Testing in unsupervised learning (ICA, really), with applications to brain imaging

Aapo Hyvärinen

Dept of Computer Science & Dept of Mathematics and Statistics
University of Helsinki, Finland

with
Pavan Ramkumar (Aalto University, Finland)

Abstract

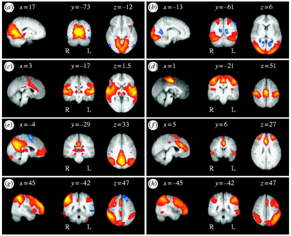
- Theory of independent component analysis (ICA) almost exclusively about estimation
- Here, we propose fundamental testing methods
- Which independent components are reliable/significant?
- Test can be about the mixing matrix or the component values
- We propose a null hypothesis based on the idea of intersubject consistency

Independent component analysis

- ICA is widely used for analyzing neuroimaging data
- One of the main methods in resting-state analysis
- Finds easily resting-state networks in fMRI (Beckmann et al 2005, Van de Ven 2005), recently similar results in MEG (Hyvärinen et al 2010, Brookes et al 2011)
- Decomposes data matrix **X** into a mixing matrix **A** and component matrix **S**

$$\mathbf{X} = \mathbf{A}\mathbf{S}$$

- Maximizing independence or non-gaussianity of the rows of **S**.



(Beckmann et al, 2005)

Importance of testing independent components

- How do we know that an estimated component is not just a random effect?
- ICA algorithms give a fixed number of components and do not tell which ones are reliable (statistically significant)
- Algorithmic artifacts also possible (local minima)
- In general, any estimation method should be complemented by a testing method
- Previously, testing zeros in the mixing matrix was prosed by Shimizu et al. (2006), but often zeros are not priviledged.

Testing using intersubject or intersession consistency

- We propose the following approach:
  - We assume we have a number of similar datasets available
  - Do ICA separately on each of them
  - A component is significant if it appears in two or more datasets in a sufficiently similar form
- Different datasets can come from different subjects, or sessions.
- Similarity could be about components in **S** or columns of mixing matrix **A**
- Key question: How to quantify the case of complete randomness, i.e. null hypothesis

Testing the mixing matrix: Null hypothesis

- ICA is a rotation of whitened data $\mathbf{X}$: after whitening, we have

$$\tilde{\mathbf{X}} = \mathbf{US} \tag{1}$$

- Assume all the subjects/sessions can be whitened using the same matrix.
- Under null hypothesis, spatial patterns of different subjects are "*completely random*" rotations in the PCA subspace (uniformly distributed in the set of orthogonal matrices).
- This models both the actual randomness in the data (differences in brain anatomy) and errors in ICA estimation.

Definition and significance of similarities

- Consider columns of the mixing matrix $\mathbf{p}_{jl}$ of components j and dataset l.
- Compute similarities of spatial patterns using Mahalanobis metric

$$\gamma_{ij,kl} = \frac{|\mathbf{p}_{ik}^T \mathbf{M} \mathbf{p}_{jl}|}{\sqrt{\mathbf{p}_{ik}^T \mathbf{M} \mathbf{p}_{ik}} \sqrt{\mathbf{p}_{jl}^T \mathbf{M} \mathbf{p}_{jl}}} \tag{2}$$

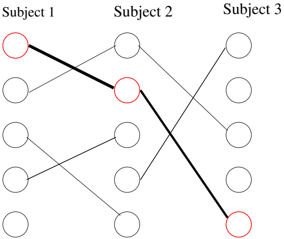
  with $\mathbf{M}$ is (stabilized) inverse of covariance matrix of $\mathbf{p}$
- Under null hypothesis, marginal distribution of γ can be obtained in closed form: e.g.

$$t = \frac{\gamma\sqrt{d-1}}{\sqrt{1-\gamma^2}} \tag{3}$$

  follows a Student's t-distribution with $d-1$ DOF.

Clustering

- Once significances have been computed, use them in clustering
- Prune connections (similarities) which are not significant
- No more than one component per subject
- Similar to hierarchical clustering $\Rightarrow$ Single-linkage vs. complete-linkage strategies

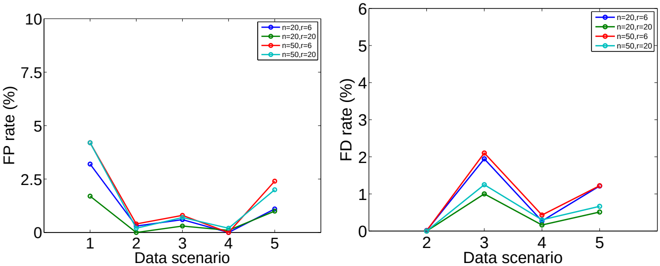


Corrections for multiple testing

- We are testing over many connections, so false positive rates have to be corrected
- We use two different corrections
- For initial creation of cluster: Bonferroni correction
  - Probability of having any false positive clusters $< \alpha$.
  - We don't want to have any false positive clusters
- For adding more components to cluster: false discovery rate
  - Percentage of false positive components $< \alpha$.
  - A few false positive components is not too serious, and we don't want to be too conservative.
  - Can be computed by Simes' procedure, or using a simple formula:

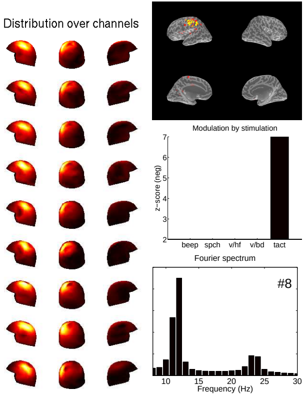
$$\alpha_{corr} = \frac{\alpha}{\text{number of subjects}} \tag{4}$$

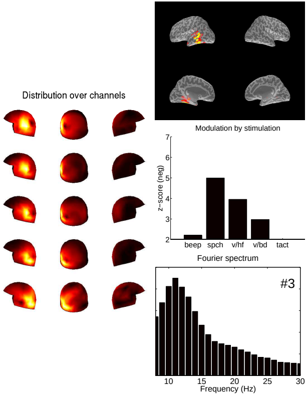
Simulations on artificial data



False positive rates and false discovery rates for simulated data. The desired rates were set to 5%.

Testing ICs: results

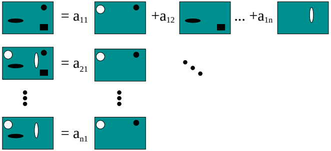




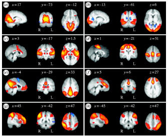
11 subjects, PCA dimension 64, $\alpha = 0.05$, 43 clusters found

Testing independent components themselves

- Spatial ICA: scans at different time points are linear sums of “source images”



- Almost always used in fMRI



- Can also be useful with MEG (Ramkumar et al, 2011)
- The independent components (rows of $\mathbf{S}$) are similar over datasets in $\mathbf{X} = \mathbf{AS}$

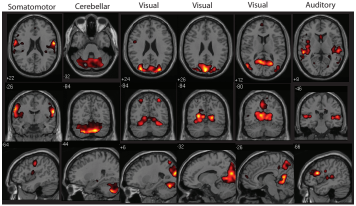
Empirical approach to null distribution

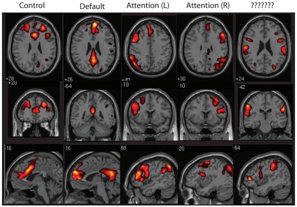
- Same basic approach:
  - ICA separately on multiple datasets k, l (subjects/sessions)
  - Compute similarities $\gamma = \mathbf{S}_k \mathbf{S}_l^T$
  - Null hypothesis: random orthogonal rotation
- But: Independent components contain a lot of noise
  $\Rightarrow$ Similarities necessarily small
- Null distribution modelled empirically
- For random orthogonal matrices, we have

$$\gamma^2 \sim \mathrm{Beta}(1/2, \beta) \tag{5}$$

  where β equals the dimension of the space.
- Here, we *estimate* β by fitting to the empirical distribution

Resting- state networks on fMRI data (preliminary)

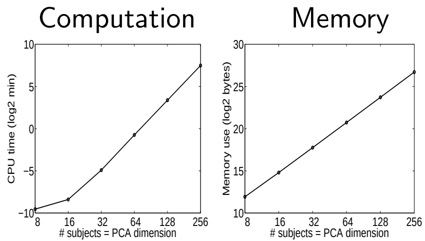




- 11 subjects
- PCA dimension 75
- $\alpha = 0.001$
- 56 clusters found

Computational complexity

- Main difficulty is memory: We need to store similarity matrix
- This can be reduced by storing just the strongest similarity from each components
- We can handle 100-200 subjects with 100-200 components on a desktop computer


Computation

Memory

Special bonus slide for Algodan

- The method could be applied in general unsupervised learning
- Assume the features live in a set of finite volume (compact), e.g. the unit sphere
- Then we can define the null hypothesis
- Consider e.g. clustering, where data is normalized to unit sphere
- Any data set can be divided into n subsets and learning can be performed for each data set
- Maybe you can apply this testing for your own method?

Conclusion

- We introduces methods for testing of which independent components are reliable, i.e. statistically significant
- We can test columns of the mixing matrix, or the values of the independent components themselves
- Based on doing ICA separately on many datasets, i.e. different subjects or sessions
- Null hypothesis defined as orthogonal rotations in whitened space
- Null distribution obtained analytically for mixing matrix case
  Empirical approximation needed for ICs
- Application on MEG and fMRI promising