The Impact on Inter-Subject Discriminability of Dimensionality Reduction in fMRI

Eric Bridgeford (ebridge2@jhu.edu), Tanay Agarwal (tagarwa2@jhu.edu)

November 4, 2016

1 Abstract

The twenty first century has seen the rapid ascent of technology into everyday life. From high performance desktops, to smart phones, to super computers, many industries rely on the power and consistency of computing. The psychiatry field, on the other hand, continues to rely on psychological examinations almost exclusively when determining factors such as rate of development, mental disorder, and overall mental health in millions of patients each year. Standards of care and diagnosis vary widely between psychiatrists, and no quantitative tools exist to aid in their tasks. Recently, connectomics has emerged as a growing branch of computational neuroscience that seeks to ultimately map the brain and enable quantitative analyses of psychiatric function. fMRI, one such method, measures the rate at which neurons absorb oxygen as part of the haemodynamic response. Throughout the course of an fMRI scanning session, individual scans may be corrupted by a wide variety of noise variables with respect to the brain activation at a point in time; notably, scanning related factors (head motion, scanner error, etc), and processing-imposed factors (assumptions made throughout the course of analysis that add bias or variance) [1]. This corruption can disrupt the true recovered signal for a particular individual, and the removal of it has prompted many inquiries into the ability one has of discriminating recovered signals for individuals [2]. Through this project, the investigators will expand upon a summer project by Eric Bridgeford, the FNGS (Functional Neurodata Graphs Service) pipeline, and examine the impact of dimensionality reduction on the discriminability of the graphs obtained.

2 Methods

2.1 Machine Learning Methods

1. PCA (Principal Component Analysis): can concentrate the majority of the signal of a particular scan into the first few principal components, and later components can be removed to reduce the dimensionality (and correspondingly, since the later components have less signal, they can be thought of as higher noise components) [3].
2. CompCor (Component-based Noise Correction): major components are calculated from regions of the brain known to be higher noise (such as the cerebrospinal fluid) and removed from the scan, since these regions will contribute very little relevant signal to the timeseries [4].

2.2 Analysis Statistics

1. Discriminability: The researchers will use a statistic developed by Wang et. al [2], the discriminability, to compare the scan-scan reliability of the fMRI graphs obtained with and without di-

mensionality reduction. The discriminability separates this study as largely novel; while previous investigations have studied PCA and CompCor for fMRI processing on an individual graph basis, the discriminability allows a dataset-wide comparison of the degree to which individuals scanned repeatedly best resemble themselves (essentially, fingerprinting of brain scans).

3 Resources

1. The project will add a module to the FNGS pipeline, which is a python library for fMRI processing developed by Eric Bridgeford. The pipeline is written in python, and utilizes many python libraries; particular libraries that will be explicitly necessary for this module will be numpy and scipy (acquisition of the brain graphs). Post processing will be done using R (analysis of brain graphs and the discriminability). This pipeline (an fMRI pipeline), is currently being merged into our Diffusion MRI pipeline built by a graduate student, which together will become the NDMG pipeline [5].

2. The researchers will use several existing multi scan fMRI datasets (part of the CoRR [6] collection), notably, the KKI dataset, the NKI24 dataset, the HNU1 dataset, the DC1 datset, and the SWU4 dataset. The project will only include the BNU1 dataset, and the deadline by end of January (outside of the class) will be all 5 datasets.

3. Simulated data developed from the randomts R package [7].

4. MNI lateral ventricles, csf, and gray matter masks [8].

4 Final Writeup

The final writeup will feature the following organization:

1. fMRI, their relevance to the psych community, what's been done so far with fMRI processing, FNGS pipeline in existing state

2. explanation of the relevance of PCA and CompCor to fMRI processing, and how they reduce noise.

3. Think of a simulation in which this approach would be beneficial (ie, when we have a strong signal concentrated to few components), and one where it could be detrimental (signal about equal throughout all components) and show the impact on the explained variance (ie, is there an elbow when we have the signal concentrated in 1 or 2 components? is it lower for more components if we have signal spread relatively equally throughout all components?) (timeseries simulated with RandomTS package).

4. Real-data results (data from the CoRR dataset).

5. Research questions to be answered by real data:

   (a) research questions: How can we best choose principal components? a particular proportion of variance explained (ie, thresholding to take however many components explain 80% of the true variance), or by a particular number of components (ie, statically taking the top 5 components)?

   (b) how should we do CompCor (ie, which masks do we use for the brain)?

6. How we will answer these:

   (a) Testing Strategy: brute force. Take any possible options we could choose in our processing pipeline, and run every other test, run with every other possible option, for each item in the particular parameter we are looking at. For example, if we have 2 options, each with 2 possible features (on or off), we would run all data through 4 pipelines.

i. Naive Approach
   A. PCA: run with and without PCA (choosing an arbitrary number of components). (2 options)
   B. CompCor: run with and without CompCor (using the gray matter, csf, and lateral ventricles masks) (2 options).
   C. number of pipelines: $2 * 2 = 4$.

ii. As Time Allows Approach
   A. PCA approach: Run PCA with a medium number of components (ie, taking the top 6 components), PCA with a thresholded amounted of explained variance (ie, taking the n components that lead to 70% of the variance explained, and no PCA. (3 feature options).
   B. Comp Cor: Run CompCor with ventricles, gray matter, and csf masks (all at once) and compare to without comp cor.
   C. number of pipelines: $3 * 2 = 6$.

iii. Advanced Approach
   A. PCA approach: Run 5 independent PCA options on the real data. One will be with no PCA, one will be with PCA and taking a fixed number of components (ie, take the top 4 components), another will be with PCA and taking a larger number of components (ie, taking top 8 components), another will be with a low threshold of explained variance (ie, take the top n components that explain 60% of the variance), and finally one with a high threshold of explained variance (ie, take the top n components that explain 80% of the variance).
   B. CompCor approach: Each option is defined as including or excluding a particular mask from comp cor (where no masks is no compcor, all masks is running compcor for each mask individually and removing all nuisance components). These are the sets of length one options (each individually), the 3 element set of options (all masks at once), and no CompCor. Here we have 5 options.
   C. number of pipelines: $5 * 5 = 25$.

(b) Answers will be interpreted as:

   i. will obtain a discriminability score for each pipeline. The best option from each category (PCA and CompCor) will be identified, and the best overall strategy (combination of options with the best discriminability score) will be reported.

* Note that figures we cannot fit in a 6 page range will be included in a supplement for thoroughness.

5 Bibliography

References

[1] Thomasson D Soltysik, D and, S Rajan, and Biassou N. Improving the use of principal component analysis to reduce physiological noise and motion artifacts to increase the sensitivity of task-based fMRI. *Neuroscience Methods*, February 2015.

[2] Shangsi Wang, Zhi Yang, Xi-Nian Zuo, Michael Milham, Cameron Craddock, Gregory Kiar, William Gray Roncal, Eric Bridgeford, Carey E Priebe, and Joshua T Vogelstein. Optimal Decisions for Discovery Science via Maximizing Discriminability: Applications in Neuroimaging. *request for preprint*.

[3] Dimensionality Reduction. *Wikipedia*. Accessed: 2016-11-04.

[4] Y Behzadi, K Restom, J Liau, and T T Liu. A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *Neuroimage*.

[5] Gregory Kiar, Eric Bridgeford, William R Roncal, Randal Burns, Joshua T Vogelstein, et al. NDMG: he NeuroData Mri Graphs. *GitHub*, 2016.

[6] K J Gorgolewski, N Mendes, D Wilfling, E Wladimirow, C.J. Gauthier, T Bonnen, R Trampel, P L Bazin, and D S Margulies. Measuring variability of human brain activity at rest - a high resolution 7-Tesla test-retest dataset.

[7] Eric W Bridgeford. Random Timeseries R Package, 2016.

[8] G Grabner, A L Janke, MM Budge, D Smith, J Pruessner, and D L Collins. Symmetric atlasing and model based segmentation: an application to the hippocampus in older adults. *Med Image Comput Comput Assist Interv Int Conf Med Image Comput Comput Assist Interv*, 9, 2006.