Cohort-level brain mapping: learning cognitive atoms to single out specialized regions

G. Varoquaux[123], Y. Schwartz[12], P. Pinel[23], B. Thirion[12]

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

Abstract. Functional Magnetic Resonance Imaging (fMRI) studies map the human brain by testing the response of groups of individuals to carefully-crafted and contrasted tasks in order to delineate specialized brain regions and networks. The number of functional networks extracted is limited by the number of subject-level contrasts and does not grow with the cohort. Here, we introduce a new group-level brain mapping strategy to differentiate many regions reflecting the variety of brain network configurations observed in the population. Based on the principle of functional segregation, our approach singles out functionally-specialized brain regions by learning group-level functional profiles on which the response of brain regions can be represented sparsely. We use a dictionary-learning formulation that can be solved efficiently with on-line algorithms, scaling to arbitrary large datasets. Importantly, we model inter-subject correspondence as structure imposed in the estimated functional profiles, integrating a structure-inducing regularization with no additional computational cost. On a large multi-subject study, our approach extracts a large number of brain networks with meaningful functional profiles.

1 Introduction

Using fMRI, the systematic study of which areas of the brain are recruited during various experiments has led to accumulation of activation maps related to specific tasks or cognitive concepts in an ever growing literature. Mapping a given population requires careful crafting of a set of tasks that are contrasted to reveal networks. These networks form a natural representation of brain function and are of particular interest to study its variability in a population, for instance to correlate it to pathologies or genetic information. However, each subject can only perform a small number of tasks in a scanner; particularly so for disabled subjects. As a result, in a given study the number of networks that is identified by standard task-activation mapping is small and limited by the number of contrasts of the study. On the other hand, it is not uncommon to scan a large number of subjects. Indeed, clinical studies must often resort to larger sample sizes due to the intrinsic variability of pathologies. Massive cohorts can be acquired, *e.g.* to learn diagnosis markers for Alzheimer's disease [10], or in neuroimaging-genetics.

In large cohorts, a small set of contrasts reveals effects throughout the whole brain [16]. This observation suggests that more information can be extracted at

the cohort level. In this paper, we address precisely this challenge by decomposing brain activity and experimental conditions at the group level to assign a specific cognitive function to each voxel. For this purpose, inter-subject variability is a blessing as functional variability reveals *functional degeneracy*, *i.e.* that different networks sustain the same cognitive function across individuals [9]. However, this variability is also a curse when it arises from spatial realignment error.

Compressed spatial representations were put forward for group studies by Thirion *et al.* [15] using clustering of the activation maps. This early work did not address the functional specificity of the clusters. Conversely, Lashkari *et al.* [7] discard spatial information and focus on extracting common functional profiles across subject, removing the need for spatial normalization. Following this idea of functional correspondence across subject, although not leading to the definition of regions, Sabuncu *et al.* [12] use this correspondence for inter-subject alignment. Linear models such as independent component analysis (ICA) have been used to extract modes of brain function across subjects [2] before clustering approaches. Laird *et al.* [6] have recently shown that the modes that it extracts from task-activation data capture meaningful structure in the space of cognitive processes. Beyond ICA, Varoquaux *et al.* [18] use dictionary learning to segment a functional parcellation from resting-state. Very interesting preliminary work by Chen *et al.* [3] integrates spatial normalization with dictionary learning to estimate jointly an inter-subject warping and functional regions.

The present paper combines ideas from this prior art in a new inter-subject model with an associated computationally-scalable estimation algorithm. Our contributions are *i)* a joint model of the position and functional tuning of brain networks, *ii)* explicit separation of the variance into intra-subject and inter-subject components, *iii)* a fast and scalable algorithm that can impose this particular variance structure. We show with simple simulations that controlling inter-subject variance is crucial, as unsupervised learning approaches such as dictionary learning or clustering will fit this variance and extract modes reflecting inter-subject variability. The paper is organized as followed. We start by giving a multi-subject model combining random effects (RFX) with functional segregation hypothesis. In section 3, we introduce an on-line and computationally-efficient algorithm to estimate this model. In section 4, we present a simulation study, and in section 5 results on an fMRI dataset comprising 150 subjects.

2 A multi-subject sparse-coding model of brain response

Sparse coding brain response Our model is based on two basic neuroscience principles: *i) functional segregation* which states that brain territories are formed of elementary, functionally-specific units [17] and *ii) functional degeneracy* which states that a particular function may recruit different networks across subjects [9]. We combine these principles at the subject and group level to learn the correct basis to describe the macroscopic level of brain organization.

Experimental stimuli and contrasts do not correspond simply to elementary cognitive processes. For instance to isolate brain regions involved in a calcula-

tion tasks, instructions to perform arithmetics will be given to a subject, however these instructions are given via a modality: auditory or visual, and will induce a word-comprehension task in addition to the calculation. Investigators use *contrast maps* to cancel out secondary effects and focus on *word – calculations*, but these contrasts can carry also some auditory, visual, or language effects as the stimuli content in the different tasks are not perfectly matched.

A typical fMRI experiment thus yields a set of task-specific contrast maps: for each subject s, $\boldsymbol{X}^s \in \mathbb{R}^{t \times n}$, where t is the number of tasks and n the number of voxels. Based on the principles of functional specialization, we stipulate that the tasks used are formed of elementary cognitive processes associated with a set of corresponding sparse neural substrates: there exist combinations of tasks $\boldsymbol{D} = \{\boldsymbol{d}_j\}$ such that each $\boldsymbol{d}_j$ is expressed on a small number of brain regions:

$$\boldsymbol{X}^s = \boldsymbol{D}\boldsymbol{A}^{s\mathsf{T}}, \qquad \text{where } \boldsymbol{A}^s \text{ is sparse.} \tag{1}$$

We are interested in learning a dictionary of k functional profiles $\boldsymbol{D} \in \mathbb{R}^{t \times k}$ and the associated sparse spatial code $\boldsymbol{A}^s \in \mathbb{R}^{n \times k}$, that we call *functional networks*. The number of atomic cognitive functions recruited by the tasks explored in an fMRI experiment is most likely much larger than the number of experimental conditions t. Drawing from a large number of subjects can help to estimate more functional profiles, as subjects will resort to different *cognitive strategies*, engaging differently atomic cognitive functions. To give a clichéd image, right-handed and left-handed subjects could rely on different visuo-spatial representations to perform a hand motion task. In practice, variability in cognitive strategy is often very subtle and can be related to variability in attention, engagement to the task, background processes, rather than high-level strategies [9]. Modeling this inter-subject variability should improve the quantification of population-level estimates and enable the separation of atoms of brain function.

Multi-subject modeling We introduce subject-specific expressions of the functional profiles:

$$\boldsymbol{F}^s = (\boldsymbol{I} + \boldsymbol{\Delta}^s)\boldsymbol{D}, \qquad \text{where } \boldsymbol{\Delta}^s \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}_t), \;\; \boldsymbol{\Delta}^s \in \mathbb{R}^{t \times t}, \;\; \boldsymbol{F}^s \in \mathbb{R}^{t \times k} \tag{2}$$

An approach commonly used when dealing with such unsupervised learning problem on multi-subject fMRI data is to concatenate the data spatially [2, 15], learning an augmented dictionary,

$$\overline{\boldsymbol{F}} \;=\; \left[\boldsymbol{F}^{1\mathsf{T}} \dots \boldsymbol{F}^{s\mathsf{T}}\right]^{\mathsf{T}} \;=\; \left[(\boldsymbol{I} + \boldsymbol{\Delta}^1)^{\mathsf{T}}, \dots (\boldsymbol{I} + \boldsymbol{\Delta}^s)^{\mathsf{T}}\right]^{\mathsf{T}} \boldsymbol{D} \;\in \mathbb{R}^{st \times k}. \tag{3}$$

The multi-subject model resulting from (1) and (2) can then be written as a standard dictionary-learning problem: $\overline{\boldsymbol{X}} = \overline{\boldsymbol{F}}\boldsymbol{A}^{\mathsf{T}}$, with $\overline{\boldsymbol{X}} \in \mathbb{R}^{st \times n}$ the spatial concatenation of the data and $\boldsymbol{A}$ functional networks independent of the subject. By learning a dictionary spanning multiple datasets, it can estimate inter-subject loadings that reveal the different cognitive strategies, drawing from the *spatial correspondence* of the coding of the information. However, estimating high-dimensional dictionaries has two major drawbacks: *i)* it is more challenging

from the statistical standpoint because the residuals implicit in eq. 3 are non white and *ii)* this approach is fragile to errors in inter-subject correspondence.

To remove the need for spatial matching, Lashkari *et al.* [7] cluster the activity profiles, grouping voxels that respond similarly to the tasks across subjects. This *functional correspondence* hypothesis leads to a functional concatenation of the data: $\underline{\boldsymbol{X}} = [\boldsymbol{X}^1, \dots \boldsymbol{X}^{s\mathsf{T}}]^\mathsf{T} \in \mathbb{R}^{t \times sn}$. The multi-subject model is then written $\underline{\boldsymbol{X}} = \boldsymbol{D}\underline{\boldsymbol{A}}^\mathsf{T}$ with $\underline{\boldsymbol{A}} = [(\boldsymbol{I}_k + \boldsymbol{\Delta}^1)\boldsymbol{A}^1, \dots (\boldsymbol{I}_k + \boldsymbol{\Delta}^s)\boldsymbol{A}^{s\mathsf{T}}]^\mathsf{T} \in \mathbb{R}^{k \times sn}$, which amounts to learning a dictionary common to all subjects and different spatial maps.

Modeling Random effects Both spatial and functional concatenation approaches lead to a simple formulation in terms of learning a dictionary of functional profiles and spatial code. However a naive resolution of these dictionary learning problems neglects that both spatial code and functional profiles share information across subjects. In functional neuroimaging data analysis, the standard way to model both common effects and variability across datasets relies on hierarchical linear models, often mixed- or random-effects (RFX) models that assume that the effect has two components of variance: inter-subject and intra-subject [19]. We can adapt this model to enhance the spatial correspondence approach by constraining the ratio of the intra- and inter-subject variance of the functional profiles in the augmented dictionary $\overline{\boldsymbol{F}}$. For this purpose, we introduce a *common effect matrix* made of s $k \times k$ identity matrices concatenated: $\boldsymbol{C} = \frac{1}{s}[\boldsymbol{I}_k, \dots \boldsymbol{I}_k]^\mathsf{T} \in \mathbb{R}^{k \times sk}$ and the *differential effects matrix* $\boldsymbol{C}_\perp \in \mathbb{R}^{(s-1)k \times sk}$, which is an orthogonal completion of $\boldsymbol{C}$. To impose an RFX structure on the dictionary, we present in section 3 an algorithm controlling $\|\overline{\boldsymbol{f}}_i\boldsymbol{C}\|_2^2 / \|\overline{\boldsymbol{f}}_i\boldsymbol{C}_\perp\|_2^2$, where $i \in [1, t]$ is the index of a dictionary element.

Proposition 1. *$\boldsymbol{C}$ and $\boldsymbol{C}_\perp$ isolate* i) *group-level profiles:* $\mathbb{E}\big[\overline{\boldsymbol{f}}_i\boldsymbol{C}\big] = \boldsymbol{d}_i$,
ii) *intra-subject variance:* $\mathbb{E}\big[\|\overline{\boldsymbol{f}}_i\boldsymbol{C}\|_2^2\big] = \big(1 + \frac{\sigma^2}{s}\big)\|\boldsymbol{d}_i\|_2^2 \sim \|\boldsymbol{d}_i\|_2^2$,
iii) *inter-subject variance:* $\mathbb{E}\big[\|\overline{\boldsymbol{f}}_i\boldsymbol{C}_\perp\|_2^2\big] = \big(\sigma^2 - \frac{\sigma^2}{s}\big)\|\boldsymbol{d}_i\|_2^2 \sim \sigma^2\|\boldsymbol{d}_i\|_2^2$.

The first and the second equalities stem from Eq. (3), while the last one follows from the fact that $\|\overline{\boldsymbol{f}}_j\|_2^2 = \|\overline{\boldsymbol{f}}_j\boldsymbol{C}\|_2^2 + \|\overline{\boldsymbol{f}}_j\boldsymbol{C}_\perp\|_2^2$, as $[\boldsymbol{C}^\mathsf{T}, \boldsymbol{C}_\perp^\mathsf{T}]$ forms a basis of $\mathbb{R}^{sk}$.

3 Efficient learning of RFX-structured dictionaries

State-of-the-art dictionary learning algorithm A general approach to learn dictionaries for sparse coding is to optimize the dictionary so that is leads to a sparse regression on train data, using an ℓ_1 penalty on the code [8]:

$$\hat{\boldsymbol{D}} = \underset{\boldsymbol{A}, \boldsymbol{D},\ \boldsymbol{D} \in \mathcal{C}}{\operatorname{argmin}} \|\boldsymbol{X} - \boldsymbol{D}\boldsymbol{A}^\mathsf{T}\|_2^2 - \lambda\|\boldsymbol{A}\|_1, \tag{4}$$

where $\boldsymbol{X}, \boldsymbol{D}, \boldsymbol{A}$ should be replaced by $\overline{\boldsymbol{X}}, \overline{\boldsymbol{F}}, \boldsymbol{A}$ or $\underline{\boldsymbol{X}}, \boldsymbol{D}, \underline{\boldsymbol{A}}$ depending on the choice of correspondence. Note that the dictionary $\boldsymbol{D}$ is constrained to a convex set $\mathcal{C}$, typically by bounding the ℓ_2 norm of its atoms: $\|\boldsymbol{d}_i\|_2 \leq 1$. This constraint is technical, as without it the penalty on $\boldsymbol{A}$ could be made arbitrarily small by

scaling up $\boldsymbol{D}$ and down $\boldsymbol{A}$ and thus keeping the data-fit term constant. Let us rewrite the optimization problem:

$$\hat{\boldsymbol{D}} = \underset{\boldsymbol{D},\, \boldsymbol{D}\in\mathcal{C}}{\operatorname{argmin}} \sum_{v} \min_{\boldsymbol{a}_v} \left(\left\| \boldsymbol{x}_v - \boldsymbol{D}\boldsymbol{a}_v^{\mathsf{T}} \right\|_2^2 + \lambda \|\boldsymbol{a}_v\|_1 \right). \tag{5}$$

This new expression highlights that, when learning the dictionary, the objective function is the sum over a large number of different realizations of the same problem, here sparse coding a simple voxel activation profile $\boldsymbol{x}_v$. The optimization problem can thus be tackled using stochastic gradient descent with on-line or mini-batch strategies [8]: small numbers of voxels randomly drawn from the data are successively considered and a corresponding sparse code $\boldsymbol{a}_v$ is learned by solving a Lasso-type problem. The dictionary can then be updated to minimize the data-fit error given the code. The algorithm iterates over small batches of voxels (hundreds) to incrementally improve the dictionary. When the number of voxels is large, such an approach can be orders of magnitude faster than the alternate optimization strategies used by [18, 3], because these require solving brain-wide sparse regression for each update of the dictionary.

Szabo *et al.* [14] extend this approach to structured dictionaries by replacing the ℓ_1 norm on $\boldsymbol{\alpha}_v$ with a structure-inducing norm, such as the ℓ_{21} norm used in the group lasso. However, the corresponding algorithms to learn the sparse code $\boldsymbol{a}_v$ are much more costly as they rely in general on optimizing augmented problems over auxiliary variables [14]. On the opposite, efficient algorithms to solve the ℓ_1 problem benefit from the sparsity of the solution and can be much less costly than a least-square estimate for very sparse problems [4].

Imposing RFX-structured dictionaries We introduce a simple modification to the on-line algorithm [8] to impose an RFX structure on the dictionary. Our approach is based on spatial correspondence to learn an augmented dictionary $\overline{\boldsymbol{F}}$ and sets different intra and inter-subject variance using proposition 1: controlling the ratio of the norm of $\overline{\boldsymbol{F}}\boldsymbol{C}$ and $\overline{\boldsymbol{F}}\boldsymbol{C}_\perp$. For this purpose, we use a careful choice of constraint set $\mathcal{C}$ on the dictionary; namely, we impose on each atom

$$\Omega(\overline{\boldsymbol{f}}_i) \leq 1, \qquad \text{with } \Omega(\overline{\boldsymbol{f}}_i) = \max(\|\overline{\boldsymbol{f}}_i\boldsymbol{C}\|_2^2,\ \mu\|\overline{\boldsymbol{f}}_i\boldsymbol{C}_\perp\|_2^2), \tag{6}$$

where μ controls the ratio of intra to inter subject variance. Because of the penalty on $\boldsymbol{A}$, it is highly likely that the constraint will be saturated. This constraint is an ℓ_∞ norm, which tends to enforce equality when saturated[4]: $\|\overline{\boldsymbol{f}}_i\boldsymbol{C}\|_2^2 = \mu\|\overline{\boldsymbol{f}}_i\boldsymbol{C}_\perp\|_2^2$.

In the on-line dictionary learning algorithm, this constraint is enforced by an Euclidean projection (see algorithm 2 of [8]): at each iteration

$$\boldsymbol{d}_{n+1} \leftarrow \underset{\boldsymbol{d}}{\operatorname{argmin}} \|\boldsymbol{d}_n + \boldsymbol{d}\|_2^2 \quad \text{subject to } \Omega(\boldsymbol{d}) \leq 1. \tag{7}$$

[4] Indeed, combined with an ℓ_2 loss, an ℓ_∞ constraint tends to saturate at its *kinks* enforcing equality between variables, as an ℓ_1 constraint enforces sparsity.

The *max* operator in Ω imposes that $\|\overline{\boldsymbol{f}}_i \boldsymbol{C}\|_2^2 \leq 1$ and $\|\overline{\boldsymbol{f}}_i \boldsymbol{C}_\perp\|_2^2 \leq \frac{1}{\mu}$. As $\boldsymbol{C}$ and $\boldsymbol{C}_\perp$ span orthogonal subspaces, the Euclidean distance decomposes in two independent optimization problems on those subspaces: the projection on a ball of radius 1 (resp. $\frac{1}{\mu}$), $\boldsymbol{c}_{n+1} \leftarrow \boldsymbol{c}_n / \|\boldsymbol{c}_n\|_2$, where $\boldsymbol{c}$ is the restriction of $\boldsymbol{d}$ to the subspace spanned by $\boldsymbol{C}$ (resp. $\boldsymbol{C}_\perp$). In practice, to implement this projection, we apply the dictionary-update algorithm after rotating the dictionary and the code to express them in the basis of $\mathbb{R}^{sk}$ formed by $[\boldsymbol{C}^\mathsf{T}, \boldsymbol{C}_\perp^\mathsf{T}]$, and for the sparse-coding step, we rotate back the dictionary to the basis that leads to sparse codes. With this strategy, the Euclidean projection Eq. (7) has the same computational cost with norm Ω than with the standard ℓ_2 norm proposed in [8]. As the computational cost of the dictionary update step is already quadratic in the length of the atoms, this strategy to impose an RFX structure on the dictionary does not change the overall algorithmic complexity of dictionary learning, neither asymptotically nor for small dictionaries.

Parameter choice and initialization Our algorithm has two parameters: λ, that controls the sparsity of the spatial maps, and μ that controls the ratio of intra-subject to inter-subject variance. We set that ratio to 10. Typically in fMRI study, inter-subject variance is 4 to 9 times larger than intra-subject variance [19], thus we are over-penalizing. However, in statistics, over-penalization is considered as preferable to under-penalization, as the former leads to bias, here to a common effect, while the later can easily lead to an explosion of variance. With regards to λ, the natural scaling factor is $\lambda \propto \frac{1}{\sqrt{p}}\varepsilon$ where p is the size of the atoms, and ε^2 the variance of the residuals [1]. We assume that $\varepsilon \propto \operatorname{std} \boldsymbol{X}$ and use the simple choice $\lambda = \frac{1}{\sqrt{p}} \operatorname{std} \boldsymbol{X}$. Similar scalings are suggested in [8]. They lead to having a number of non-zero constant on average in the code $\boldsymbol{A}$. In other words, each voxel is coded on the same number of maps, independently of the size of the problem (number of maps extracted, number of contrasts).

The dictionary learning problem is not convex. The starting point is important because a good choice can significantly speed up the convergence, and also determine the final results. We use spatially-constrained clustering on spatially-concatenated data [15] to learn an initial parcellation and associated dictionary.

4 Results on simulated data

Synthetic data generation We generate a simple and well-understood synthetic dataset to illustrate how the different approaches work, as well as the impact of spatial variability. We study the scenario in which two observed contrasts are generated from three functional networks, each one of them made of a single blob (Fig. 1, top left). Group-level loadings are generated from a uniform $[0, 1]$ distribution, and for each subject one cognitive strategy out of two, corresponding to a variation in 20% of the weights, is affected randomly. Finally, Gaussian-distributed noise is added with a variance of 0.1. We generate images of size 50×50 for 32 subjects. Optionally, we add spatial variability across subjects with Gaussian noise of 3 pixel standard deviation on the positions of the blobs.

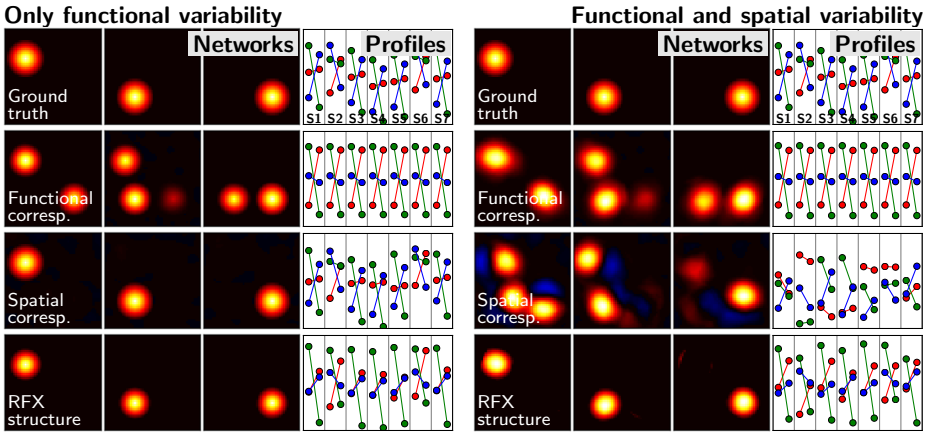

Fig. 1. Simulations: functional networks and subject-level profiles as estimated by different dictionary learning strategies – right column: with only functional variability – left column: with spatial variability. On the ground-truth profile plot the second cognitive strategy can be seen from the red loadings in the second and sixth subjects.

Results Without spatial variability, spatial correspondence and RFX-structure are very successful at singling out the blobs, however the functional correspondence strategy is less so (see Fig. 1). This is not surprising, as in the functional correspondence case, the dictionary learning task amounts to separating out 3 vectors (functional profiles) in a 2-dimensional space, which corresponds to an under-determined source separation problem. The under-determined problem is much harder than the over-determined problem, as in the spatial correspondence approach. Indeed, learning an augmented dictionary across subjects can benefit from inter-subject functional variability to tease out networks. However, in the presence of spatial variability, the simple spatial correspondence fits this variability and the estimated maps exhibit *adjustment* modes, combining different networks with negative regions that correspond to network spatial derivatives. Indeed, the loadings show little consistency across subjects, as the spatial maps learning are combined to compensate for spatial fluctuations. The RFX structure prevents such a combination to happen via a shrinkage to common factors. As a result the spatial maps are more faithful to the true networks. Note that the inter-subject profiles are overly shrunk. This an expected consequence of strong regularization: suppressing the variance comes to the cost of a bias. However this bias is not detrimental to the mean profile or the spatial maps.

5 Learning a cognitive brain atlas from fMRI

Functional localizer dataset We use a functional localizer that targets a wide spectrum of cognitive processes, namely visual, auditory and sensorimotor processes, as well as reading, language comprehension and mental calculation. This

protocol [11] lasts only 5 minutes, in order to be performed routinely on top of other protocols. We use 151 subjects that were acquired on the same 3T SIEMENS Trio scanner. 6 contrast maps best represent the brain activity for the cognitive processes recruited in this protocol. The contrast maps are both a combination of several conditions (e.g., sentence reading, calculation), and a difference of those conditions (e.g., right click versus left click) to draw out the effect of interest. For instance, the map "calculation - words" aims to isolate the effect of calculation by canceling out the modality of the stimulus (auditory or visual), and the residual effect of the comprehension of the stimulus (reading or listening). The effect of words is then encoded by negative loadings.

Networks and profiles extracted Fig. 2 shows some functional networks and profiles extracted using $k = 50$. The profiles are represented by their loadings on the contrasts of the original experiment, that oppose one type of brain function to another. Some networks extracted correspond across methods: for instance the network corresponding to a left click (a1, b1 and c1), for which the spatial map highlights the hand area of the motor cortex and the functional profiles are concentrated on the motor and left contrasts. As finger movement gives very strong activations, this network is reliable across subjects: standard errors on contrast loadings are small and the inter-subject functional profiles (Fig. 3) are similar across subjects even without enforcing structure. Note that a similar right-click network is also extracted (not shown). Extracting such a network is no surprise, as it maps well to a task performed in the study. More interestingly, networks corresponding to higher-level cognition are also extracted, *e.g.* the language network (a2, b2 and c2) and the dorsal-attentional network (a3, a4 and, b3 and c3), or a salience network (a4) [13]. We report a qualitative comparison of all the networks extracted for the different multi-subject approaches. As in the simulations, some maps learned by spatial correspondence have loadings that are not reproducible across subjects (b4 on Fig. 2 –note the large error bars– and on Fig. 3). Functional correspondence tends to mix well-known networks and produce degenerate maps. For instance, it extracts for the dorsal-attentional attentional network two components (a3 and a4) that are not well differentiated and include other regions. Indeed, the dorsal-attentional network is made of the intra-parietal sulci and the frontal eye fields and is well known for high-level visuo-spatial tasks, for instance during eye saccades. Maps a4 and a3 also outline the visual area MT (V5) and the dorsal ACC, part of respectively the visual system and the salience network. The corresponding functional profiles indeed stray away from the accepted functions of this network: a3 does not present any visual loading, while a4 shows right motor clicks and a preference for horizontal checkerboards. On the opposite, the RFX-structure approach selects only the frontal eye field and the intra-parietal sulci on the spatial map. The cognitive loadings are limited to visual and calculation tasks. While it may seem surprising to find calculation in a visuo-spatial network, this specific network has recently been reported as recruited in mental arithmetics [5]. Finally, we find that all the networks extracted by the RFX-structure approach outline known structure and have sensible cognitive loadings.

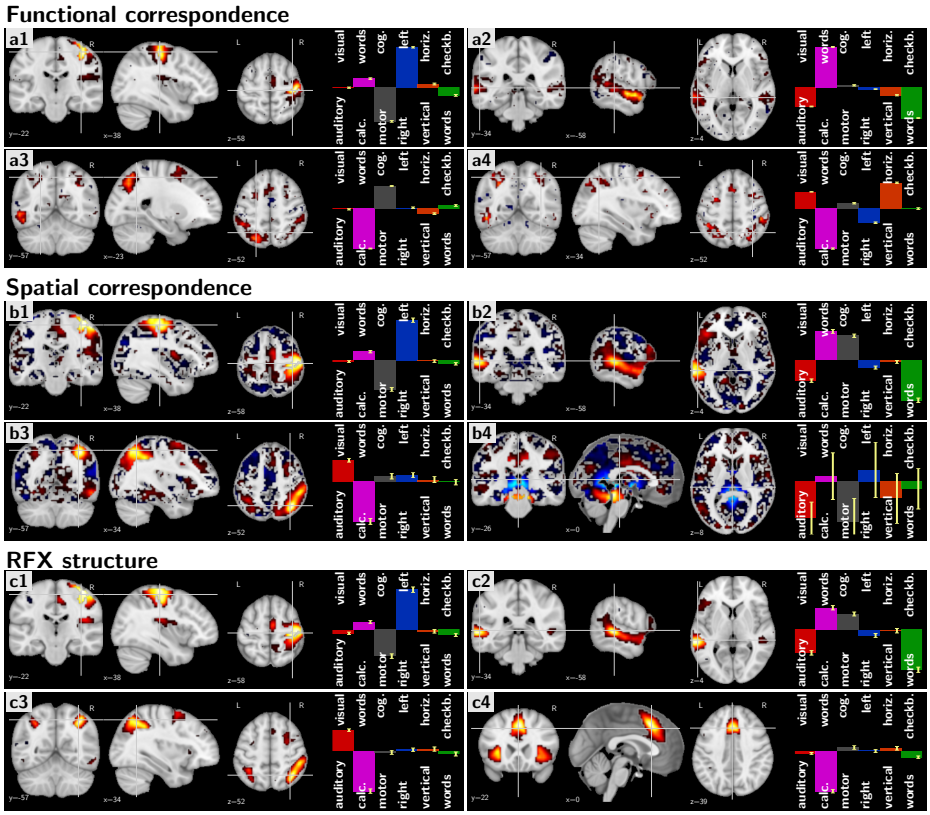

Fig. 2. Networks learned on the localizer dataset with different strategies. Each box represents the functional network and the group-level profile as loadings on the contrasts of the study: auditory - visual, calculation - word, motor - cognition, right click - left click, vertical checkerboard - horizontal checkerboard, and words - checkerboard. The standard error across the group is displayed as a yellow bar for each loading. **a1**, **b1** and **c1** correspond to the left hand region of the motor cortex, **a2**, **b2** and **c2** to the language network, **a3**, **a4**, **b3, c3** to the dorsal-attentional network, and **c4** to a salience network. **b4** is likely a noise pattern.

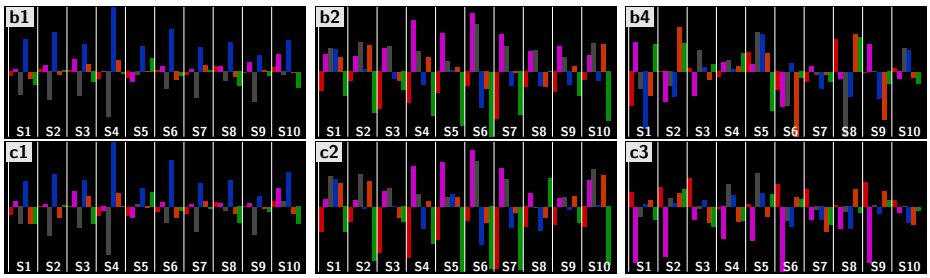

Fig. 3. Inter-subject functional profiles $\underline{\boldsymbol{D}}$ for the first 10 subjects, for spatial correspondence –top row– and RFX structure –bottom row. A white line separates subjects.

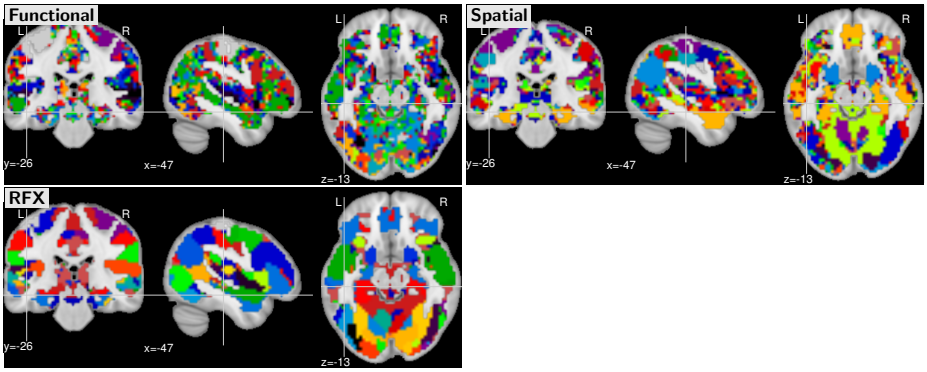

Fig. 4. Parcellations for the different strategies. The colors are random.

Towards a cognitive brain atlas To evaluate the overall spatial layout of the networks extracted we turn the decomposition in a hard assignment: we assign each voxel to the component for which it has the highest value in the spatial map. This procedure retrieves a cognitive label for each voxel and thus establishes a cognitively-informed brain parcellation. The maps extracted by functional correspondence often lack spatial structure and segment redundant regions across the different components (as with a3 and a4), as a result the corresponding parcellation appears noisy (see Fig. 4). The parcellations for spatial correspondence show more regularity, and even more so for the RFX-structured approach. The later gives sensible divisions of well-known parts of the cortex, such as the motor cortex, or the ventral visual stream.

Functional richness of the profiles The corresponding functional profiles are summarized by computing the t-value (mean effect divided by standard error) per network and contrast, across subjects. These values, clipped to [-10, 10], are presented in Fig. 5(left), which shows that the RFX model achieves an intermediate level of sensitivity between spatial correspondence, that yields smaller t values, and functional correspondence that exhibits high t-values.

A way of assessing the functional significance of these decompositions is to quantify how specific the encoding of functional profiles into networks is. To do so, we label each network as showing negative, none or positive activation, by thresholding the t values, and compute the entropy of the resulting assignment. Fig 5 (right) presents the results for a standard range of thresholds, obtained through 100 bootstrap replications of the t values and entropy computation. In a range of values that is usable in practice (t values between 2. and 4.) the RFX model yields a more efficient encoding than the other decompositions; the spatial decomposition dominates for very low t-values while the functional decomposition outperforms the others for extremely high t values. Altogether, this suggests that the RFX model encodes efficiently the possible functional profiles, while the spatial model is more sensitive to between-subject variability and the

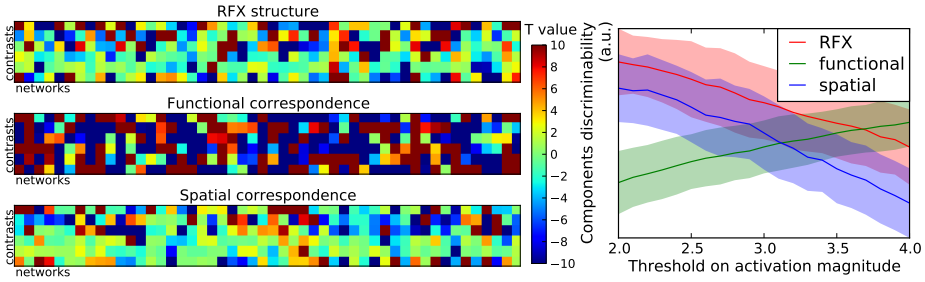

Fig. 5. Extracted functional profiles. (Left) These profiles summarize the functional activation per network (columns) and contrast (lines) of interest through a t-value per network and contrast, across subjects. The contrasts are identical to those in Fig. 2. The color scale, clipped to [-10, 10], shows that the RFX model achieves an intermediate level of sensitivity. (right) The specificity of the encoding of cognitive contrasts into networks is summarized by the entropy of an assignment to negative, none or positive activation: for most thresholds the RFX model yields the most efficient encoding.

functional model underestimates the group-level variance and thus overestimates the functional specificity of brain networks.

6 Conclusion

We have introduced a multi-subject model for task-induced fMRI activations that combines the principles of functional segregation and inter-subject degeneracy in a structured sparse coding problem. Technically, a major contribution of our formulation is to bound the ratio of inter-subject to intra-subject variance as it prevents extracting maps from non-reproducible variability. On a mid-sized cohort (150 subject, 6 contrasts) our model extracts a large number brain networks that are meaningful both in terms of cognitive content and of spatial maps. Applying this approach to larger studies should reveal richer and more specific effects. For larger cohorts, it can easily be extended to multi-level model specification, for instance in multi-centric studies, adding a center effect. An exciting direction of future research is to use this possibility to combine multiple studies in a meta analysis. Importantly, our approach is very computationally efficient: it is $\mathcal{O}\left(n^2\right)$ in the number of subjects, and the analysis presented in this paper runs in 10 mn on a single CPU, compared to several hours for non on-line learning. It is thus applicable to mining of massive datasets. Altogether, our results provide the basis of a framework to derive a synthetic and optimized representation of large amount of multi-subject fMRI data in terms of specialized brain regions.

This work was supported by the ANR grants BrainPedia ANR-10-JCJC 1408-01 and IRMGroup ANR-10-BLAN-0126-02.

References

1. Bickel, P., Ritov, Y., Tsybakov, A.: Simultaneous analysis of lasso and dantzig selector. The Annals of Statistics 37, 1705 (2009)
2. 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)
3. Chen, G., Fedorenko, E., Kanwisher, N., Golland, P.: Deformation-invariant sparse coding for modeling spatial variability of functional patterns in the brain. Machine Learning and Interpretation in Neuroimaging p. 68 (2012)
4. Efron, B., Hastie, T., Johnstone, I., Tibshirani, R.: Least angle regression. Ann. Statist. 32, 407 (2004)
5. Knops, A., Thirion, B., Hubbard, E., Michel, V., Dehaene, S.: Recruitment of an area involved in eye movements during mental arithmetic. Science 324, 1583 (2009)
6. Laird, A., Fox, P., Eickhoff, S., et al.: Behavioral interpretations of intrinsic connectivity networks. J Cog Neurosci 23, 4022 (2011)
7. Lashkari, D., Golland, P.: Exploratory fMRI analysis without spatial normalization. In: Information Processing in Medical Imaging. p. 398 (2009)
8. Mairal, J., Bach, F., Ponce, J., Sapiro, G.: Online learning for matrix factorization and sparse coding. Journal of Machine Learning Research 11, 19 (2010)
9. Noppeney, U., Friston, K., Price, C.: Degenerate neuronal systems sustaining cognitive functions. J Anat 205, 433 (2004)
10. Petersen, R., Aisen, P., Beckett, L., et al.: Alzheimer's disease neuroimaging initiative (ADNI) clinical characterization. Neurology 74, 201 (2010)
11. Pinel, P., Thirion, B., Meriaux, S., Jobert, A., Serres, J., Le Bihan, D., Poline, J., Dehaene, S.: Fast reproducible identification and large-scale databasing of individual functional cognitive networks. BMC neuroscience 8, 91 (2007)
12. Sabuncu, M., Singer, B., Conroy, B., Bryan, R., Ramadge, P., Haxby, J.: Function-based intersubject alignment of human cortical anatomy. Cereb Cortex 20, 130 (2010)
13. Seeley, W., Menon, V., Schatzberg, A., Keller, J., Glover, G., Kenna, H., Reiss, A., Greicius, M.: Dissociable intrinsic connectivity networks for salience processing and executive control. J neurosci 27, 2349 (2007)
14. Szabó, Z., Póczos, B., Lorincz, A.: Online group-structured dictionary learning. In: CVPR. p. 2865 (2011)
15. Thirion, B., Flandin, G., Pinel, P., Roche, A., Ciuciu, P., Poline, J.: Dealing with the shortcomings of spatial normalization: Multi-subject parcellation of fMRI datasets. Hum brain map 27, 678 (2006)
16. Thyreau, B., Schwartz, Y., Thirion, B., et al.: Very large fMRI study using the imagen database: Sensitivity–specificity and population effect modeling in relation to the underlying anatomy. NeuroImage 61, 295 (2012)
17. Tononi, G., McIntosh, A., Russell, D., Edelman, G.: Functional clustering: Identifying strongly interactive brain regions in neuroimaging data. Neuroimage 7, 133 (1998)
18. Varoquaux, G., Gramfort, A., Pedregosa, F., Michel, V., Thirion, B.: Multi-subject dictionary learning to segment an atlas of brain spontaneous activity. In: Inf Proc Med Imag. p. 562 (2011)
19. Worsley, K., Liao, C., Aston, J., Petre, V., Duncan, G., Morales, F., Evans, A.: A general statistical analysis for fMRI data. NeuroImage 15, 1 (2002)