convex in $\mathbf{V}^s$ and $(\mathbf{U}^s, \mathbf{V})$. Our

[4] There exist noisy ICA approaches, but they all assume that the contribution of the noise to the observed data is small.

optimization strategy relies on alternating optimizations of $\mathbf{V}^s$, $\mathbf{U}^s$, $\mathbf{V}$, keeping other parameters constant. In the following we give the mathematical analysis of the optimization procedure; the exact operations are detailed in algorithm 1.

Following [18], we use a block coordinate descent, to minimize $\mathcal{E}$ as a function of $\mathbf{U}_s$. Solving Eq. (3) as a function of $\mathbf{V}^s$ corresponds to a ridge regression problem on the variable $(\mathbf{V}^s - \mathbf{V})^T$, the solution of which can be computed efficiently (line 9, algorithm 1). Minimizing $\mathcal{E}$ as a function of $\mathbf{V}$ corresponds to minimizing $\sum_{s=1}^{S} \frac{1}{2}\|\mathbf{v}^s - \mathbf{v}\|_2^2 + \frac{\lambda}{\mu}\,\Omega(\mathbf{v})$ for all column vectors $\mathbf{v}$ of $\mathbf{V}$. The solution is a proximal operator [8], as detailed in lemma 1.

Lemma 1. $\operatorname*{argmin}_{\mathbf{v}}\big(\sum_{s=1}^{S} \frac{1}{2}\|\mathbf{v}^s - \mathbf{v}\|_2^2 + \gamma\,\Omega(\mathbf{v})\big) = \operatorname*{prox}_{\gamma/S\ \Omega} \bar{\mathbf{v}}$, *where* $\bar{\mathbf{v}} = \frac{1}{S}\sum_{s=1}^{S} \mathbf{v}^s$.

The proof of lemma 1 follows from the fact that $\sum_{s=1}^{S} \|\mathbf{v}^s - \mathbf{v}\|_2^2 = S\sum_{s=1}^{S} \|\bar{\mathbf{v}} - \mathbf{v}\|_2^2 + \sum_{s=1}^{S} \|\bar{\mathbf{v}} - \mathbf{v}^s\|_2^2$, as the second term at the right hand side is independent from $\mathbf{v}$, the minimization problem simplifies to minimizing the first term, which corresponds to the problem solved by the proximal operator on $\bar{\mathbf{v}}$.

Algorithm 1 Solving optimization problem given in Eq. (3)

Input: $\{\mathbf{Y}^s \in \mathbb{R}^{n\times p}, s = 1, \ldots, S\}$, the time series for each subject; k, the number of maps; an initial guess for $\mathbf{V}$.

Output: $\mathbf{V} \in \mathbb{R}^{p\times k}$ the group-level spatial maps, $\{\mathbf{V}^s \in \mathbb{R}^{p\times k}\}$ the subject-specific spatial maps, $\{\mathbf{U}^s \in \mathbb{R}^{n\times k}\}$ the associated time series.

1: $E_0 \leftarrow \infty$, $E_1 \leftarrow \infty$, $i \leftarrow 1$ (initialize variables).
2: $\mathbf{V}^s \leftarrow \mathbf{V}$, $\quad \mathbf{U}_s \leftarrow \mathbf{Y}^s\mathbf{V}(\mathbf{V}^T\mathbf{V})^{-1}$, $\quad$ for $s = 1 \ldots S$
3: **while** $E_i - E_{i-1} > \varepsilon E_{i-1}$ **do**
4: **for** s=1 to S **do**
5: **for** l=1 to k **do**
6: Update $\mathbf{U}^s$: $\quad \mathbf{u}_l^s \leftarrow \mathbf{u}_l^s + \|\mathbf{v}_l^s\|_2^{-2}(\mathbf{Y}^s - \mathbf{U}^s\mathbf{V}^{sT})\mathbf{v}_l^s$ $\quad$ (following [15])
7: $\mathbf{u}_l^s \leftarrow \mathbf{u}_l^s / \max(\|\mathbf{u}_l^s\|_2, 1)$
8: **end for**
9: Update $\mathbf{V}^s$ (ridge regression): $\mathbf{V}^s \leftarrow \mathbf{V} + (\mathbf{Y}^s - \mathbf{U}^s\mathbf{V}^T)^T\mathbf{U}^s(\mathbf{U}^{sT}\mathbf{U}^s + \mu\mathbf{I})^{-1}$
10: **end for**
11: Update $\mathbf{V}$ using lemma 1: $\quad \mathbf{V} \leftarrow \operatorname*{prox}_{\lambda/S\mu\ \Omega}\big(\frac{1}{S}\sum_{s=1}^{S} \mathbf{V}^s\big)$.
12: Compute value of energy: $\quad E_i \leftarrow \mathcal{E}(\mathbf{U}^s, \mathbf{V}^s, \mathbf{V})$
13: $i \leftarrow i + 1$
14: **end while**

Choice of initialization The optimization problem given by Eq. (3) is not convex, and thus the output of algorithm 1 depends on the initialization. As ICA applied to fMRI data extracts super Gaussian signals and thus can be used for sparsity recovery [26], we initialize $\mathbf{V}$ with maps extracted with the fastICA algorithm [14], initialized with a random mixing matrix. However, as not all spatial maps estimated by ICA are super-Gaussian, we run ICA with an

increasing model order until it selects k maps with a positive kurtosis[5]. Note that ICA also solves a non convex problem. We find empirically that this initialization procedure strongly reduces the number of iterations required for convergence compared to initialization with the results of a PCA or to random initialization.

Algorithmic complexity The complexity of one iteration of algorithm 1 can be decomposed in the three contributions, corresponding to the update of each term. First, updating $\mathbf{U}^s$ for each $s \in \{1 \dots S\}$ costs $\mathcal{O}\left((p+n)\,k^2 + k\,n\,p\right)$. Second, updating $\mathbf{V}^s$ requires $\mathcal{O}\left(k^3 + k\,n\,p\right)$ operations. Last, the computational cost of updating $\mathbf{V}$ is given by the cost of computing the proximal operator for Ω, $T_\Omega(p)$, times the number of dictionary elements. Thus the total cost of one iteration is in $\mathcal{O}\left(S\left(k\,n\,p + (p+n)\,k^2 + k^3\right) + k\,T_\Omega(p)\right)$. This expression is linear in the total number of samples $S\,n$. This is important in population studies as, unlike previous work, our approach uses the full multi-subject data set to perform simultaneously denoising and latent factor estimation. In addition, there exist closed-form expressions or efficient algorithms for computing the proximal operator for many interesting choices of Ω, in which case algorithm 1 also scales well with p and can be applied to high-resolution brain images. Note that, due to the size of the datasets, we use randomized projection algorithms to perform truncated SVDs required for initialization and model selection (detailed later).

Imposing smooth sparsity Sparsity is typically induced with an ℓ_1 prior. However using the ℓ_1 norm disregards the fact that the penalized variables are actually spatial maps, hence have an inherent grid structure. In order to promote sparse and spatially coherent maps for $\mathbf{V}$, and consequently for the $\mathbf{V}^s$, we propose to use a Smooth-Lasso (SL) penalty [13]. The SL amounts to adding to the Lasso term an ℓ_2 penalty on the gradient, which straightforwardly yields:

$$\Omega_{\mathrm{SL}}(\mathbf{v}) = \|\mathbf{v}\|_1 + \frac{1}{2}\mathbf{v}^T\mathbf{L}\mathbf{v} \tag{4}$$

where $\mathbf{L}$ is the Laplacian operator defined on the 3D grid of voxels. The computation of the proximal operator associated to $\gamma\Omega_{\mathrm{SL}}$ is detailed in Algorithm 2. It can be optimized with a fast first order method called FISTA [2] after noticing that the cost function to optimize can be divided in 2 terms. A convex smooth term with Lipschitz gradient formed by $\frac{1}{2}\|\mathbf{v}-\bar{\mathbf{v}}\|^2 + \frac{\gamma}{2}\mathbf{v}^T\mathbf{L}\mathbf{v}$ and a convex term $\gamma\|\mathbf{v}\|_1$. The Lipschitz constant of the smooth term is given by $1 + \gamma\|\mathbf{L}\|$, where $\|\mathbf{L}\|$ stands for the spectral norm of the Laplacian operator.

4 Model selection and choice of parameters

In this section we detail how the different parameters of the model are set.

Setting the subject-level penalization For a given choice of number of components k, we set the group-level penalization constant μ from the data,

[5] There is no guarantee that this procedure converges. We have observed empirically that, on fMRI data, high model-order ICA can extract at least 100 super-Gaussian maps. If the procedure does not converge, we suggest reducing k.

Algorithm 2 Proximal operator for the smooth-lasso with FISTA

Input: Spatial map $\mathbf{v}$
Output: Denoised map $\mathbf{v}^\star$

$\mathbf{z} = \mathbf{v}^\star = \mathbf{v}$, $\tau = 1$, $0 < \kappa < (1 + \gamma\|\mathbf{L}\|)^{-1}$
for l=1 to k **do**
  $\mathbf{v}_o = \mathbf{v}^\star$
  $\mathbf{v}^\star = s_{\kappa\gamma}\left(\mathbf{z} - \kappa(\mathbf{z} - \mathbf{v} + \gamma\mathbf{L}\mathbf{z})\right)$
  $\tau_o = \tau$
  $\tau = \frac{1+\sqrt{1+4\tau^2}}{2}$
  $\mathbf{z} = \mathbf{v}^\star + \frac{\tau_o - 1}{\tau}(\mathbf{v}^\star - \mathbf{v}_o)$
end for

s is the element-wise soft-thresholding operator: $s_{\kappa\gamma}(\cdot) = \mathrm{sign}(\cdot)\max(|\cdot| - \kappa\gamma, 0)$ [11]

by computing estimates of the intra-subject variance σ and the inter-subject variance ζ. We first compute a lower bound e on $n\,\sigma$ using the variance of the residuals of a PCA of order k performed on each subject datasets $\mathbf{Y}^s$. Indeed a PCA gives the solution to Eq. (1) minimizing σ. Then, we note that, replacing Eq. (2) in Eq. (1), we have $\{\mathbf{Y}^s = \mathbf{U}^s\mathbf{V} + \mathbf{U}^s\mathbf{F}^s + \mathbf{E}^s, s = 1 \dots S\}$. Thus, we can have a lower bound f on the sum of square of $\{\mathbf{U}^s\mathbf{F}^s + \mathbf{E}^s, s = 1 \dots S\}$ by running a PCA on the concatenated $\{\mathbf{Y}^s, s = 1 \dots S\}$. If we consider that $\mathbf{F}^s$ and $\mathbf{E}^s$ are independent, we have $f \approx s\,(n\,\sigma + k\,\zeta)$, thus we set $\lambda = k/n\,(f/(s\,e) - 1)^{-1}$.

Setting population-level penalization and model order At the subject level, we can compute the likelihood of new data based on the probabilistic model given in Eq. (1). This model corresponds the probabilistic PCA model [24], and relies on the multivariate Gaussian likelihood:

$$\mathcal{L}(\mathbf{Y}^s|\mathbf{V}^s, \mathbf{\Sigma}_\mathbf{U}, \sigma) = -\frac{1}{2}\log|\mathbf{\Sigma}_{\text{model}}| - \frac{1}{2n}\mathrm{tr}\big(\mathbf{Y}^s\mathbf{\Sigma}_{\text{model}}^{-1}\mathbf{Y}^{sT}\big) + \text{cste}, \quad (5)$$

$$\text{with} \qquad \mathbf{\Sigma}_{\text{model}} = \mathbf{V}^s\mathbf{\Sigma}_\mathbf{U}\mathbf{V}^{sT} + \sigma\mathbf{I} \quad (6)$$

Note that while the matrix $\mathbf{\Sigma}_\mathbf{U}$ can be very large ($p \times p$), it is of low rank (its rank is given by the rank of $\mathbf{V}^s$, k). Thus the above likelihood can be computed on large data at a low computational and memory cost using the Woodbury matrix identity and the matrix determinant lemma.

We set the amount of population-level penalization by choosing the parameter λ amongst a grid to maximize the likelihood of left-out data in a 3-fold cross-validation scheme. We apply stratified cross-validation: in each fold, we learn the model on two-thirds of each subject's dataset, and test on the left out third. We choose to split the data for each subject rather than splitting the population by leaving some subjects out, as we are interested in learning maps that give good models for each subject, rather than a model of variability across the population. We also apply cross-validation to select the model order k, i.e. the number of dictionary elements. However, as already noted in [21], setting a high model order may only lead to a saturation of the likelihood of left-out data, and not a decrease. In addition, at high model-order the model can learn patterns that account for subject-to-subject variability, as we will see on simulated data.

5 Simulation study

Synthetic data generation We generate latent spatial maps that match the spatial structure of functional networks or artifact present in the fMRI signal. From these maps, we generate observations by mixing them with random time series and adding random spatially-correlated Gaussian noise. To create the latent spatial maps, we use a blob model: each map is made of a few cones-shaped localized patterns, the position and size of which are chosen randomly (see Fig. 1). The number of blobs on each map is given by a binomial distribution corresponding to 3 trials with 50% success rate. Maps are generated one after the other with blob positions chosen randomly, but avoiding overlap between blobs across maps. In addition, we generate subject-specific latent spatial maps by adding an isotropic, Gaussian-distributed, jitter on the blobs position and width.

Empirical results on synthetic data Synthetic datasets of 12 subjects, with 5 latent spatial maps, 150 time points, and 50×50 spatial grids were generated. In Fig. 1, we represent generated population maps for a smoothness of 2 pixels and a jitter of 3 pixels, as well as the corresponding estimates by different methods: thresholded ICA for sparse recovery [26], ℓ_1-penalized sparse PCA (SPCA), and MSDL. On this dataset, the three methods find similar maps, but thresholded ICA displays more high-frequency noise. In Fig. 2, we display the subject-specific maps estimated by our method. We can see that the method makes a compromise between fitting the observed noisy subject-level data, and finding the same latent factors across subjects. As a result, the subject-specific maps capture the inter-subject variability, but also some observation noise. To quantify the quality of the recovery, we have generated 30 datasets, with varying amount of between-subject spatial jitter and spatial smoothness of the observation noise. For each of these datasets, we compute the best assignment matching the estimated maps with the ground truth using the Kuhn-Munkres algorithm[20] to maximize cross-correlation. In Fig 5 we report the average cross-correlation across all maps and synthetic datasets for the population-level maps, but also for the subject-specific maps for which the ground truth is compared to the corresponding estimated subject-specific map in the case of the MSDL approach, and to the estimated population mean for other methods. We find that thresholded ICA is always the worst performer. If there is no spatial jitter across subjects, MSDL and SPCA perform similarly for the recovery of population-level maps, but SPCA outperforms MSDL for the recovery of subject-level maps. This can be explained by the fact that, in this case, the subject-level specificities learned are observation noise: our heuristic over-estimates μ. When there is indeed subject variability, MSDL outperforms only slightly SPCA for the estimation of population maps, but more clearly for the estimation of individual maps.

In Fig. 3, we report the likelihood of left-out data in a 3-fold cross-validation for varying model order, with a synthetic dataset comprising 5 spatial latent factors. We can can see that if the subject-specific observation noise is not smooth, and there is no spatial jitter across subjects, the likelihood reaches a plateau for a specified model order corresponding to the ground truth. However, if the

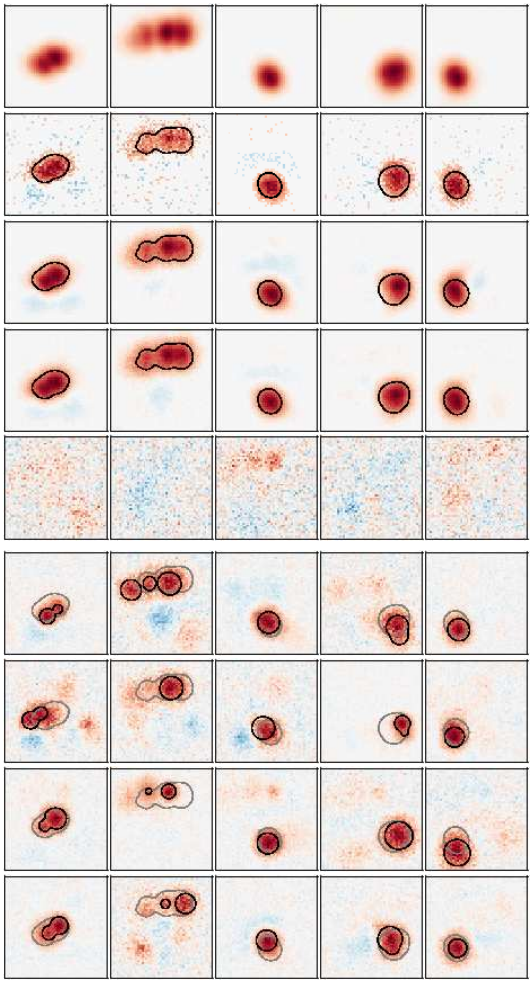


Fig. 1. Population-level maps: **V**.
a. Ground truth
b. Maps estimated by thresholded ICA, the ground truth is outlined in black.
c. Maps estimated by sparse PCA
d. Maps estimated by our multi-subject dictionary learning model
e. Raw observations

Fig. 2. Subject-specific maps estimated by our multi-subject dictionary learning model: each row corresponds to a different subject. The subject-specific ground truth is outlined in black, while the population-level ground truth is outlined in light gray.

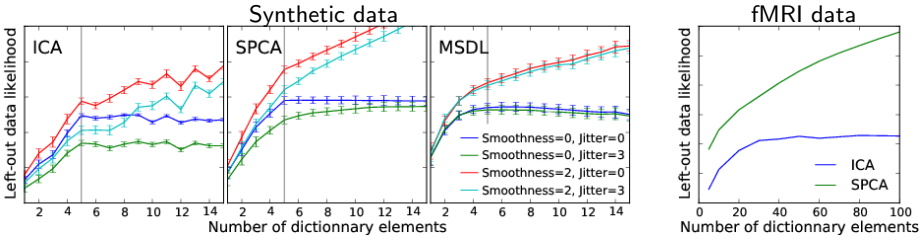


Fig. 3. Likelihood of left out data in a 3-fold stratified cross-validation with the different methods. Left: synthetic data for varying level of observation noise spatial smoothness and of subject-variability jitter; right: resting-state fMRI data.

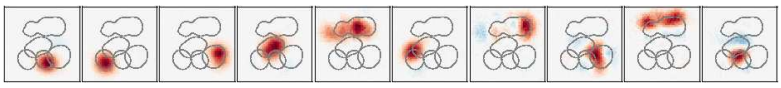


Fig. 4. High model order.

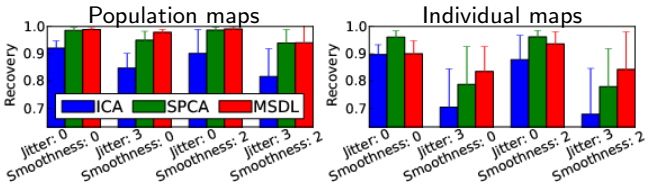


Fig. 5. Synthetic data: correlation between the recovered maps and the ground truth.

observation noise is spatially smooth, and thus violates the i.i.d. hypothesis, the likelihood always increases with model order. In the presence of between-subject spatial jitter, the maps learned are structured and correspond to spatial gradients of the blobs (see Fig 4), i.e. patterns that fit the between-subject variability.

6 An atlas of functional regions for spontaneous activity

We apply our multi-subject dictionary learning method to an fMRI dataset of 20 healthy subjects scanned twice in a resting task, eyes closed. Each session is comprised of 244 brain volumes acquired with a repetition time of 2.4 s. After correction for slice timing differences, motion correction and inter-subject spatial normalization using SPM5, we extract the time series on a mask of the brain, resulting in roughly 500 samples per subject, and 25 000 features – a dataset of 2 Go. We center and variance normalize the time series before dictionary learning.

We have measured by 3-fold cross validation the likelihood of left-out data as a function of model order for ICA and SPCA, but because of lack of time and computing resource not for MSDL (see Fig. 3). We find that the likelihood saturates at a model order of 40 with ICA, thus in the following we choose $k = 40$. In Fig. 7 we display an outline at 33% of the maximum value for all the dictionary elements learned using MSDL. We find that these spatial maps segment the major sources of signal in the Echo-Planar Imaging volumes, that is the gray-matter functional regions that generate the Blood Oxygen-Level Dependent signal (Fig. 6 a and b), as well as sources of artifacts such as the ventricles, the circle of Willis (Fig. 6 c), or the white matter (Fig. 6 d). The estimated maps separate well these salient features from the background as they are mostly sparse. This is in contrast with ICA maps, such as those presented in [9] that are most-often thresholded to a high-value, hiding a noisy background as can be seen in unthresholded maps reported by [26].

For functional studies without known experimental design, such as resting-state, the interesting point is that the gray matter is divided in a set of localized spatial maps that can be seen as an atlas of functional regions present in brain spontaneous fully estimated from the fMRI data. In addition, this division of the brain is consistent with current anatomo-functional knowledge. For instance, in the primary areas, it follows the known topographic maps (see Fig. 7). It is also interesting to note that the parietal lobe, considered as a major functional hub, yields the finest parcellation. As a reference, we also report the corresponding spatial maps estimated with ICA and SPCA on Fig. 8. We find that the maps estimated by ICA outline similar regions but display more high spatial frequency

noise with many different connected components. A few SPCA maps display a scattered, salt-and-pepper like sparsity pattern. This can be explained by the presence of highly spatially-correlated noise: the ℓ_1 penalization tends to choose only one voxel to represent a group of highly-correlated features.

In Fig. 9, we display the outline of the dictionary element corresponding to the Calcarine sulcus for 5 subjects, showing the population-level latent map in addition to the subject-specific map. We can see that the subject-specific map matches better the outline of a gray-matter segmentation –performed using SPM5 on anatomical T1 images– than the population-level map, although at no point anatomical scans were used in the dictionary-learning procedure.

7 Conclusion

The contributions of this work are two-fold. In a statistical learning context, we formulate a new problem, that of learning subject-specific latent factors that match an inter-subject model with explicit variability terms, and we give an efficient algorithm to solve this problem. Importantly, this algorithm separates the problem of learning loadings and subject-level latent factors, from denoising population-level latent factors using proximal operators. There is a rich literature concerned with proximal-based image denoising [8] and the problem we introduce should be further addressed with other proximal operators, such as total variation or structured sparsity [19] for which there exist efficient algorithms.

In a brain imaging context, we address the challenge of segmenting brain regions from spontaneous activity and pave the path for establishing a reference probabilistic functional atlas of this on-going brain activity. Unlike previous methods, our approach controls the amount of signal not fitted by the model at the subject level and at the population level. Using realistic simulations as well as resting-state fMRI data, we have shown that our procedure can *i) at the population level* extract contrasted spatial maps that are consistent with current anatomo-functional knowledge and *ii) at the individual level* adapt these maps to individual spatial configurations. Given the population-level atlas, individual maps for a new subject can be estimated at a small computational cost with the same subject-level model. This procedure thus gives a principled method for defining subject-specific regions to use in a functional-connectivity study, for instance to guide diagnosis or prognosis.

References

1. Archambeau, C., Bach, F.: Sparse probabilistic projections. Adv NIPS (2008)
2. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM Journal on Imaging Sciences 2(1), 183 (2009)
3. Beckmann, C.F., Smith, S.M.: Probabilistic independent component analysis for functional magnetic resonance imaging. Trans Med Im 23, 137 (2004)
4. Bell, A., Sejnowski, T.: An information-maximization approach to blind separation and blind deconvolution. Neur. Comp 7, 1129 (1995)

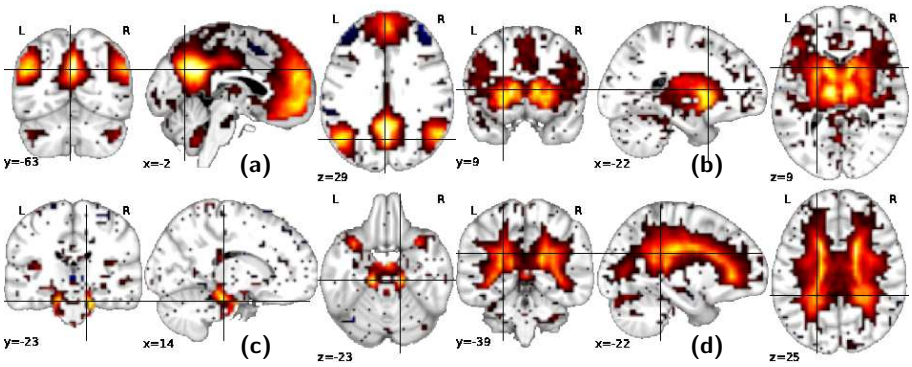
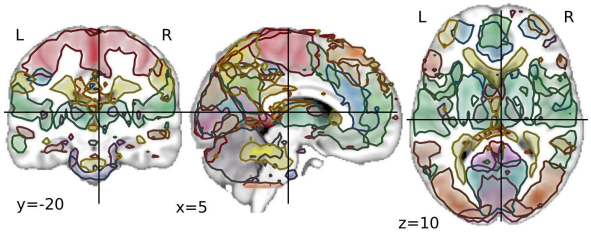


Fig. 6. Population-level maps estimated by MSDL: **(a)** default mode network, **(b)** thalamus, pallidum and caudate, **(c)** vascular system, including the circle of Willis, **(d)** white matter. The maps are not thresholded but reflect the sparsity pattern.

Fig. 7. Outlines at 33% of all dictionary elements estimated by MSDL for 2 different set of cutting planes. The motor system is divided in (**1**) dorsal, (**2**) lateral, and (**3**) ventral regions. Similarly, the visual system is divided in (**4**) a primary region centered on the Calcarine sulcus, overlapping with (**5**) a region centered on the striate cortex, and (**6**) extrastriate regions. (**7**), (**8**): fine details of the vascular system segmented in several maps.



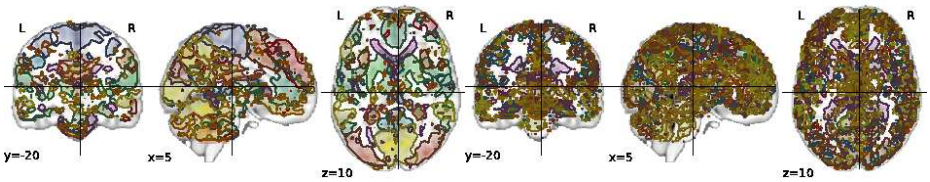
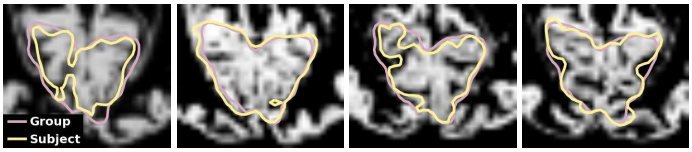


Fig. 8. Outlines at 33% of all ICA (left) and SPCA (right) dictionary elements.

Fig. 9. Dictionary element corresponding to the Calcarine sulcus overlaid on the gray matter map for different subjects.



5. Biswal, B., Mennes, M., Zuo, X., Gohel, S., Kelly, C., Smith, S., Beckmann, C., Adelstein, J., Buckner, R., Colcombe, S., et al.: Toward discovery science of human brain function. Proc Ntl Acad Sci 107, 4734 (2010)
6. Calhoun, V.D., Adali, T., Pearlson, G.D., Pekar, J.J.: A method for making group inferences from fMRI data using independent component analysis. Hum Brain Mapp 14, 140 (2001)
7. Chen, S., Donoho, D., Saunders, M.: Atomic decomposition by basis pursuit. SIAM review 43, 129 (2001)
8. Combettes, P., Pesquet, J.: A proximal decomposition method for solving convex variational inverse problems. Inverse Problems 24, 065014 (2008)
9. Damoiseaux, J.S., Rombouts, S.A.R.B., Barkhof, F., Scheltens, P., Stam, C.J., Smith, S.M., Beckmann, C.F.: Consistent resting-state networks across healthy subjects. Proc Natl Acad Sci 103, 13848 (2006)
10. Daubechies, I., Roussos, E., Takerkart, S., Benharrosh, M., Golden, C., D'Ardenne, K., Richter, W., Cohen, J.D., Haxby, J.: Independent component analysis for brain fMRI does not select for independence. Proc Natl Acad Sci 106, 10415 (2009)
11. Donoho, D.: De-noising by soft-thresholding. Trans Inf Theory 41, 613 (1995), http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=382009
12. Golland, P., Golland, Y., Malach, R.: Detection of spatial activation patterns as unsupervised segmentation of fMRI data. In: MICCAI (2007)
13. Hebiri, M., Van De Geer, S.A.: The Smooth-Lasso and other $\ell_1 + \ell_2$-penalized methods. ArXiv:1003.4885 (2010), http://arxiv.org/abs/1003.4885v1
14. Hyvärinen, A., Oja, E.: Independent component analysis: algorithms and applications. Neural Networks 13, 411 (2000)
15. Jenatton, R., Obozinski, G., Bach, F.: Structured sparse principal component analysis. In: Proc. AISTATS (2010)
16. Kettenring, J.R.: Canonical analysis of several sets of variables. Biometrika 58, 433 (1971)
17. Kreutz-Delgado, K., Murray, J., Rao, B., Engan, K., Lee, T., Sejnowski, T.: Dictionary learning algorithms for sparse representation. Neur. Comp. 15, 349 (2003)
18. Mairal, J., Bach, F., Ponce, J., Sapiro, G.: Online learning for matrix factorization and sparse coding. Journal of Machine Learning Research 11, 19 (2010)
19. Mairal, J., Jenatton, R., Obozinski, G., Bach, F.: Network flow algorithms for structured sparsity. Adv NIPS (2010)
20. McHugh, J.: Algorithmic Graph Theory. Prentice Hall (1990)
21. Minka, T.: Automatic choice of dimensionality for PCA. Adv NIPS p. 598 (2001)
22. Sigg, C., Buhmann, J.: Expectation-maximization for sparse and non-negative PCA. Proc. ICML (2008)
23. Smith, S., Fox, P., Miller, K., Glahn, D., Fox, P., Mackay, C., Filippini, N., Watkins, K., Toro, R., Laird, A., et al.: Correspondence of the brain's functional architecture during activation and rest. Proc Natl Acad Sci 106, 13040 (2009)
24. Tipping, M., Bishop, C.: Probabilistic principal component analysis. Journal of the Royal Statistical Society. Series B, Statistical Methodology p. 611 (1999)
25. Varoquaux, G., Baronnet, F., Kleinschmidt, A., Fillard, P., Thirion, B.: Detection of brain functional-connectivity difference in post-stroke patients using group-level covariance modeling. In: MICCAI (2010)
26. Varoquaux, G., Keller, M., Poline, J., Ciuciu, P., Thirion, B.: ICA-based sparse features recovery from fMRI datasets. In: ISBI. p. 1177 (2010)
27. Wainwright, M.: Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming. Trans Inf Theory 55, 2183 (2009)