A Generative-Discriminative Basis Learning Framework to Predict Clinical Severity from Resting State Functional MRI Data

Niharika Shimona D'Souza[1]([✉]), Mary Beth Nebel[2,3], Nicholas Wymbs[2,3], Stewart Mostofsky[2,3,4], and Archana Venkataraman[1]

[1] Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, USA
`shimona.niharika.dsouza@jhu.edu`
[2] Center for Neurodevelopmental Medicine and Research, Kennedy Krieger Institute, Baltimore, USA
[3] Department of Neurology, Johns Hopkins School of Medicine, Baltimore, USA
[4] Department of Pediatrics, Johns Hopkins School of Medicine, Baltimore, USA

Abstract. We propose a matrix factorization technique that decomposes the resting state fMRI (rs-fMRI) correlation matrices for a patient population into a sparse set of representative subnetworks, as modeled by rank one outer products. The subnetworks are combined using patient specific non-negative coefficients; these coefficients are also used to model, and subsequently predict the clinical severity of a given patient via a linear regression. Our generative-discriminative framework is able to exploit the structure of rs-fMRI correlation matrices to capture group level effects, while simultaneously accounting for patient variability. We employ ten fold cross validation to demonstrate the predictive power of our model on a cohort of fifty eight patients diagnosed with Autism Spectrum Disorder. Our method outperforms classical semi-supervised frameworks, which perform dimensionality reduction on the correlation features followed by non-linear regression to predict the clinical scores.

1 Introduction

Resting state fMRI (rs-fMRI) allows us to assess brain activity and localize critical functions through steady state patterns of co-activation [1]. Building predictive models at the patient level remains an open challenge due to the high data dimensionality and to the considerable inter-subject variability. Predictive analysis methods usually follow a two step procedure. First, feature selection is applied to the raw correlation values; examples include graph theoretic measures, statistical measures and embedding features obtained from unsupervised learning techniques such as PCA, k-PCA or ICA [2]. As a second step, conventional regression techniques such as Random Forests or Support Vector Regression are applied to the feature space representation to predict the clinical severity. These

strategies adequately capture the group-averaged functional connectivity across the cohort but fail to account for individual variability. Consequently, the generalization power of these techniques is limited.

The recent success of Bayesian [3] and dictionary learning [4] studies on rs-fMRI data is largely based on their ability to simultaneously model the patient and group level information. [4] introduces a basis learning framework for patient subtype classification, which reduces the dimensionality of T1 MR voxel based morphometry data while preserving the anatomical interpretability. [5] introduces a correlation matrix decomposition strategy, where multiple rank one matrix outer products capturing the underlying 'generative' basis are combined using patient specific coefficients. The sparse basis networks identify meaningful co-activation patterns common to all the patients, and the coefficients model the patient variability. Our main contribution lies in exploiting the 'discriminative' nature of rs-fMRI correlation matrices. We estimate the clinical severity of every patient by constructing a regression model which maps the behavioral scores to the functional data space. We jointly optimize for each of the hidden variables in the model, i.e. the basis, coefficients and regression weights. We refine the validation process by quantifying the model generalizability in terms of the regression performance on unseen data, as opposed to the correlation fit measure in [5]. Hence, our framework is less prone to overfitting.

We validate our framework on a population study of Autism Spectrum Disorder (ASD). Patient variability manifests as a spectrum of impairments, typically quantified by a "behavioral score" of clinical severity obtained from an expert assessment. Identifying sub-networks predictive of ASD severity is the key link to understanding the social and behavioral implications of the disorder. Our inclusion of behavioral data into the optimization framework guides the identification of representative networks specific to resting state ASD characterization.

2   A Joint Model for Connectomics and Clinical Severity

Let N be the number of patients and M be the number of regions in our brain parcellation. We decompose the patient correlation matrices $\mathbf{\Gamma}_n \in \mathcal{R}^{M \times M}$ into a non-negative combination of a K basis subnetworks $\mathbf{b}_k \mathbf{b}_k^T$. The sparse vector $\mathbf{b}_k$ indicates the relative contribution of each brain region to network k. The vector $\mathbf{c}_n$ denotes the non-negative contribution of each subnetwork for patient n. The coefficients $\mathbf{c}_n$ are subsequently used to model the clinical severity score y_n via the regression weight vector $\mathbf{w} \in \mathcal{R}^K$. We concatenate the subnetworks into a basis matrix $\mathbf{B} \in \mathcal{R}^{M \times K}$, the coefficients into the matrix $\mathbf{C} \in \mathcal{R}^{K \times N}$, and the scores into a vector $\mathbf{y} \in \mathcal{R}^N$. Our combined objective can be written as follows:

$$\mathcal{J}(\mathbf{B}, \mathbf{C}, \mathbf{W}) = \sum_n ||\mathbf{\Gamma}_n - \mathbf{B}\mathbf{diag}(\mathbf{c}_n)\mathbf{B}^T||_F^2 + \gamma||\mathbf{y} - \mathbf{C}^T\mathbf{w}||_2^2 \quad s.t. \ \ \mathbf{c}_{nk} \geq 0, \ (1)$$

Here, γ is the tradeoff between the behavioral and functional data terms, and $\mathbf{diag}(\mathbf{c}_n)$ is a matrix with the elements of $\mathbf{c}_n$ on its leading diagonal, and off diagonal elements as 0. We impose an ℓ_1 penalty upon the matrix $\mathbf{B}$ in order to

recover a sparse set of subnetworks. Since the objective in Eq. (1) is ill posed, we add quadratic penalty terms on $\mathbf{C}$ and $\mathbf{w}$ which act as regularizers.

$$\lambda_1||\mathbf{B}||_1 + \lambda_2||\mathbf{C}||_2^2 + \lambda_3||\mathbf{w}||_2^2. \tag{2}$$

Equation (2) is added to the overall objective with λ_1, λ_2 and λ_3 being the sparsity, norm penalty on $\mathbf{C}$, and the penalty on $\mathbf{w}$ respectively.

2.1   Optimization Strategy

We employ a fixed point alternating minimization strategy to optimize $\mathbf{B}$, $\mathbf{C}$ and $\mathbf{w}$. At every iteration, the optimal solution for one variable is calculated assuming the other variables are held constant. Proximal gradient descent [6] is an effective strategy of optimizing a non-differentiable sparsity penalty such as the one in Eq. (2), when the supporting terms are convex in the variable of interest. However, the expansion of the first Frobenius norm gives rise to non-convex bi-quadratic terms in $\mathbf{B}$, which prevents us from directly computing a proximal solution. Therefore, we introduce N constraints of the form $\mathbf{D}_n = \mathbf{Bdiag}(\mathbf{c}_n)$, which are enforced by an Augmented Lagrangian penalty:

$$\mathcal{J}(\mathbf{B}, \mathbf{C}, \mathbf{w}, \mathbf{D}_n, \mathbf{\Lambda}_n) = \sum_n ||\mathbf{\Gamma}_n - \mathbf{D}_n\mathbf{B}^T||_F^2 + \sum_n \text{Tr}\left[\mathbf{\Lambda}_n^T(\mathbf{D}_n - \mathbf{Bdiag}(\mathbf{c}_n))\right]$$

$$+ \sum_n \frac{1}{2}||\mathbf{D}_n - \mathbf{Bdiag}(\mathbf{c}_n)||_F^2 + \gamma||\mathbf{y} - \mathbf{C}^T\mathbf{w}||_2^2 \;\; s.t. \;\; \mathbf{c}_{nk} \geq 0 \tag{3}$$

where, each $\mathbf{\Lambda}_n$ is a matrix of Lagrangians and each of the supporting Frobenius norm terms are regularizers on the Lagrangian constraints. The objective in Eq. (3) is convex in $\mathbf{B}$ and the set $\{\mathbf{D}_n\}$ separately. Our optimization begins by randomly initializing $\mathbf{B}$, $\mathbf{C}$ and $\mathbf{w}$ and setting $\mathbf{D}_n = \mathbf{Bdiag}(\mathbf{c}_n)$ and $\mathbf{\Lambda}_n = \mathbf{0}$. We then iterate through the following four steps until global convergence.

Step 1 - Optimizing B via Proximal Gradient Descent. Given the fixed learning rate parameter t, the proximal update for $\mathbf{B}$ is:

$$\mathbf{B}^{k+1} = \mathbf{sgn}(\mathbf{X}).^*(\mathbf{max}(|\mathbf{X}| - t, \mathbf{0})) \;\; s.t. \;\; \mathbf{X} = \mathbf{B}^k - (t/\lambda_1)\frac{\partial\mathcal{J}}{\partial\mathbf{B}} \tag{4}$$

The derivative of $\mathcal{J}$ with respect to $\mathbf{B}$, where $\mathbf{V}_n = \mathbf{diag}(\mathbf{c}_n)$, is computed as:

$$\frac{\partial\mathcal{J}}{\partial\mathbf{B}} = \sum_n \left[2\left[\mathbf{BD}_n^T\mathbf{D}_n - \mathbf{\Gamma}_n\mathbf{D}_n\right] - \mathbf{D}_n\mathbf{V}_n + \mathbf{BV}_n^2 - \mathbf{\Lambda}_n\mathbf{V}_n\right] \tag{5}$$

As seen, the non-smoothness of the $||\mathbf{B}||_1$ penalty is handled by performing iterative shrinkage thresholding applied on a locally smooth quadratic model.

Step 2 - Optimizing C using Quadratic Programming. The objective is quadratic in $\mathbf{C}$ when $\mathbf{B}$ and $\mathbf{w}$ are held constant. Furthermore, the $\mathbf{diag}(\mathbf{c}_n)$

term decouples the updates for $\mathbf{c}_n$ across patients. Hence, we use N quadratic programs of the form below to solve for the vectors $\{\mathbf{c}_n\}$:

$$\frac{1}{2}\mathbf{c}_n^T\mathbf{H}_n\mathbf{c}_n + \mathbf{f}_n^T\mathbf{c}_n \quad s.t. \quad \mathbf{A}_n\mathbf{c}_n \geq \mathbf{b}_n \tag{6}$$

The Quadratic Programming parameters for our problem are given by:

$$\mathbf{H}_n = \mathbf{diag}(\mathbf{B}^T\mathbf{B}) + 2\gamma\mathbf{w}\mathbf{w}^T + 2\lambda_2\mathcal{I}_K$$
$$\mathbf{f}_n = -\mathbf{diag}(\mathbf{D}_n^T\mathbf{B}) - \mathbf{diag}(\mathbf{\Lambda}_n^T\mathbf{B}) - 2\gamma y_n\mathbf{w}; \quad \mathbf{A}_n = -\mathcal{I}_K \quad \mathbf{b}_n = \mathbf{0}$$

This strategy helps us find the globally optimal solutions for $\mathbf{c}_n$, as projected onto the K dimensional space of positive real numbers.

Step 3 - Closed Form Update for $\mathbf{w}$. The global minimizer of $\mathbf{w}$ computed at the first order stationary point can be expressed as:

$$\mathbf{w} = (\mathbf{C}\mathbf{C}^T + \frac{\lambda_3}{\gamma}\mathcal{I}_K)^{-1}(\mathbf{C}\mathbf{y}) \tag{7}$$

Step 4 - Optimizing the Constraint Variables $\mathbf{D}_n$ **and** $\mathbf{\Lambda}_n$. Each of the primal variables $\{\mathbf{D}_n\}$ has a closed form solution given by:

$$\mathbf{D}_n = (\mathbf{diag}(\mathbf{c}_n)\mathbf{B}^T + 2\mathbf{\Gamma}_n\mathbf{B} - \mathbf{\Lambda}_n)(\mathcal{I}_K + 2\mathbf{B}^T\mathbf{B})^{-1} \tag{8}$$

In contrast, we update the dual variables $\{\mathbf{\Lambda}_n\}$ using gradient ascent:

$$\mathbf{\Lambda}_n^{k+1} = \mathbf{\Lambda}_n^k + \eta_k(\mathbf{D}_n - \mathbf{B}\mathbf{diag}(\mathbf{c}_n)) \tag{9}$$

The updates for $\mathbf{D}_n$ and $\mathbf{\Lambda}_n$ ensure that the proximal constraints are satisfied with increasing certainty at each iteration. The learning rate parameter η_k for the gradient ascent step of the augmented Lagrangian is chosen to guarantee sufficient decrease for every iteration of alternating minimization. In practice, we initialize this value to 0.001, and scale it by 0.75 at each iteration.

In all of our derivations, $\mathrm{Tr}(\mathbf{M})$ is the trace operator and gives the sum of the diagonal elements of a matrix $\mathbf{M}$, and $\mathcal{I}_K$ is the $K \times K$ identity matrix.

2.2   Predicting Symptom Severity

We use cross validation to evaluate the predictive power of our model. Specifically, we compute the optimal $\{\mathbf{B}^\star, \mathbf{w}^\star\}$ based on the training dataset. We can then estimate the coefficients $\mathbf{c}_{test}$ for a new patient by re-solving the quadratic program in **Step 2** using the previously computed $\{\mathbf{B}^\star, \mathbf{w}^\star\}$. Notice that we must set the data term $\gamma||\mathbf{C}^T\mathbf{w} - \mathbf{y}||_2^2$ to 0 in the testing experiments, since the severity y_{test} is unknown. Also, we assume that the constraint $\mathbf{D}_{test} = \mathbf{B}^\star\mathbf{diag}(\mathbf{c}_{test})$ is satisfied exactly for the conditions of the proximal operator to hold. Finally, $y_{test} = \mathbf{c}_{test}^T\mathbf{w}^\star$ is the estimate of the behavioral score for the unseen test patient.
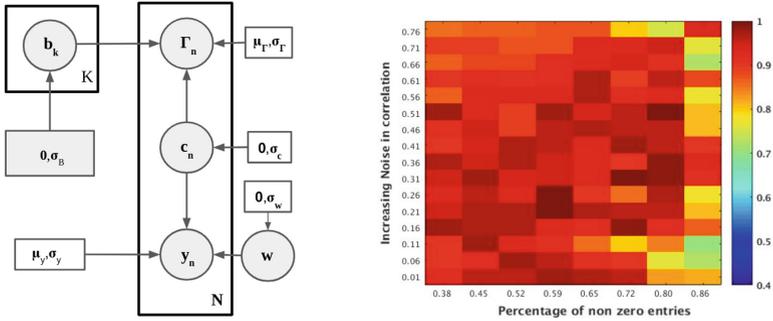
Fig. 1. (a) The graphical model from which we generate synthetic data (b) The network recovery performance of our algorithm for varying levels of sparsity and correlation matrix noise variance. For our experiments, we fix the rest of the parameters of the model at $\sigma_{\mathbf{B}} = 0.2$, $\mu_{\boldsymbol{\Gamma}_n} = \mathbf{B}\mathbf{diag}(\mathbf{c}_n)\mathbf{B}^T$, $\sigma_{\mathbf{c}} = 0.1$, $\mu_{\mathbf{y}} = \mathbf{c}_n^T\mathbf{w}$, $\sigma_{\mathbf{y}} = 0.2$ and $\sigma_{\mathbf{w}} = 0.1$

2.3   Baseline Comparison Methods

We compare our algorithm with a standard machine learning pipeline to predict the target severity score. We first perform dimensionality reduction to concentrate the $\frac{M \times (M-1)}{2}$ rs-fMRI correlation pairs into a small number of basis elements. Then, we construct a non-linear regression model to predict clinical severity. We consider two dimensionality reduction/regression combinations:

1. Principal Component Analysis on the correlation coefficients followed by a Random Forest Regression on the projected features
2. Kernel Principal Component Analysis on the correlation coefficients followed by a Random Forest Regression on the embedding features.

3   Experimental Results

Evaluating Robustness on Synthetic Data. Our optimization problem in Eq. (1) suggests an underlying graphical model, depicted in Fig. 1(a). Notice that the ℓ_1 penalty on $\mathbf{B}$ translates to a Laplacian prior with σ_B controlling the potentially overlapping level of sparsity. In contrast, the ℓ_2 penalties translate into Gaussian distributions, with the mean corresponding to the ℓ_2 argument and the variance related to the regularization parameters. We use this model to sample the correlation matrices $\{\boldsymbol{\Gamma}_n\}$ and the behavioral scores $\{y_n\}$, and then infer the latent networks generating the data. Figure 1(b) indicates the performance of network recovery from our algorithm. We quantify the peformance in terms of average inner-product similarity between recovered networks and generating networks, both normalized to unit norm. The number of generating and recovery networks is chosen to be 4. Unsurprisingly, increasing the overlap in the sparsity patterns across networks and increasing the noise in the correlation estimates worsens the recovery performance. However, our optimization procedure

is robust in the noise regime estimated from our real-world rs-fMRI correlation matrices (0.01–0.2) and for recovered sparsity levels (0.1–0.4). The experiment also helps us identify stable parameter settings for the next section.

rs-fMRI Dataset and Preprocessing. We evaluate our method on a cohort of 58 children with high-functioning ASD (Age: 10.06 ± 1.26, IQ: 110 ± 14.03). We acquired rs-fMRI scans on a Phillips 3T Achieva scanner using a single-shot, partially parallel gradient-recalled EPI sequence (TR/TE $= 2500/30$ ms, flip angle $= 70°$, res $= 3.05 \times 3.15 \times 3$ mm, 128 or 156 time samples).

Rs-fMRI preprocessing [3] consisted of slice time correction, rigid body realignment, and normalization to the EPI version of the MNI template using SPM. We use a CompCorr strategy to remove the spatially coherent noise from the white matter, ventricles, and six rigid body realignment parameters. We then spatially smoothed the data (6 mm FWHM Gaussian kernel) and band-pass filtered the time series (0.01–0.1 Hz). We use the Automatic Anatomical Labeling (AAL) atlas to define 116 cortical, subcortical and cerebellar regions. Empirically, we observed a consistent noise component having nearly constant contribution from all the brain regions and low predictive power. Consequently, we subtract out the contribution of the first eigenvector from the correlation matrices and used the residuals $\{\mathbf{\Gamma}_n\}$ as inputs for all the methods.

We consider two measures of clinical severity: Autism Diagnostic Observation Schedule (ADOS) total raw score [7], which captures the social and communicative interaction deficits of the patient along with repetitive behaviors (dynamic range: 0–30), and the Social Responsiveness Scale (SRS) total raw score [7] which characterizes social responsiveness (dynamic range: 70–200).

Predicting ASD Severity. We employ a ten fold cross validation strategy for each of the methods, whereby, we train the model on a 90% data split and evaluate the performance on the unseen 10% test data. We perform a grid search to find the optimal parameter setting for each method. Based on these results, we fix the regression tradeoff at $\gamma = 1$, and the three regularization parameters at $\{\lambda_1 = 40, \lambda_2 = 2, \lambda_3 = 1\}$ for SRS, and $\{\lambda_1 = 30, \lambda_2 = 0.2, \lambda_3 = 1\}$ for ADOS, and the learning rate at $t = 0.001$ for proximal gradient descent. The number of components was fixed at 15 for PCA and 10 for k-PCA. For k-PCA, we use an RBF kernel with the coefficient parameter varied between 0.01–10.

As seen from the Fig. 2, the baseline methods have poor validation performance and track the mean value of the held out data (shown by the black line). In comparison, our method not only consistently fits the training set more faithfully, but also generalizes much better beyond the training data. The major shortcoming of the baseline data-driven analysis techniques is in their failure to identify representative patterns of behavior from the correlation features. In contrast, our basis learning technique exploits the underlying structure of the correlation matrices and leverages patient specific information to map the ASD behavioral space, thus improving the prediction performance. As reported in Table 1, our method quantitatively outperforms the baselines approaches, both in terms of the root median square error (rMSE) and the R^2 performance.

Table 1. Performance evaluation using **root median square error (rMSE)** & $\mathbf{R^2}$ fit, both for testing & training. Lower MSE & higher R^2 score indicate better performance.

Score	Metric	Our method	PCA + RF Reg	k-PCA + RF Reg
ADOS	rMSE train	**0.088**	1.07	1.017
	R^2 train	**0.99**	0.94	0.96
	rMSE test	**2.53**	2.93	2.70
	R^2 test	**0.096**	0.031	0.01
SRS	rMSE train	**0.13**	6.43	6.90
	R^2 train	**0.99**	0.95	0.97
	rMSE test	**13.26**	20.51	20.30
	R^2 test	**0.052**	0.023	0.008
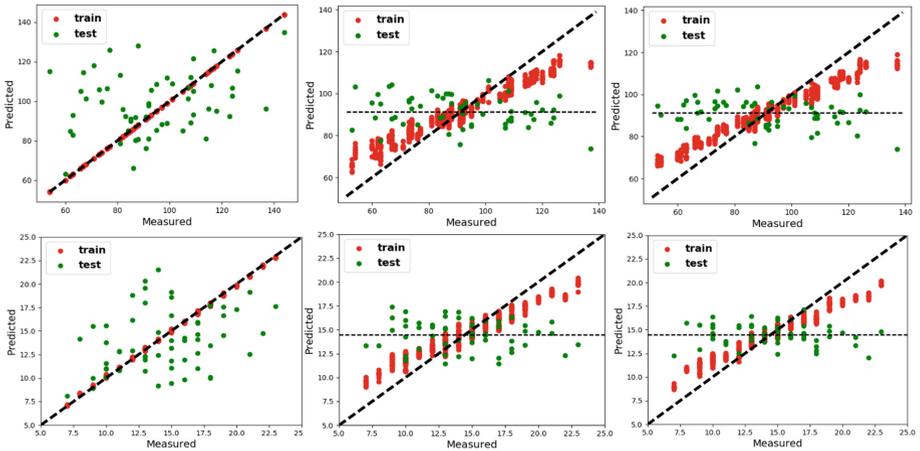
Fig. 2. Prediction performance of each method for ADOS (TR) & SRS (BR) **Left:** Our Method ($K = 8$) **Middle:** PCA ($comp = 15$) & RF Regression on the projected data **Right:** k-PCA ($comp = 10$, rbf $C = 0.1$) & RF Regression on the embedding features. Red & Green points correspond to the training & testing performance respectively

Subnetwork Identification. Figure 3 illustrates the basis subnetworks in **B** trained on the ADOS data. The colorbar indicates subnetwork contribution to the AAL regions. Regions storing negative values are anticorrelated with regions storing positive ones. Subnetwork 1 includes competing i.e. anticorrelated contributions from regions of the default mode network (DMN) and somatomotor network (SMN). Abnormal connectivity within the DMN and SMN has been previously reported in ASD [8]. Additionally, subnetwork 5 appears to be comprised of competing contributions from SMN regions and higher order visual processing areas in the occipital and temporal lobes, consistent with behavioral reports of reduced visual-motor integration in ASD. Subnetwork 2 includes competing contributions from prefrontal and subcortical regions (mainly the thalamus,
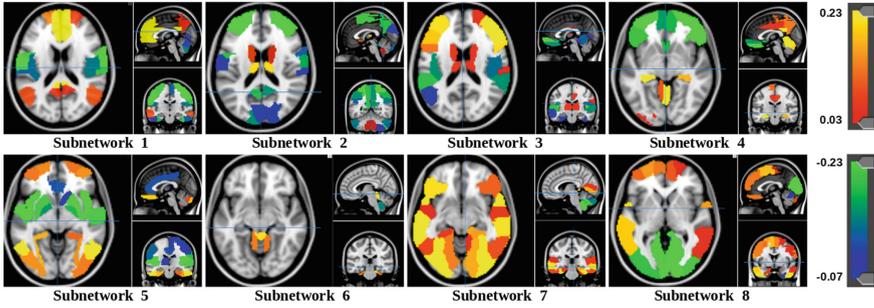
Fig. 3. Eight subnetworks identified by our model from ADOS prediction. The blue & green regions are anticorrelated with the red & orange regions for each subnetwork.

amygdala and hippocampus), that may be important for social-emotional regulation in ASD. Finally, subnetwork 3 is comprised of competing contributions from the central executive control network and the insula, which is thought to be critical for switching between self-referential and goal-directed behavior [9].

4   Conclusion

Unlike generic machine learning analysis, our matrix decomposition elegantly combines multimodal information from the imaging and behavioral domains. The key to our model is its ability to capture and learn from the structure of correlation matrices. Conventional analysis methods dramatically fall short of unifying the two data viewpoints reliably enough to implicate predictive functional patterns in the brain. Our joint optimization framework robustly identifies brain networks characterizing ASD and provides a key link to quantifying and interpreting the spectrum of manifestation of the disorder across a wide range of population. In the future, we will explore extensions of this model that jointly classify patients versus controls in addition to predicting symptom severity.

References

1. Fox, M.D., et al.: Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. Nat. Rev. Neurosci. **8**(9), 700 (2007)
2. Murphy, K.P.: Machine Learning: A Probabilistic Perspective. MIT Press, Cambridge (2012)
3. Venkataraman, A., et al.: Bayesian community detection in the space of group-level functional differences. IEEE Trans. Med. Imaging **35**(8), 1866–1882 (2016)

4. Batmanghelich, N.K., et al.: Generative-discriminative basis learning for medical imaging. IEEE Trans. Med. Imaging **31**(1), 51–69 (2012)
5. Eavani, H., et al.: Identifying sparse connectivity patterns in the brain using resting-state fMRI. Neuroimage **105**, 286–299 (2015)
6. Parikh, N., Boyd, S., et al.: Proximal algorithms. Found. Trends® Optim. **1**(3), 127–239 (2014)
7. Payakachat, N., et al.: Autism spectrum disorders: a review of measures for clinical, health services and cost-effectiveness applications. Expert. Rev. Pharmacoeconomics Outcomes Res. **12**(4), 485–503 (2012)
8. Nebel, M.B., et al.: Intrinsic visual-motor synchrony correlates with social deficits in autism. Biol. Psychiatry **79**(8), 633–641 (2016)
9. Sridharan, D., et al.: A critical role for the right fronto-insular cortex in switching between central-executive and default-mode networks. Proc. Natl. Acad. Sci. **105**(34), 12569–12574 (2008)