

Determining biotypes of PTSD patients using sparse canonical-correlation analysis between clinical and resting-state fMRI functional connectivity measures

Paul Zhutovsky

Mirjam van Zuiden

Leonardo Cerliani

Rajat Thomas

Miranda Olff

Guido van Wingen

Academic Medical Center, Amsterdam

**Background:** Posttraumatic stress disorder (PTSD) is a heterogeneous disorder which can be diagnosed in many different ways according to the DSM-IV or DSM-5 manual [1]. Because of this clinical heterogeneity it is important to determine subtypes of PTSD patients that share clinical and biological properties. A recent paper [2] showed that using canonical-correlation analysis (CCA) [3] between clinical measures and functional connectivity (FC) estimated from resting-state(rs-) fMRI one can identify neurophysiological subtypes ("biotypes") in a fully data-driven way for patients with major depression disorder (MDD). Similar to MDD, PTSD patients also have shown abnormalities in rs-fMRI connectivity within the default mode (DMN), saliency network (SN) and the central executive network (CEN) [4, 5] which points to rs-fMRI as a promising measure in determining biotypes in PTSD. In addition the altered activity of DMN, SN and CEN has been proposed to be associated with different profiles of clinical symptoms [5]. Therefore the combination of rs-fMRI and clinical data seems to be an optimal way to tackle the heterogeneity of PTSD and determine sensible subtypes.

**Goal:** Using a similar approach to [2] we will utilize the full resting-state fMRI data of the ENIGMA-PGC PTSD consortium by:

1. Finding multivariate associations between the clinical data (measured via the CAPS interview) and rs-fMRI functional connectivity (FC) using sparse CCA of PTSD patients
2. Cluster the canonical FC space to obtain biotypes of PTSD patients
3. Determine clinical and FC differences of the individual biotypes
4. Verify the discovered differences in hold-out data to determine test-retest reliability

**Methods:** Rs-fMRI will be preprocessed according to the ENIGMA rs-fMRI pipeline [6] and parcellated using a high-resolution parcellation (such as the Power parcellation [7] adjusted to include subcortical regions such as amygdala, ACC, etc.). Functional connectivity (FC) will be calculated using pairwise Pearson linear correlations between the individual ROIs and Fisher's r-to-z transformations will be applied. Nuisance regression with sites, age, etc. will be performed and the cleaned FC measures will be entered into a sparse-CCA analysis [8-10] with individual CAPS scores as clinical measures. Only patients with available CAPS scores will be included in this part of the analysis.

CCA determines multivariate linear combinations of clinical and FC data which are maximally correlated. However, in the case where more features than subjects are present in the data CCA cannot find unique canonical weights and becomes unstable [9, 10]. Therefore, we will use sparse-CCA which includes a penalty for the canonical weights forcing some of their values to become exactly zero. This also increases the interpretability of the model because not all FC/CAPS scores will be part of the canonical variable. The level of sparsity can be determined in a data-driven way by either performing permutation testing [8, 9] or cross-validation [11] for each modality individually. Also the amount of canonical variables to choose can be found in a formalized way using permutation testing [12]. The FC data will then be transformed into its canonical space creating a single brain score per canonical weight and clustering will be performed. We will apply common clustering algorithms such as hierarchical clustering or k-means clustering and estimate the number of clusters in a data-driven way using measures such as the Silhouette score [13] or the gap statistic [14].

After the optimal number of cluster is identified we will assign the individual patients to their biotypes and perform statistical tests on their clinical and FC measures. In that way we will try to determine group differences between patients belonging to different biotypes. The discovered differences will then be verified in patients which were not used for the discovery of biotypes.

**Preliminary results:** As a proof-of-concept we applied the proposed analysis plan to data of 32 PTSD patients (16 male/16 female) from the AMC Amsterdam cohort. The sample is too small to obtain any reliable results and the results are presented here for visualization purposes only. Figure 1 shows the results of the sparse-CCA analysis using 2 factors with their respective canonical correlations. In Figure 2 we show the obtained clusters for a 2 cluster solution using hierarchical clustering. Figure 3 shows the average FC of patients in the two clusters. These preliminary results show the feasibility of the approach. However, these results cannot be seen as robust because of the small sample size and therefore require the full ENIGMA-PGC PTSD cohort to enhance our understanding of the heterogeneity of PTSD.

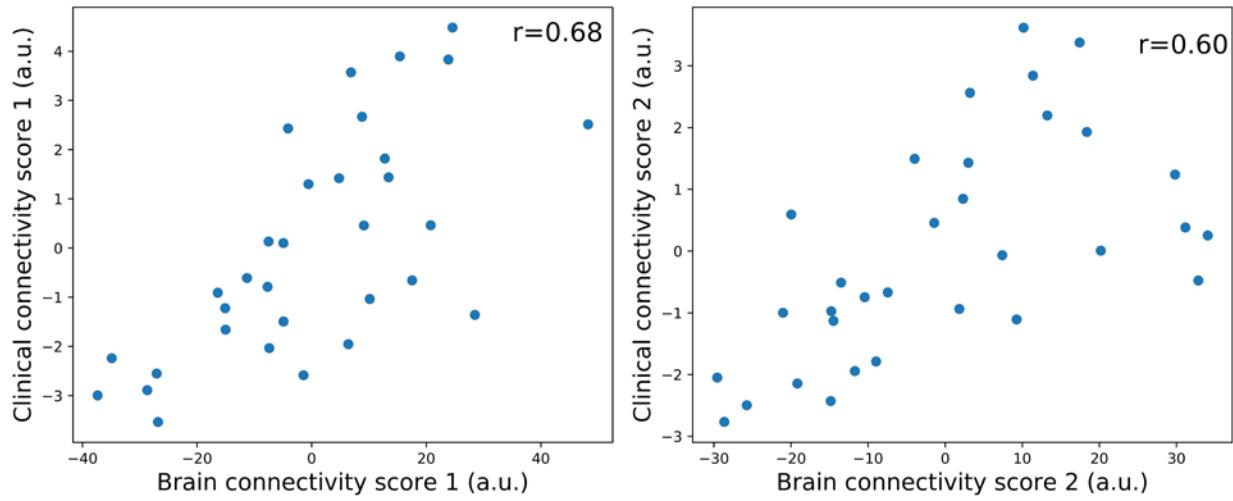


Figure 1: Results of the sparse CCA analysis with 2 factors

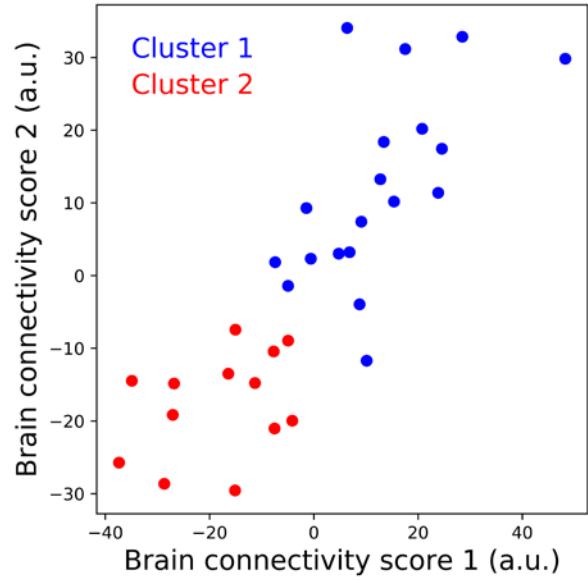
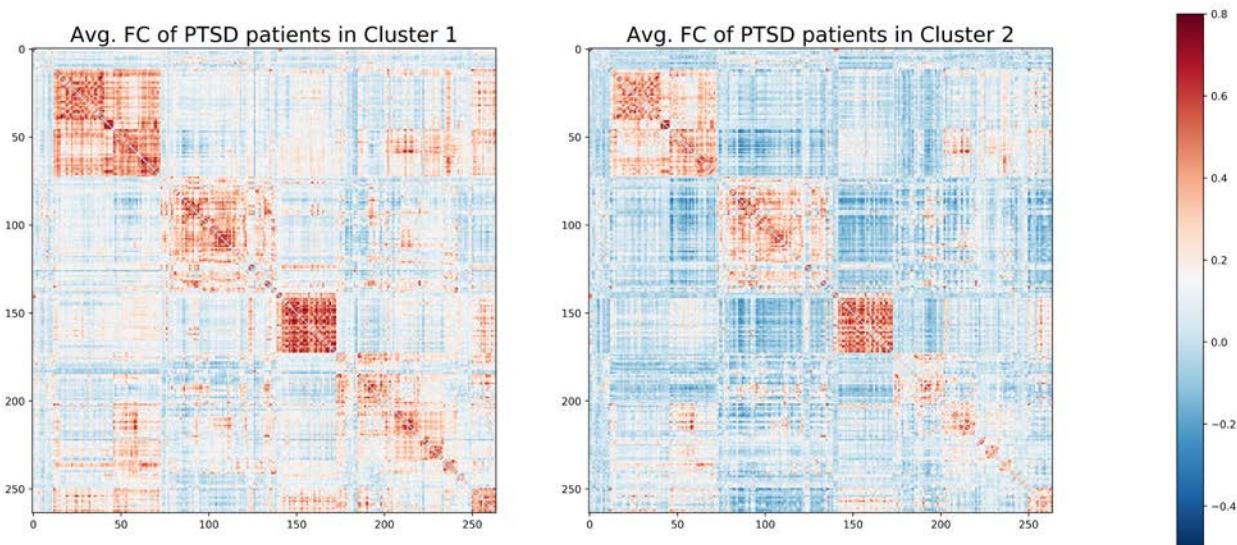


Figure 2: Clustering results based on two sparse-CCA FC factors



**Figure 3: The FC of individual clusters**

**References:**

1. Galatzer-Levy, I.R. and R.A. Bryant, *636,120 Ways to Have Posttraumatic Stress Disorder*. Perspectives on Psychological Science, 2013. **8**(6): p. 651-662.
2. Drysdale, A.T., et al., *Resting-state connectivity biomarkers define neurophysiological subtypes of depression*. Nature Medicine, 2016. **23**(1): p. 28.
3. Hotelling, H., *Relations Between Two Sets of Variates*. Biometrika, 1936. **28**(3/4): p. 321.
4. Koch, S.B., et al., *Aberrant Resting-State Brain Activity in Posttraumatic Stress Disorder: A Meta-Analysis and Systematic Review*. Depress Anxiety, 2016. **33**(7): p. 592-605.
5. Lanius, R.A., et al., *Restoring large-scale brain networks in PTSD and related disorders: a proposal for neuroscientifically-informed treatment interventions*. Eur J Psychotraumatol, 2015. **6**: p. 27313.
6. Adhikari, B.M., et al., *Heritability estimates on resting state fMRI data using ENIGMA analysis pipeline*. Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing, 2018. **23**: p. 307-318.
7. Power, J.D., et al., *Functional network organization of the human brain*. Neuron, 2011. **72**(4): p. 665-78.
8. Witten, D.M., R. Tibshirani, and T. Hastie, *A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis*. Biostatistics, 2009. **10**(3): p. 515-34.
9. Witten, D.M. and R.J. Tibshirani, *Extensions of sparse canonical correlation analysis with applications to genomic data*. Stat Appl Genet Mol Biol, 2009. **8**(1): p. Article28.
10. Avants, B.B., et al., *Sparse canonical correlation analysis relates network-level atrophy to multivariate cognitive measures in a neurodegenerative population*. Neuroimage, 2014. **84**: p. 698-711.
11. Bilenko, N.Y. and J.L. Gallant, *Pyrcca: Regularized Kernel Canonical Correlation Analysis in Python and Its Applications to Neuroimaging*. Front Neuroinform, 2016. **10**: p. 49.

12. Rosa, M.J., et al., *Estimating multivariate similarity between neuroimaging datasets with sparse canonical correlation analysis: an application to perfusion imaging*. Front Neurosci, 2015. **9**(366): p. 366.
13. Rousseeuw, P.J., *Silhouettes - a Graphical Aid to the Interpretation and Validation of Cluster-Analysis*. Journal of Computational and Applied Mathematics, 1987. **20**: p. 53-65.
14. Tibshirani, R., G. Walther, and T. Hastie, *Estimating the number of clusters in a data set via the gap statistic*. Journal of the Royal Statistical Society Series B-Statistical Methodology, 2001. **63**(2): p. 411-423.