Generalized Canonical Correlation Analysis with applications to fMRI reproducibility

Ronan Perry

Johns Hopkins University

August 15, 2019

Statistical Parametric Maps (SPMs)

- An experiment yields subject data matrices $Y_k \in \mathbb{R}^{n \times t_k}$, $1 \leq k \leq N$ for some set of experimental conditions.

Statistical Parametric Maps (SPMs)

- An experiment yields subject data matrices $Y_k \in \mathbb{R}^{n \times t_k}$, $1 \leq k \leq N$ for some set of experimental conditions.
- Assumption: the activity of each voxel is, under the null, distributed according to a known density (usually t- or f-distributions)
- Can compare control and experimental groups by performing univariate voxel-wise tests for significance

SPMs: General Linear Model

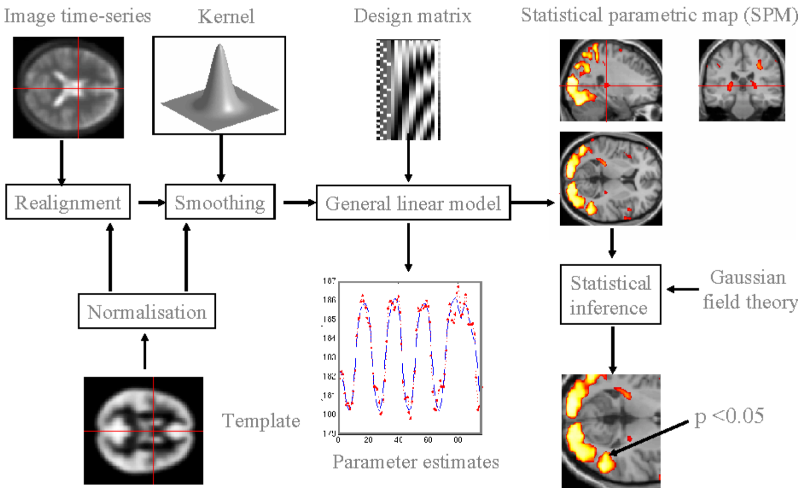

Figure: General linear model and random field theory for statistical inference.

Motivation (applications paper link)

- A key to experiment reproducibility is that the same spatial maps be generated across replications
- Studies often seek significant p-values for activity detection, but usually ignore the need for reproducible spatial patterns
- One problem is that they often parameterize the BOLD response function, not consistent across individuals.

Reproducibility of processing pipelines

- Many ways to preprocess and analyze fMRI data
- Attempts to improve reproducibility
  - Extensions to univariate approaches
  - Multivariate approaches
- Authors' assumptions: the subjects share an unknown spatial map but show different temporal responses to a task.
- Goal is to use a multivariate approach to learn a reproducible spatial map shared by each subject

Multiview Learning

- Given some data, we want to learn a representation
- But if there are data from multiple views (ie. image and text), the learning should account for similarities and differences between the views.

Multiview Learning

- Given some data, we want to learn a representation
- But if there are data from multiple views (ie. image and text), the learning should account for similarities and differences between the views.
- Alignment: each view maps to the close-to-same representation

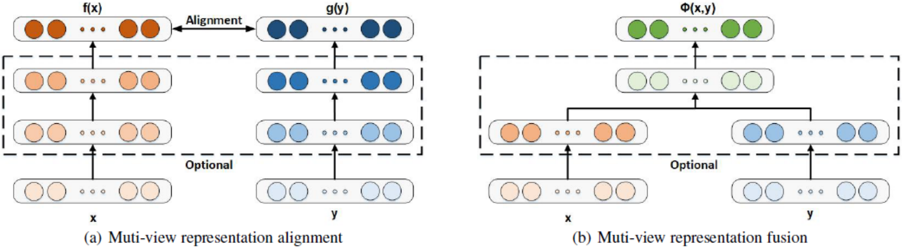

Figure: Alignment vs. Fusion methods

Canonical Correlation Analysis (CCA)

- Given data matrices $X_1 \in \mathbb{R}^{n \times t_1}, X_2 \in \mathbb{R}^{n \times t_2}$

Canonical Correlation Analysis (CCA)

- Given data matrices $X_1 \in \mathbb{R}^{n \times t_1}, X_2 \in \mathbb{R}^{n \times t_2}$
- Goal is to projections of X_1, X_2 whose correlations are maximized

Canonical Correlation Analysis (CCA)

- Given data matrices $X_1 \in \mathbb{R}^{n \times t_1}, X_2 \in \mathbb{R}^{n \times t_2}$
- Goal is to projections of X_1, X_2 whose correlations are maximized
- Let $z_1 = X_1 a_1$ and $z_2 = X_2 a_2$

Canonical Correlation Analysis (CCA)

- Given data matrices $X_1 \in \mathbb{R}^{n \times t_1}, X_2 \in \mathbb{R}^{n \times t_2}$
- Goal is to projections of X_1, X_2 whose correlations are maximized
- Let $z_1 = X_1 a_1$ and $z_2 = X_2 a_2$
  $(a_1, a_2) = argmax\left(\frac{z_1^T z_2}{\|z_1\|\|z_2\|}\right)$

Canonical Correlation Analysis (CCA)

- Given data matrices $X_1 \in \mathbb{R}^{n \times t_1}, X_2 \in \mathbb{R}^{n \times t_2}$
- Goal is to projections of X_1, X_2 whose correlations are maximized
- Let $z_1 = X_1 a_1$ and $z_2 = X_2 a_2$
  $(a_1, a_2) = argmax\left(\frac{z_1^T z_2}{\|z_1\|\|z_2\|}\right)$
- Equivalent to
  $(a_1, a_2) = argmax(a_1^T C_{12} a_2)$
  s.t. $a_1^T C_{11} a_1 = a_2^T C_{22} a_2 = 1$
- Comes down to solving an eigenvalue decomposition problem

Extensions

- Sparse CCA
  - Force sparsity of the projections a_1, a_2
  - Incorporate regularization terms

Extensions

- Sparse CCA
  - Force sparsity of the projections a_1, a_2
  - Incorporate regularization terms
- Kernel CCA
  - Incorporate nonlinearities

Extensions

- Sparse CCA
  - Force sparsity of the projections a_1, a_2
  - Incorporate regularization terms
- Kernel CCA
  - Incorporate nonlinearities
- CCA for more than two data matrices

Generalized Canonical Correlation Analysis (GCCA)

- What there are more than two data matrices, i.e. X_k for $1 \leq k \leq N$

Generalized Canonical Correlation Analysis (GCCA)

- What there are more than two data matrices, i.e. X_k for $1 \leq k \leq N$
- We seek a generalization, equivalent to CCA in the two-sample case.
- One way is to maximize the sum of pair-wise correlations (SUMCOR).

Generalized Canonical Correlation Analysis (GCCA)

- What there are more than two data matrices, i.e. X_k for $1 \leq k \leq N$
- We seek a generalization, equivalent to CCA in the two-sample case.
- One way is to maximize the sum of pair-wise correlations (SUMCOR).
- Optimization becomes
  $(a_1, ..., a_k) = argmax(a^T(C - D)a)$
  s.t. $\frac{1}{N}\sum_{k=1}^{N} a_k^T C_{kk} a_k = 1$
  where $C_{ij} = Corr(X_i, X_j)$ and $D_{ii} = Corr(X_i, X_i)$

GCCA and fMRI

- Authors' idea: GCCA operates on inconsistent temporal responses to tasks while still able to maximize the correlations of the latent spatial maps

GCCA and fMRI

- Authors' idea: GCCA operates on inconsistent temporal responses to tasks while still able to maximize the correlations of the latent spatial maps
- Let X_k be some $n \times t_k$ fMRI data matrix

GCCA and fMRI

- Authors' idea: GCCA operates on inconsistent temporal responses to tasks while still able to maximize the correlations of the latent spatial maps
- Let X_k be some $n \times t_k$ fMRI data matrix
- $z_k = X_k a_k$ "individual spatial map"

GCCA and fMRI

- Authors' idea: GCCA operates on inconsistent temporal responses to tasks while still able to maximize the correlations of the latent spatial maps
- Let X_k be some $n \times t_k$ fMRI data matrix
- $z_k = X_k a_k$ "individual spatial map"
- $z = \frac{1}{N} \sum_{k=1}^{N} z_k$ "population spatial map"

Evaluating reproducibility

- NPAIRS (nonparametric prediction, activation, influence, and reproducibility resampling)
- A framework for evaluating the reproducibility and prediction capabilities of preprocessing pipelines

Evaluating reproducibility

- NPAIRS (nonparametric prediction, activation, influence, and reproducibility resampling)
- A framework for evaluating the reproducibility and prediction capabilities of preprocessing pipelines
- Algorithm design
  1. Partition the fMRI data into half
  2. Use GCCA to separately extract population spatial maps for each half
  3. Compare the two maps to calculate correlation and signal to noise ratio (SNR)

NPAIRs Algorithm

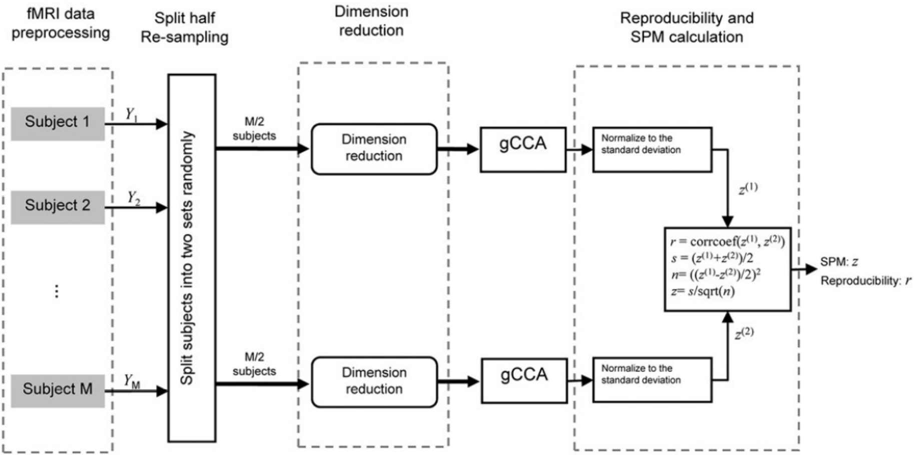

Figure: NPAIRs algorithm for reproducibility and inference

Results

- Comparisons of GCCA to GLM and CVA (canonical variate analysis)
- GCCA finds better spatial map
  - Seems to find the Default Mode Network (DMN)
  - Can't reproduce the BOLD signal
  - Not necessarily useful if attempting to extract task-specific network
- Suggest addition of penalty term to tune spatial/temporal reproducibility

References (links)

- Enhancing reproducibility of fMRI statistical maps using generalized canonical correlation analysis in NPAIRS framework
- Statistical Parametric Maps
- Multiview learning survey