Extracting Brain Regions from Rest fMRI with Total-Variation Constrained Dictionary Learning

Alexandre Abraham[1,2], Elvis Dohmatob[1,2], Bertrand Thirion[1,2], Dimitris Samaras[3,4], and Gael Varoquaux[1,2]

[1] Parietal Team, INRIA Saclay-Île-de-France, Saclay, France
alexandre.abraham@inria.fr
[2] CEA, DSV, I^{2}BM, Neurospin bât 145, 91191 Gif-Sur-Yvette, France
[3] Stony Brook University, NY 11794, USA
[4] Ecole Centrale, 92290 Châtenay Malabry, France

Abstract. Spontaneous brain activity reveals mechanisms of brain function and dysfunction. Its population-level statistical analysis based on functional images often relies on the definition of brain regions that must summarize efficiently the covariance structure between the multiple brain networks. In this paper, we extend a network-discovery approach, namely dictionary learning, to readily extract brain regions. To do so, we introduce a new tool drawing from clustering and linear decomposition methods by carefully crafting a penalty. Our approach automatically extracts regions from rest fMRI that better explain the data and are more stable across subjects than reference decomposition or clustering methods.

Keywords: dictionary learning, clustering, resting state fMRI.

1 Introduction

The covariance structure of functional networks, observed at rest using functional Magnetic Resonance Imaging (fMRI) signals, is a promising source of diagnostic or prognostic biomarkers, as it can be measured on several impaired subjects, such as stroke patients [12]. However, statistical analysis of this structure requires the choice of a reduced set of brain regions [12]. These should *i)* cover the main resting-state networks [3,15]; *ii)* give a faithful representation of the original signal, *e.g.* in the sense of compression or explained variance; *iii)* be defined in a way that is resilient to inter-subject variability.

Independent Component Analysis (ICA) is the reference method to extract underlying networks from rest fMRI [3]. Promising developments rely on penalized dictionary learning to output more contrasted maps [13]. However, if the maps highlight salient localized features, post-processing is required to extract connected regions. [8] use ICA maps to manually define this parcellation from resting-state networks. A complementary approach is to rely on voxel clustering that creates hard assignments rather than continuous maps [15,4].

This paper bridges the gap between the two strategies. The main contributions are *i)* the adaptation of dictionary learning to produce well-formed brain

K. Mori et al. (Eds.): MICCAI 2013, Part II, LNCS 8150, pp. 607–615, 2013.

regions and *ii)* the computational improvement to the corresponding estimation procedures. We also bring to light the main trade-offs between clustering and decomposition strategies and show that our approach achieves better region extraction in this trade-off space. The paper is organized as follows. In section 2, we present our new region-extraction method. Section 3 presents experiments to compare different approaches. Finally section 4 summarizes the empirical results.

2 A Dictionary Learning Approach to Segment Regions

Prior Methods Used to Extract Regions. Various unsupervised methods are routinely used to extract structured patterns from resting-state fMRI data. These patterns are then interpreted in terms of functional networks or regions. *Spatial Group Independent Component Analysis* (Group ICA) is the most popular method to process resting-state fMRI. It is based on a linear mixing model to separate different signals and relies on a principal component analysis (PCA) to reject noise [3]. *K-Means* is the de facto approach to learn clusters minimizing the ℓ_2 reconstruction error: it learns a hard assignment for optimal compression. *Ward Clustering* also seeks to minimize ℓ_2 error, but using agglomerative hierarchical clustering. The benefits are that imposing a spatial constraint comes at no cost and it has been extensively used to learn brain parcellations [4].

Multi-Subject Dictionary Learning. Our approach builds upon the Multi-Subject Dictionary Learning (MSDL) formulation [13]. The corresponding learning strategy is a minimization problem comprising a subject-level data-fit term, a term controlling subject-to-group differences, and a group-level penalization:

$$\underset{\mathbf{U}_s,\mathbf{V}_s,\mathbf{V}}{\operatorname{argmin}} \frac{1}{S}\sum_{s\in S}\frac{1}{2}\left(\left\|\mathbf{Y}_s-\mathbf{U}_s\mathbf{V}_s\right\|_{\mathrm{Fro}}^2+\mu\left\|\mathbf{V}_s-\mathbf{V}\right\|_{\mathrm{Fro}}^2\right)+\mu\,\alpha\,\Omega(\mathbf{V}) \qquad (1)$$

where $\mathbf{Y}^s \in \mathbb{R}^{n\times p}$ are the n-long time series observed on p voxels for subject s, $\mathbf{U}^s \in \mathbb{R}^{n\times k}$ are the k time series associated to the subject maps $\mathbf{V}^s \in \mathbb{R}^{p\times k}$, $\mathbf{V} \in \mathbb{R}^{p\times k}$ is the set of group-level maps and Ω is a regularization function. μ is a parameter that controls the similarity between subject-level and group-level maps while α sets the amount of regularization enforced on the group-level maps.

This problem is not jointly convex with respect to $\{\mathbf{U}^s\}$, $\{\mathbf{V}^s\}$ and $\mathbf{V}$, but it is separately convex and [13] relies on an alternate minimization strategy, optimizing separately (1) with regards to $\{\mathbf{U}^s\}$, $\{\mathbf{V}^s\}$ and $\mathbf{V}$ while keeping the other variables fixed. Importantly, the optimization step with regards to $\mathbf{V}$ amounts to computing a proximal operator, also used in *e.g.* image-denoising:

$$\mathrm{prox}_{\alpha\Omega}(\mathbf{w}) \stackrel{\mathrm{def}}{=} \underset{\mathbf{v}}{\operatorname{argmin}} \left\|\mathbf{w}-\mathbf{v}\right\|_2^2+\alpha\,\Omega(\mathbf{v}) \qquad (2)$$

Sparse TV Penalization to Enforce Compact Regions. We want to define a small set of regions that represent well brain-activity signals. Dictionary learning does not produce in itself regions, but continuous maps. Enforcing sparsity, *e.g.* via

an ℓ_1 penalty ($\Omega(\mathbf{v}) = \|\mathbf{v}\|_1$) on these maps, implies that they display only a few salient features that may not be grouped spatially. [13] use a smoothness prior (ℓ_2 norm of the image gradient) in addition to the sparsity prior to impose spatial structure on the extracted maps. However, while smoothness is beneficial to rejecting very small structures and high-frequency noise, it also smears edges and does not constrain the long-distance organization of the maps.

The simplest convex relaxation of a segmentation problem is the minimization of the total variation (TV) [7] that tends to produce plateaus. Briefly, the total variation is defined as the norm of the gradient of the image: $\mathrm{TV}(\mathbf{v}) = \sum_i \sqrt{(\nabla_x v)_i^2 + (\nabla_y v)_i^2 + (\nabla_z v)_i^2}$. Considering the image gradient as a linear operator $\nabla : \mathbf{v} \in \mathbb{R}^p \rightarrow (\nabla_x \mathbf{v}, \nabla_y \mathbf{v}, \nabla_z \mathbf{v}) \in \mathbb{R}^{3p}$, $\mathrm{TV}(\mathbf{v}) = \|\nabla \mathbf{v}\|_{21}$, where the ℓ_{21}-norm groups [9] are the x, y, z gradient components at one voxel position.

Going beyond TV regularization, we want to promote regions comprising many voxels, but occupying only a fraction of the full brain volume. For this we combine ℓ_1 regularization with TV [1]. The corresponding proximal operator is

$$\operatorname*{argmin}_{\mathbf{v}} \left\|\mathbf{w} - \mathbf{v}\right\|_2^2 + \alpha \left(\|\nabla \mathbf{v}\|_{21} + \rho \|\mathbf{v}\|_1\right) = \operatorname*{argmin}_{\mathbf{v}} \left\|\mathbf{w} - \mathbf{v}\right\|_2^2 + \alpha \left\|\tilde{\nabla}_\lambda \mathbf{v}\right\|_{21} \quad (3)$$

where $\tilde{\nabla}_\lambda$ is an augmented operator $\mathbb{R}^p \rightarrow \mathbb{R}^{4p}$, consisting of a concatenation of the operator ∇ and the scaled identity operator $\rho \mathbf{I}$, and the ℓ_{21} norm uses an additional set of groups on the new variables. Note that the structure of the resulting problem is exactly the same as for TV, thus we can rely on the same efficient algorithms [2] to compute the proximal operator. Finally, in an effort to separate as much as possible different features on different components, we impose positivity on the maps. This constraint is reminiscent of non-negative matrix factorization [10] but also helps removing background noise formed of small but negative coefficients (as on the figures of [13]). It is enforced using an algorithm for constrained TV [2]. The optimality of the solution can be controlled using the dual gap [6], the computation of which can be adapted from [11]:

$$\delta_{\mathrm{gap}}(\mathbf{v}) = \|\mathbf{w} - \mathbf{v}\|_2^2 + \alpha \left\|\tilde{\nabla}_\lambda \mathbf{v}\right\|_{21} - \|\mathbf{w}\|_2^2 - \|\mathbf{v}\|_2^2 \quad (4)$$

Computational Efficiency. We introduce three improvements to the original optimization algorithm of MSDL: stochastic coordinate descent rather than cycling block coordinate descent, computing image gradients on rectangular geometries, and an adaptive dual gap control on the proximal operator solver.

The algorithm outlined in [13] to minimize (1) is an alternate minimization using a cyclic block coordinate descent. The time required to update the $\{\mathbf{U}^s, \mathbf{V}^s\}$ parameters grows linearly with the number of subjects, and becomes prohibitive for large populations. For this reason, rather than a cyclic choice of coordinates to update, we alternate selecting a random subset of subjects to update $\{\mathbf{U}^s, \mathbf{V}^s\}$ and updating $\mathbf{V}$. This stochastic coordinate descent (SCD) strategy draws from the hypothesis that subjects are similar and a subset brings enough representative information to improve the group-level maps $\mathbf{V}$ while bringing

the computational cost of an iteration of the outer loop of the algorithm down. More formally, the justification of this strategy is similar to the stochastic gradient descent approaches: the loss term in (1) is a mean of subject-level term [5] over the group; we are interested in minimizing the expectation of this term over the population and, for this purpose, we can replace the mean by another unbiased estimator quicker to compute, the subsampled mean.

The computation of spatial regularization, whether it be with smooth lasso or TV penalization, implies computing spatial gradients of the images. However, in fMRI, it is most often necessary to restrict the analysis to a mask of the brain: out-of-brain volumes contain structured noise due *e.g.* to scanner artifacts. This masking imposes to work with an operator ∇ that has no simple expression. This is detrimental to computational efficiency because *i)* the computation of the proximal operator has to cater for border effects with the gradient for voxels on the edge of the mask –see *e.g.* [11] for more details– *ii)* applying ∇ and ∇^T imposes inefficient random access to the memory while computing gradients on rectangular image-shaped data can be done very efficiently. For these reasons, we embed the masked maps $\mathbf{v}$ into "unmasked" rectangular maps, on which the computation of the proximal term is fast: $\mathcal{M}^{-1}(\mathbf{v})$, where $\mathcal{M}$ is the masking operator. In practice, this amounts to using $\mathcal{M}(\mathbf{z})$ with $\mathbf{z} = \text{prox}_\Omega(\mathcal{M}^{-1}(\mathbf{w})))$ when computing $\text{prox}_\Omega(\mathbf{w})$, and correcting the energy with the norm of $\mathbf{z}$ outside of the mask. Indeed, in the expression of the proximal operator (2) $\|\mathcal{M}^{-1}(\mathbf{w}) - \mathbf{z}\|^2 = \|\mathbf{w} - \mathcal{M}(\mathbf{z})\|^2 + \|\mathcal{M}^{-1}(\mathcal{M}(\mathbf{z})) - \mathbf{z}\|^2$ where the first term is the term in the energy (1) and the second term is the correction factor that does not affect the remainder of the optimization problem (1).

Finally, we use the fact that in an alternate optimization it is not always necessary to optimize to a very small tolerance all the terms for each execution of the outer loop. In particular, the final steps of convergence of TV-based problems can be very slow. The dual-gap (4) gives an upper bound of the distance of the objective to the optimal. We introduce an adaptive dual gap (ADG) strategy: at each iteration of the alternate optimization algorithm, we record how much the energy was decreased by optimizing on $\{\mathbf{U}^s, \mathbf{V}^s\}$ and stop the optimization of the proximal operator when the dual gap reaches a third of this value.

Extracting Regions. Decomposition methods such as Group ICA or MSDL produce continuous maps that we must turn into regions. For this purpose, we choose a threshold so that, on average, each voxel is non-zero in only one of the maps (keeping as many voxels as there are in the brain). For this, we consider the $^1/_k{}^{\text{th}}$ quantile of the voxel intensity across all the maps $\mathbf{V}$. This choice of threshold is independent of the map sparsity, or kurtosis, which is the relevant parameter in the case of ICA. It can thus be used with all models and penalization choices. Drawing from the simple picture that brain networks tend to display homologous inter-hemispheric regions that are strongly correlated and hard to separate, we aim at extracting $2\,k$ regions, and take the largest connected components in the complete set of maps after thresholding. Importantly, some maps can contribute more than two regions to the final atlas, while some might contribute none.

Finally, in order to compare linear decomposition and clustering methods on an equal footing, we also convert the extracted maps to a hard assignment by assigning each voxel to the map with the highest corresponding value.

3 Experiments

Evaluation Metrics. Gaging the success of an unsupervised method is challenging because its usefulness in application terms is not well defined. However, to form a suitable tool to represent brain function, a set of regions must be stable with regards to the subjects used to learn them, and must provide an adequate basis to capture the variance of fMRI data. To measure stability across models, we rely on the normalized mutual information [14] –as standard clustering stability score– computed from a hard assignment. Data fidelity is evaluated by learning, using least square fitting, the time series associated to the model maps and by computing the explained variance of these series over the original ones.

Dataset. We use the freely-available Autism Brain Imaging Database Exchange[1] dataset. It is a fMRI resting state dataset containing 539 subjects suffering of autism spectrum disorders and 573 typical controls. To avoid site-related artifacts, we restrict our study to data from University of Leuven. From the original 59 subjects, 11 have been removed because the top of the scans were cut. We apply our model evaluation metrics by performing 20 runs taking 36 random subjects as a train set to learn regions and the 12 remaining subjects as test set.

Parameter Choice. Following [13], we use a dimensionality $k = 42$, which implies that we extract 84 regions. With regards to the PCA, which does not produce contrasts maps, we consider each map as a region. Parameter μ controls subject-to-group differences. It has only a small impact on the resulting group-level maps and we set it to 1. ρ and α control the overall aspect of the maps. Maximizing explained variance on test data leads to setting $\rho = 2.5$ and $\alpha = 0.01$. However optimizing for explained variance always privileges low-bias models that fit close to the data, *i.e.* under-penalizing. These are not the ideal settings to extract well-formed regions as the corresponding maps are not very contrasted. We also investigate settings with α 20 times larger, to facilitate region extraction.

4 Results

We report results for MSDL with various penalizations. In the following, "TV-MSDL" denotes MSDL with sparse TV penalization and $\alpha = 0.20$. "MSDL low regularization" is the same penalization with $\alpha = 0.01$. "Smooth MSDL" denotes the original formulation of [13] with smooth Lasso regularization and $\alpha = 0.20$.

[1] `http://fcon_1000.projects.nitrc.org/indi/abide/`

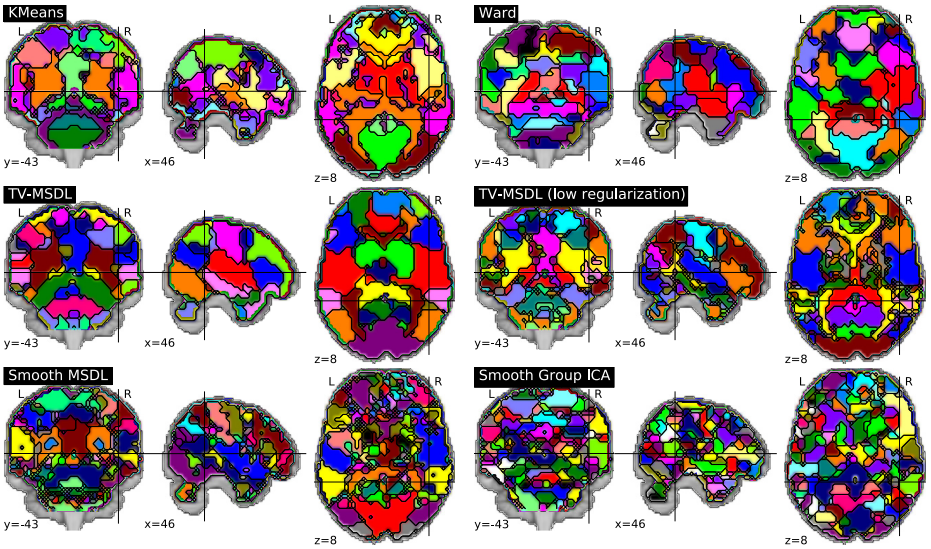


Fig. 1. Regions extracted with the different strategies (colors are random). Please note that a 6mm smoothing has been applied to data before ICA to enhance region extraction.

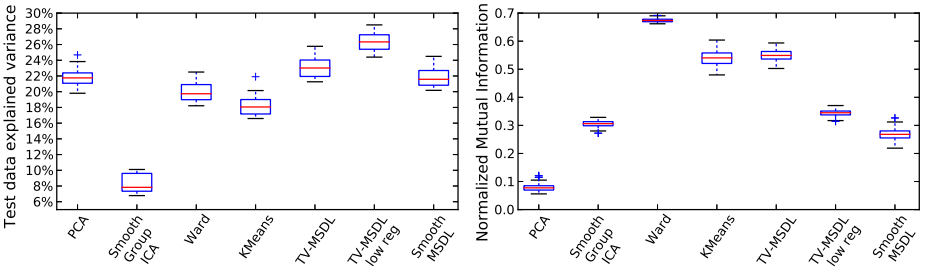
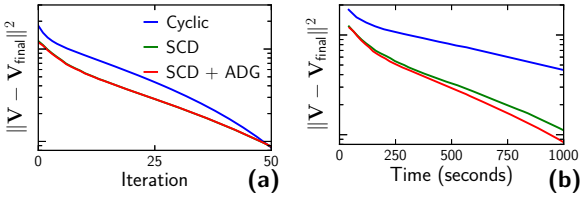


Fig. 2. (Left) Explained variance on test data for various strategies. **(Right)** Normalized mutual information between regions extracted from two different subsets.



Time per iteration

Update	$\{\mathbf{U}^s, \mathbf{V}^s\}$	$\mathbf{V}$
Cyclic	37.3s	7.6s
SCD	11.7s	8.4s
SCD + ADG	11.3s	4.0s

(c)

Fig. 3. Comparing different optimization strategies: cyclic block coordinate descent, as proposed by [13], stochastic coordinate descent (SCD), and SCD with adaptive dual gap (ADG) on the proximal term. **(a)** distance to the optimal $\mathbf{V}$ (in log scale) as a function of the number of iterations, **(b)** distance to the optimal $\mathbf{V}$ (in log scale) as a function of time, **(c)** time spent per iteration to update $\{\mathbf{U}^s, \mathbf{V}^s\}$ or $\mathbf{V}$.

Qualitative Assessment of the Brain Regions. Fig. 1 represents an hard assignment of the regions created with different methods. First, we note that, unsurprisingly, algorithms with spatial constraints give more structured parcellations. This behavior is visible in the TV-MSDL regions when regularization increases, but also when comparing to smooth MSDL, that does not enforce long-range structure. Regions extracted by TV-MSDL segment best well-known structures, such as the ventricles or gray nuclei. Finally, their strong symmetry matches neuroscientific knowledge on brain networks, even though it was not imposed by the model.

Stability-fidelity trade-offs. Fig. 2 shows our model-evaluation metrics: explained variance on test data capturing data fidelity and normalized mutual information between the assignments estimated on independent subjects measuring the methods' stability. PCA is the optimal linear decomposition method to maximize explained variance on a dataset: on the training set, it outperforms all others algorithms with about 40% of explained variance. On the other hand, on the test set, its score is significantly lower (about 23%): the components that it learned in the tail of the spectrum are representative of noise and not reproducible. In other words, it overfits the training set. This overfit also explains its poor stability: the first six principal components across the models are almost equal, however in the tail they start to diverge and eventually share no similarity. While the raw ICA maps span the same subspace as the PCA maps, when thresholded and converted to regions, their explained variance drops: Group ICA does not segment regions reliably. Both clustering techniques give very similar results, although Ward is more stable and explains test data better than K-means. They both give very stable regions but do not explain the data as well as PCA.

Thus on these reference methods, we observe a trade-off between data fidelity and stability. Linear models can explain the data well but do not yield very stable regions as opposed to clustering methods. An ideal model would maximize both fidelity and stability. TV-MSDL can improve on both aspects and explore different parts of the trade-off by controlling the amount of regularization. The regularization can be set to maximize explained variance (*TV-MSDL low reg*) to the cost of less contrast in the maps and less stability. Increasing regularization to restore contrast (*TV-MSDL*) gives regions with an explained variance greater than that of PCA, but with a stability similar to that of K-means.

Computational Speedup. Fig. 3 shows speed benchmarks realized on the full dataset (48 subjects), parallelizing the computation of $\{\mathbf{U}^s, \mathbf{V}^s\}$ on 16 cores. Profiling results (Fig 3c) show that the update $\{\mathbf{U}^s, \mathbf{V}^s\}$ is the bottleneck. Using SCD with a stochastic subset of a fourth of the dataset proportionnaly decreases the time of this step and only has a little impact on convergence rate per iteration (Fig 3a). However, the iteration time speedup brought by SCD dramatically increases overall convergence speed (Fig 3b). ADG yields an additional speed up, and altogether, we observe a speedup of a factor 2, but we expect it to increase with the group size. SCD combined with ADG enable tackling large groups.

5 Conclusion

Decomposition methods such as ICA or dictionary learning delineate functional networks that capture the variance of the signal better than parcels extracted using clustering of voxels. Regions can be recovered from the network maps by thresholding. They are however much more unstable across subjects than clustering results, and this thresholding can be detrimental to the explained variance. We have introduced a new region-extraction approach that pushes further this fidelity/stability trade-off. It uses a sparse TV penalty with dictionary learning to combine the tendency of TV to create discrete spatial patches with the ability of linear decomposition models to unmix different effects. Careful choices of optimization strategy let our method scale to very large groups of subjects. The resulting regions are stable, reveal a neurologically-plausible partition of the brain, and can give a synthetic representation of the resting-state correlation structure in a population. This representation opens the door to learning phenotypic markers from rest, *e.g.* for diagnosis of neurological disorders.

Acknowledgments. We acknowledge funding from the NiConnect project and NIDA R21 DA034954, SUBSample project from the DIGITEO Institute, France.

References

1. Baldassarre, L., Mourao-Miranda, J., Pontil, M.: Structured sparsity models for brain decoding from fMRI data. In: PRNI, pp. 5–8 (2012)
2. Beck, A., Teboulle, M.: Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems. Trans. Image Proc. 18, 2419–2434 (2009)
3. Beckmann, C.F., Smith, S.M.: Probabilistic independent component analysis for functional magnetic resonance imaging. Trans. Med. Im. 23, 137–152 (2004)
4. Blumensath, T., Behrens, T.E.J., Smith, S.M.: Resting-state FMRI single subject cortical parcellation based on region growing. In: Ayache, N., Delingette, H., Golland, P., Mori, K. (eds.) MICCAI 2012, Part II. LNCS, vol. 7511, pp. 188–195. Springer, Heidelberg (2012)
5. Bottou, L.: Stochastic learning. In: Bousquet, O., von Luxburg, U., Rätsch, G. (eds.) Machine Learning 2003. LNCS (LNAI), vol. 3176, pp. 146–168. Springer, Heidelberg (2004)
6. Boyd, S., Vandenberghe, L.: Convex optimization. Cambridge University Press
7. Chambolle, A., Caselles, V., Cremers, D., Novaga, M., Pock, T.: An introduction to total variation for image analysis. In: Fornasier, M. (ed.) Theoretical Foundations and Numerical Methods for Sparse Recovery, vol. 9, pp. 263–340 (2010)
8. Kiviniemi, V., Starck, T., Remes, J., et al.: Functional segmentation of the brain cortex using high model order group PICA. Hum. Brain Map. 30, 3865–3886 (2009)
9. Kowalski, M.: Sparse regression using mixed norms. Applied and Computational Harmonic Analysis 27, 303–324 (2009)
10. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature 401, 788–791 (1999)

11. Michel, V., Gramfort, A., Varoquaux, G., Eger, E., Thirion, B.: Total variation regularization for fMRI-based prediction of behavior. Trans. Med. Imag. 30, 1328–1340 (2011)
12. 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: Jiang, T., Navab, N., Pluim, J.P.W., Viergever, M.A. (eds.) MICCAI 2010, Part I. LNCS, vol. 6361, pp. 200–208. Springer, Heidelberg (2010)
13. Varoquaux, G., Gramfort, A., Pedregosa, F., Michel, V., Thirion, B.: Multi-subject dictionary learning to segment an atlas of brain spontaneous activity. In: Székely, G., Hahn, H.K. (eds.) IPMI 2011. LNCS, vol. 6801, pp. 562–573. Springer, Heidelberg (2011)
14. Vinh, N., Epps, J., Bailey, J.: Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. Journal of Machine Learning Research 11, 2837–2854 (2010)
15. Yeo, B., Krienen, F., Sepulcre, J., Sabuncu, M., et al.: The organization of the human cerebral cortex estimated by intrinsic functional connectivity. J. Neurophysio. 106, 1125–1165 (2011)