Sparse PLS hyper-parameters optimisation for investigating brain-behaviour relationships

F. S. Ferreira*†‡, M. J. Rosa*†‡, M. Moutoussis‡§, R. Dolan‡§, the NSPN consortium, J. Shawe-Taylor†, J. Ashburner§ and J. Mourao-Miranda*†‡

*Centre for Medical Image Computing, Department of Computer Science, University College London, UK
†Department of Computer Science, University College London, UK
‡Max Planck UCL Centre for Computational Psychiatry and Ageing Research, University College London, UK
§Wellcome Trust Centre for Neuroimaging, University College London, London, UK

***Abstract*—Unsupervised learning approaches, such as Partial Least Squares, can be used to investigate relationships between multiple sources of data, such as neuroimaging and behavioural data. In cases of high-dimensional datasets with limited number of examples (e.g. neuroimaging data) there is a need for regularisation to enable the solution of the ill-posed problem and prevent overfitting. Different approaches have been proposed to optimise the regularisation parameters in unsupervised models, however, so far, there has been no comparison between the different approaches using the same data. In this work, two optimisation frameworks (i.e. a permutation and a train/test framework) were compared using sparse PLS to investigate associations between brain connectivity and behaviour data. Both frameworks were able to identify at least one brain-behaviour associative effect. A second brain-behaviour effect was only found using the train/test framework. More importantly, the results show that the multivariate associative effects found with the train/test framework generalise better to new data, suggesting that results based on the permutation framework should be carefully interpreted.**

***Keywords*—Tuning parameters; High-dimensionality; Brain-behaviour; Regularisation; Sparse PLS**

I. Introduction

In recent years, unsupervised learning approaches have been increasingly used in neuroimaging due to the potential of these methods to explore multivariate relationships in different types of data. Canonical Correlation Analysis (CCA) and Partial Least Squares (PLS) are examples of these approaches and are used to find pairs of weight vectors that maximise the correlation or covariance, respectively, between the projections of data sources. Typically, neuroimaging data consists of a few samples that are high-dimensional. This represents an ill-posed problem which can be addressed with dimensionality reduction (e.g. PCA) or regularisation (e.g. sparse PLS) techniques. The former has been used to reduce the dimensionality of the data, for instance before applying CCA [1]. In the latter, regularisation constraints are added to the models, for instance to constrain the norm of the weights. Sparse PLS (SPLS) is a sparse version of PLS, in which L_1 and L_2 regularisations are used to impose sparsity on the weights [2]. The L_1-constraint has a hyper-parameter that controls the degree of sparsity, which will affect the number of variables selected in each data source. The contribution of this work is to compare two frameworks that have been previously proposed to optimise the SPLS hyper-parameters, the permutation framework [3] and a recent train/test framework proposed by Monteiro et al. [4]. The two strategies were compared in terms of the similarity between the SPLS weights and generalisability of the associative effects using a hold-out framework.

II. Material

A. Data

We used resting-state functional magnetic resonance imaging (rfMRI) and extensive item-level questionnaire data covering positive and negative mental health related behaviour and symptoms of 299 healthy and 33 depressed participants, comprising adolescents and young adults (14-24 years old) from the NeuroScience in Psychiatry Network (NSPN) study [5]. All MRI data were acquired on three identical 3T whole-body MRI systems (Magnetom TIM Trio; VB17 software version; Siemens Healthcare). rfMRI data from all participants were acquired using a multi-echo acquisition protocol with three volumes (echo times TE = 13, 31, 48 ms) per time point. For each participant, there were $\sim$11-minute time-series of rfMRI data with temporal resolution (TR) of 2420 ms and spatial resolution 3.8 mm isotropic. T1-weighted scans of resolution 1.0 mm isotropic were also acquired (TR = 18.70 ms) using six equidistant echo times (TE) between 2.2 and 14.7 ms, and averaged to form a single image of increased signal-to-noise ratio (SNR).

B. Data pre-processing

The initial questionnaire data comprised 380 variables (item-level). However, some variables were either removed because more than 95% of all participants had the same value (21 items) or because their standard deviation was zero (2 items). A total of 357 variables per participant were used.

The three series of rfMRI volumes were parcellated into 137 regions using an anatomical atlas. Regional mean time-series were estimated by averaging the fMRI signals over a set of voxels ($n = 58$) sampled from each region to avoid biasing the connectivity estimates by region size. The probabilistic sulci atlas from BrainVISA was used to define 123 cortical regions [6] and the Harvard-Oxford atlas was used to define 14 subcortical regions [7]. Six parameters from the realignment

978-1-5386-4291-7/18/$31.00

step of the pre-processing were regressed out of the averaged regional time-series. To estimate connection strengths between regions, partial correlations using L_2-norm ridge regression were computed [1]. Partial correlation values were converted into z-scores using Fisher's transformation (using FSLNets toolbox with the regularisation parameter $\rho = 0.01$ [1]) which has been shown to be a good option to estimate brain networks from fMRI data [8]. Partial correlation z-statistic matrices were estimated separately for each echo time-series and then averaged across the three echo times for each participant, resulting in a single functional brain connectivity matrix per participant, which was then vectorised. A total of 9316 brain connectivity variables per participant were used.

Confounding effects were regressed out of both brain and behavioural data using the following demeaned measures: *Mean frame displacement*, i.e. a summary statistic quantifying average subject head motion during the resting-state fMRI acquisitions; *site* (each MRI site was encoded as a one-hot variable); *age*; *sex*.

III. Methods

All analyses were performed using two different sources/views, i.e. brain connectivity data (data matrix $\boldsymbol{X} \in \mathbb{R}^{n \times p}$) and behavioural data (data matrix $\boldsymbol{Y} \in \mathbb{R}^{n \times q}$), where n is the number of subjects (i.e. 332 subjects), p is the number of brain connectivity features (i.e. 9316 connections) and q is the number of behavioural features (i.e. 357 variables).

A. Sparse PLS

SPLS finds a pair of sparse weight vectors $\boldsymbol{u}$ and $\boldsymbol{v}$, such that the covariance between the projections of $\boldsymbol{X}$ and $\boldsymbol{Y}$ onto these weight vectors is maximized. This allows variable selection and modelling in a one-step procedure, which potentially improves the interpretability of the results and avoids overfitting [2]. Witten et al. [3] proposed a framework called Penalized Matrix Decomposition (PMD) which was then modified to create the following sparse version of PLS:

$$\begin{gathered} \text{maximise}_{\boldsymbol{u},\boldsymbol{v}} \ \boldsymbol{u}^{\boldsymbol{T}} \boldsymbol{X}^{\boldsymbol{T}} \boldsymbol{Y} \boldsymbol{v} \\ \text{subject to} \\ \|\boldsymbol{u}\|_2^2 \leq 1, \ \|\boldsymbol{v}\|_2^2 \leq 1, \ \|\boldsymbol{u}\|_1 \leq c_u, \ \|\boldsymbol{v}\|_1 \leq c_v \end{gathered} \tag{1}$$

where the sparse weight vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ have the same length as the number of features of the corresponding view, i.e $\boldsymbol{u} \in \mathbb{R}^{p \times 1}$ and $\boldsymbol{v} \in \mathbb{R}^{q \times 1}$ and several entries equal zero. The regularisation hyper-parameters c_u and c_v control the L_1-norm constraints of $\boldsymbol{u}$ and $\boldsymbol{v}$, respectively. If c_u and c_v are sufficiently small, the L_1-norm constraints imposes sparsity on the corresponding view and consequently fewer features are included in the model. The values of c_u and c_v can be simply chosen according to the desired amount of sparsity imposed on $\boldsymbol{u}$ and $\boldsymbol{v}$ or using a grid search analysis [3], [4]. The hyper-parameters must be chosen in $1 \leq c_u \leq \sqrt{p}$ and $1 \leq c_v \leq \sqrt{q}$ to both L_1-norm and L_2-norm constraints be active [2]. Each pair of weight vectors represent a multivariate associative effect between the two views. The problem described in Equation 1 can be solved using the SPLS algorithm that can be found in [4].

978-1-5386-4291-7/18/$31.00 ©2018 IEEE

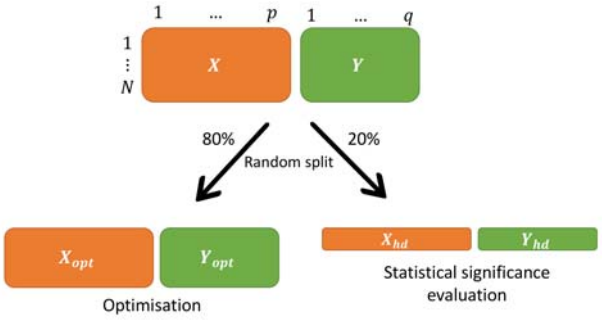

Fig. 1. Random split of data into the optimisation and hold-out sets.

B. Learning frameworks

The learning frameworks consist of 3 parts: hyper-parameter optimisation, statistical significance evaluation and matrix deflation. Only the first one is different across frameworks. To evaluate the generalisability of each framework, the data matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$ were first randomly split into an optimisation set (80% of the data, *opt*) and a hold-out set (20% of the data, *hd*) (Fig. 1). The former was used for optimising the hyper-parameters and the latter was used for validating the model.

1) Hyper-parameter optimisation: For both frameworks, the hyper-parameter values were optimised by performing a grid-search analysis, in which 20 equidistant points in $1 \leq c_u \leq \sqrt{p}$ and $1 \leq c_v \leq \sqrt{q}$ were chosen.

Permutation framework: For each hyper-parameter combination $\{c_{u_j}, c_{v_j}\}$, the weight vectors $\boldsymbol{u_j}$ and $\boldsymbol{v_j}$ are computed using the optimisation set. Then the correlation ρ_j between the projections of $\boldsymbol{X_{opt}}$ and $\boldsymbol{Y_{opt}}$ onto $\boldsymbol{u_j}$ and $\boldsymbol{v_j}$ is computed: $\rho_j = Corr(\boldsymbol{X_{opt}} \boldsymbol{u_j}, \boldsymbol{Y_{opt}} \boldsymbol{v_j})$ (Fig. 2A).

The rows of $\boldsymbol{Y_{opt}}$ are randomly permuted (number of permutations $B = 1000$) to obtain the matrix $\boldsymbol{Y^b_{opt}}$, where $b \in 1, ..., B$. Weight vectors $\boldsymbol{u}_j^b$ and $\boldsymbol{v}_j^b$ are then computed and correlations ρ_j^b between the projections of $\boldsymbol{X_{opt}}$ and $\boldsymbol{Y^b_{opt}}$ onto $\boldsymbol{u}_j^b$ and $\boldsymbol{v}_j^b$ are computed (Fig. 2A) [3]. Finally, the p-value for ρ_j is calculated:

$$p = \frac{1 + \sum_{b=1}^{B} 1_{\rho_j^b \geq \rho_j}}{B + 1} \tag{2}$$

The hyper-parameter pair with the lowest p-value ($p_{unc} < 0.001$) is chosen. However, it is likely that several combinations have the same p-value and then a second criteria must be used. Therefore, the hyper-parameter combination with the largest distance between the true correlation and the null distribution of the correlations ($d_j = \frac{\rho_j - \frac{1}{B}\sum_{b=1}^{B} \rho_j^b}{sd(\rho_j^B)}$, where $sd(\rho_j^B)$ indicates the standard deviation of $\rho_j^1, ..., \rho_j^B$) is chosen [3]. The best hyper-parameter pair is finally passed for use in the statistical significance evaluation. (Fig. 3).

Train/test framework: For each hyper-parameter combination, the optimisation set is randomly split $K = 50$ into a training set (80%) and a testing set (20%) (Fig. 2B).

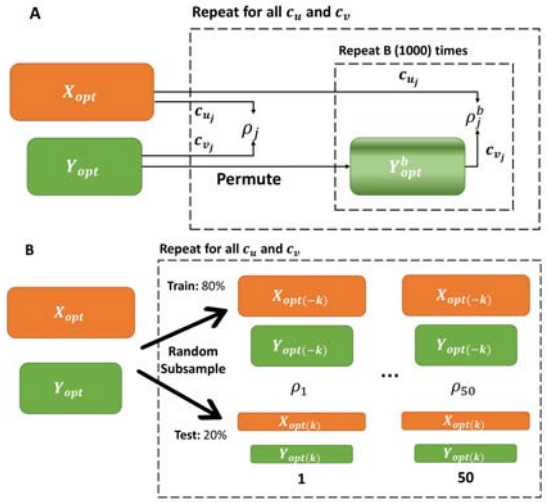

Fig. 2. Hyper-parameter optimisation step for (A) permutation and (B) train/test frameworks.

The weight vectors pairs ($\boldsymbol{u_{(-k)}}$ and $\boldsymbol{v_{(-k)}}$) are computed using the training set ($\boldsymbol{X_{opt_{(-k)}}}$ and $\boldsymbol{Y_{opt_{(-k)}}}$) and the test correlation is computed by projecting the testing data ($\boldsymbol{X_{opt_{(k)}}}$ and $\boldsymbol{Y_{opt_{(k)}}}$) onto these weights for each split k [4]:

$$\rho_k = \text{Corr}(\boldsymbol{X_{opt_{(k)}}}\boldsymbol{u_{(-k)}}, \boldsymbol{Y_{opt_{(k)}}}\boldsymbol{v_{(-k)}}) \quad (3)$$

Then, the K correlation values are averaged across splits for each hyper-parameter combination (arithmetic mean: $\bar{\rho_j} = \frac{1}{K}\sum_{k=1}^{K}\rho_k$). This procedure is repeated for all hyper-parameter combinations and the one with the highest average test correlation is selected (Fig. 2B). As multiple hyper-parameter combinations can have the same average correlation, the sparsest one is chosen [4]. The best hyper-parameter pair is then passed for use in the statistical significance step (Fig. 3).

2) Statistical significance evaluation: For both frameworks, the statistical significance of the associative effect is assessed using a permutation test. Firstly, the model is trained with the best hyper-parameter pair using the optimisation set (Fig. 3). Then, the hold-out set ($\boldsymbol{X_{hd}}$ and $\boldsymbol{Y_{hd}}$) is projected onto the new weight vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ and the hold-out correlation between the projections is calculated ($\rho_{hd} = \text{Corr}(\boldsymbol{X_{hd}u}, \boldsymbol{Y_{hd}v})$) (Fig. 3). After that, $\boldsymbol{Y_{opt}}$ is permuted and the model is trained with the best hyper-parameter pair for each permutation $\boldsymbol{m}$, where $m \in 1, ..., M$. The hold-out set is projected onto the computed weight vectors ($\boldsymbol{u_m}$ and $\boldsymbol{v_m}$) and the correlation between the projections is computed ($\rho^m_{hd} = \text{Corr}(\boldsymbol{X_{hd}u_m}, \boldsymbol{Y_{hd}v_m})$). The process is repeated $M = 10000$ times and a p-value for the hold-out correlation can be computed using Equation 2.

In neuroimaging settings, the samples sizes are small and therefore few samples may be included in the hold-out set, which can lead to high variance in the results since the validation is dependent on how the data is split. To make the model validation more robust, multiple hold-out sets (here 10 random splits of the data) are used [4]. The *omnibus hypothesis* is used to evaluate if any of the hold-out sets is statistically significant [4]. The p-values are corrected for multiple comparisons using Bonferroni correction ($\alpha = 0.05/10 = 0.005$), which means that the omnibus hypothesis is rejected if any $p_{corr} \leq 0.005$. The weight vector pair with lowest p-value is chosen to deflate the data matrices among the significant pairs.

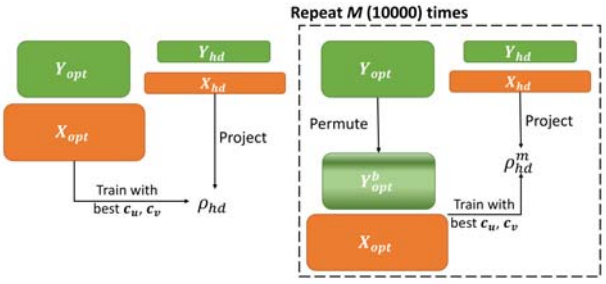

Fig. 3. Statistical significance evaluation step.

3) Matrix deflation: If any of the weight vector pairs is considered statistically significant, then the effect explained by that weight vector (e.g. $\boldsymbol{u}_h$ or $\boldsymbol{v}_h$) must be removed from the data to allow the finding of new potential effects (e.g. $\boldsymbol{u}_{h+1}$ or $\boldsymbol{v}_{h+1}$). The process is known as *matrix deflation*. Here, we used the projection deflation method proposed by Mackey [9] and tested in a similar multiple hold-out framework by Monteiro et al [4].

IV. Results

The frameworks were compared in terms of the obtained associative effects ($\boldsymbol{u}$ and $\boldsymbol{v}$) and generalisability, measured by the hold-out correlation values.

A. Weight vectors or associative effects

Two statistically significant associative effects were found using the train/test framework and only one was found with the permutation framework. Fig. 4 shows non-zero weights of the first associative effect obtained using both frameworks. For visualisation purposes, only the top 5 positive and negative behavioural variables associated with the first associative effect (the variables with the highest weights) are shown. Although the behavioural weight vectors were very similar ($\rho_{pearson} = 0.84$), there is a considerable difference in terms of sparsity in the brain connectivity weight vectors ($\rho_{pearson} = 0.41$). Indeed, only one brain connectivity variable is contributing to the associative effect in the permutation framework (Fig. 4).

B. Generalisability of the frameworks

Table I shows the hold-out correlations of the 10 different splits of the data for the first and second effect, for each framework. In both effects, the significant hold-out correlations (i.e. $p < 0.005$) were consistently higher in the train/test framework than the ones in the permutation framework.

V. Discussion

The most common framework for optimising hyper-parameters in unsupervised sparse models, such as the SPLS, is the permutation framework [3]. However, an important

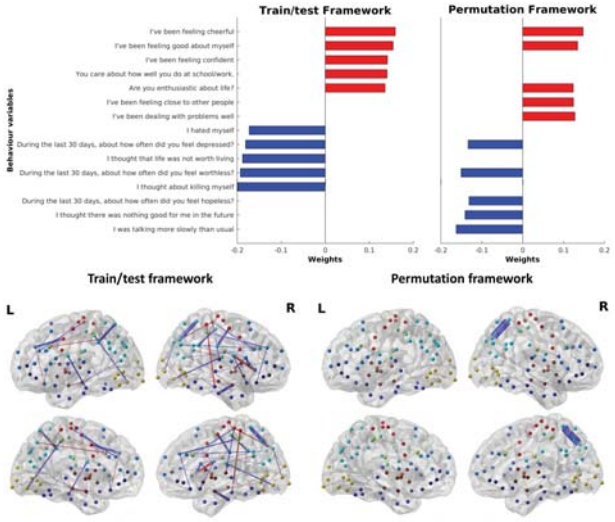

Fig. 4. The top 5 non-zero positive and negative behavioural variables (top) and all non-zero brain connections (bottom) associated with the first associative effect for both frameworks. L - left hemisphere; R - right hemisphere.

TABLE I. HOLD-OUT CORRELATIONS AND P-VALUES OF 10 DIFFERENT SPLITS OF THE DATA FOR THE FIRST AND SECOND ASSOCIATIVE EFFECT USING BOTH FRAMEWORKS.

	Train/test framework		**Permutation framework**	
	First effect	Second effect	First effect	Second effect
Split	ρ_{hd} (p_{hd})	ρ_{hd} (p_{hd})	ρ_{hd} (p_{hd})	ρ_{hd} (p_{hd})
1	**0.53 (0.0018)**	0.31 (0.0529)	0.40 (0.0220)	-0.06 (0.6967)
2	**0.70 (0.0002)**	0.42 (0.0372)	**0.37 (0.0008)**	-0.04 (0.6055)
3	0.37 (0.0214)	0.25 (0.0590)	0.39 (0.0448)	-
4	0.35 (0.0159)	0.17 (0.1741)	0.42 (0.0213)	-
5	**0.64 (0.0001)**	0.28 (0.0641)	0.39 (0.0171)	-
6	**0.51 (0.0018)**	0.53 (0.0100)	0.38 (0.0131)	-
7	**0.44 (0.0012)**	0.36 (0.0092)	0.39 (0.0665)	-
8	**0.68 (0.0003)**	0.42 (0.0454)	0.37 (0.1887)	-
9	0.45 (0.0104)	**0.38 (0.0046)**	0.39 (0.0066)	-
10	0.44 (0.0098)	**0.54 (0.0020)**	0.39 (0.1244)	-

limitation of this framework is that the whole data is used to fit the model and often an out of sample model performance (e.g. out of sample correlation) is not available. Recently a new framework based on multiple splits of the data has been proposed to address this limitation [4]. In the present work, we compared these two frameworks for optimising the SPLS parameters using a hold-out framework to access their generalisability in terms of hold-out correlation.

Both frameworks were able to identify at least one brain-behaviour associative effect. However, for the first associative effect, the weight vectors had different levels of sparsity across frameworks, particularly the brain connectivity weight vectors. The observed differences in sparsity level can be related to the different criteria used to optimize the hyper-parameters but might also be due to the signal to noise ratio in the data. The second associative effect was only found using the train/test optimisation framework. This might be explained by the fact that the permutation-based framework is not generalising well (i.e. the average hold-out correlation is close to zero for the second weight vector pair). Despite of having only one significant split, the hold-out correlation values were more stable across different splits for permutation framework. This effect is expected because, for the permutation framework, all examples in the optimisation set are used for optimising the hyper-parameters, whereas for the train/test framework, the data is randomly split into train/test many times during the optimisation, which introduces greater variability in the model. In terms of computational costs, the train/test optimisation framework (20001 computations per split) is much more efficient than the permutation optimisation framework (400400 computations per split).

In summary, a careful choice of how to optimise the model's hyper-parameters should always be made, because different criteria and frameworks might have a strong influence on the results. As expected, optimising hyper-parameters using a metric based on test data (test correlation) leads to a better generalisability than using metrics a based on the whole data.

ACKNOWLEDGEMENTS

F.F. would like to acknowledge J. Monteiro and J. Schrouff for the support provided for this work. F.F. was supported by a PhD fellowship awarded by FCT (SFRH/BD/120640/2016). M.M. receives support from the UCLH Biomedical Research Centre. R.D. is supported by a Wellcome Investigator Award (ref 098362/Z/12/Z). J.M.M. was supported by the Wellcome Trust (UK) under the grant number WT102845/Z/13/Z. All NSPN consortium authors are supported by Wellcome Strategic Award (ref 095844/7/11/Z).

REFERENCES

[1] S. M. Smith, T. E. Nichols, D. Vidaurre, A. M. Winkler, T. E. J. Behrens, M. F. Glasser, K. Ugurbil, D. M. Barch, D. C. Van Essen, and K. L. Miller, "A positive-negative mode of population covariation links brain connectivity, demographics and behavior," *Nature Neuroscience*, vol. 18, no. 11, pp. 1565–1567, 2015.

[2] D. M. Witten, R. Tibshirani, and T. Hastie, "A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis," *Biostatistics*, vol. 10, no. 3, pp. 515–534, 2009.

[3] D. M. Witten and R. J. Tibshirani, "Extensions of Sparse Canonical Correlation Analysis with Applications to Genomic Data," *Statistical Applications in Genetics and Molecular Biology*, vol. 8, no. 1, p. 29, 2009.

[4] J. M. Monteiro, A. Rao, J. Shawe-Taylor, and J. Mourão-Miranda, "A multiple hold-out framework for Sparse Partial Least Squares," *Journal of Neuroscience Methods*, vol. 271, no. 271, pp. 182–194, 2016.

[5] B. Kiddle, B. Inkster, G. Prabhu, M. Moutoussis, K. J. Whitaker, E. T. Bullmore, R. J. Dolan, P. Fonagy, I. M. Goodyer, and P. B. Jones, "Cohort profile: The NSPN 2400 Cohort: a developmental sample supporting the Wellcome Trust NeuroScience in Psychiatry Network," *International Journal of Epidemiology*, 2017.

[6] M. Perrot, D. Rivière, A. Tucholka, and J.-F. Mangin, "Joint Bayesian Cortical Sulci Recognition and Spatial Normalization." Springer, Berlin, Heidelberg, 2009, pp. 176–187.

[7] R. S. Desikan, F. Ségonne, B. Fischl, B. T. Quinn, B. C. Dickerson, D. Blacker, R. L. Buckner, A. M. Dale, R. P. Maguire, B. T. Hyman, M. S. Albert, and R. J. Killiany, "An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest," *NeuroImage*, vol. 31, no. 3, pp. 968–980, jul 2006.

[8] S. M. Smith, K. L. Miller, G. Salimi-Khorshidi, M. Webster, C. F. Beckmann, T. E. Nichols, J. D. Ramsey, and M. W. Woolrich, "Network modelling methods for FMRI," *NeuroImage*, vol. 54, no. 2, pp. 875–891, 2011.

[9] L. Mackey, "Deflation Methods for Sparse PCA." in *Advances in Neural Information Processing Systems (NIPS)*, 2008, pp. 1017–1024.