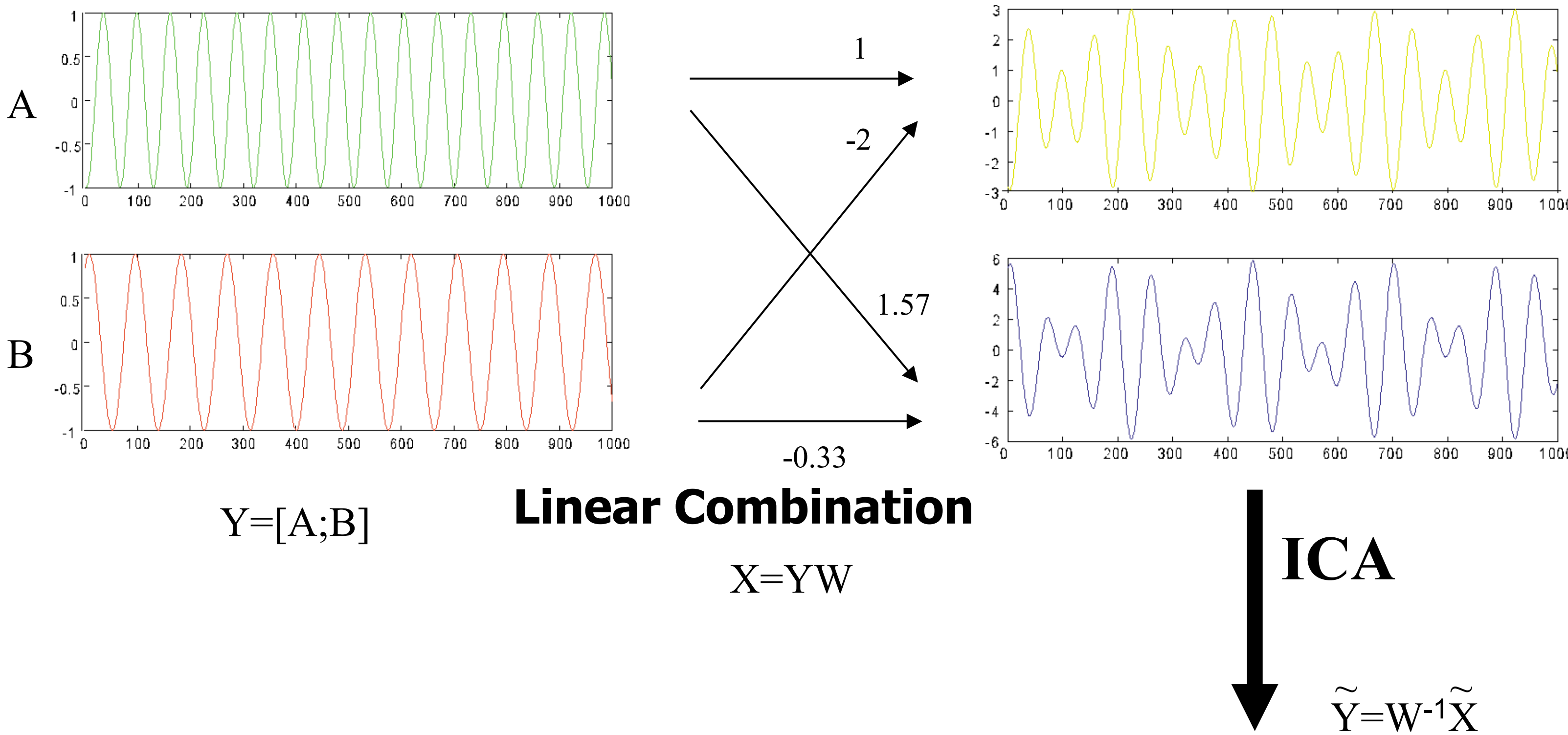
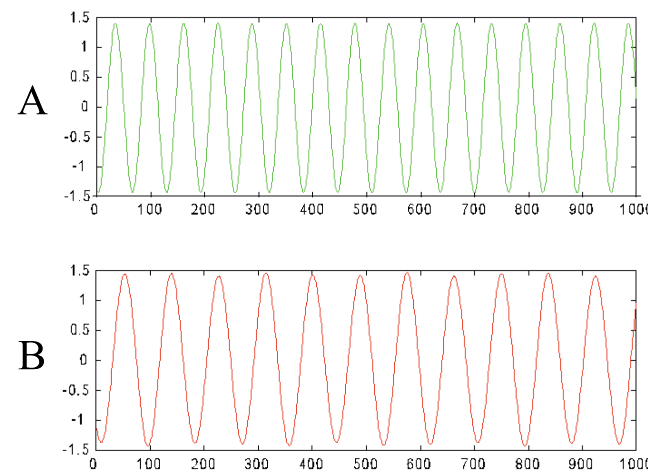
Clustering of ICA components

Arnaud Delorme

(with Julie Onton, Romain Grandchamp, Nima Bigdely Shamlo, Scott Makeig)





ICA is a method to recover a version, of the original sources by multiplying the data by a unmixing matrix

$$\mathbf{U} = \mathbf{W}\mathbf{X}$$

ICA activity → $\mathbf{U}$; Data → $\mathbf{X}$

Data X

$$\begin{bmatrix} 3 & 2 & 5 & 4 & 3 & 2 & \\ 0 & -2 & -5 & -1 & 1 & -1 & \dots \\ -1 & 2 & 0 & 1 & 0 & -3 & \end{bmatrix} \begin{matrix} \leftarrow \text{Channel 1} \\ \leftarrow \text{Channel 2} \\ \leftarrow \text{Channel 3} \end{matrix}$$

$$\begin{bmatrix} 5 & 3 & -2 \\ 1 & 2 & 4 \\ 0 & -1 & 3 \end{bmatrix} \rightarrow \begin{bmatrix} 3*5+0*3-1*(-2) & 2*5+(-2)*3+2*(-2) & \\ 3*1+0*2-1*4 & 2*1+(-2)*2+2*4 & \dots \\ 5*1-5*2+0*4 & 5*1-5*2+0*4 & \end{bmatrix} \begin{matrix} \leftarrow \text{Comp. 1} \\ \leftarrow \text{Comp. 2} \\ \leftarrow \text{Comp. 3} \end{matrix}$$

Weight matrix W

ICA activity U

$$X = W^{-1} U$$

Data

ICA activity U

$$\begin{bmatrix} 3 & 2 & 5 & 4 & 3 & 2 & \\ 0 & -2 & -5 & -1 & 1 & -1 & \dots \\ -1 & 2 & 0 & 1 & 0 & -3 & \end{bmatrix} \begin{matrix} \leftarrow \text{Comp. 1} \\ \leftarrow \text{Comp. 2} \\ \leftarrow \text{Comp. 3} \end{matrix}$$

* * *

$$\begin{bmatrix} 5 & 3 & -2 \\ 1 & 2 & 4 \\ 0 & -1 & 3 \end{bmatrix} \rightarrow \begin{bmatrix} 3*5+0*3-1*(-2) & 2*5+(-2)*3+2*(-2) & \\ 3*1+0*2-1*4 & 2*1+(-2)*2+2*4 & \dots \\ 5*1-5*2+0*4 & 5*1-5*2+0*4 & \end{bmatrix} \begin{matrix} \leftarrow \text{Chan 1} \\ \leftarrow \text{Chan 2} \\ \leftarrow \text{Chan 3} \end{matrix}$$

Inverse weight matrix W^{-1}

Data X

Historical Remarks

- Herault & Jutten ("Space or time adaptive signal processing by neural network models“, *Neural Nets for Computing Meeting*, Snowbird, Utah, 1986): Seminal paper, neural network
- Bell & Sejnowski (1995): Information Maximization
- Amari et al. (1996): Natural Gradient Learning
- Cardoso (1996): JADE

- **Applications of ICA to biomedical signals**
  - EEG/ERP analysis (Makeig, Bell, Jung & Sejnowski, 1996).
  - fMRI analysis (McKeown et al. 1998)

ICA Theory – Cost Functions

Family of BSS algorithms

- Information theory (Infomax)
- Bayesian probability theory (Maximum likelihood estimation)
- Negentropy maximization
- Nonlinear PCA
- Statistical signal processing (cumulant maximization, JADE)

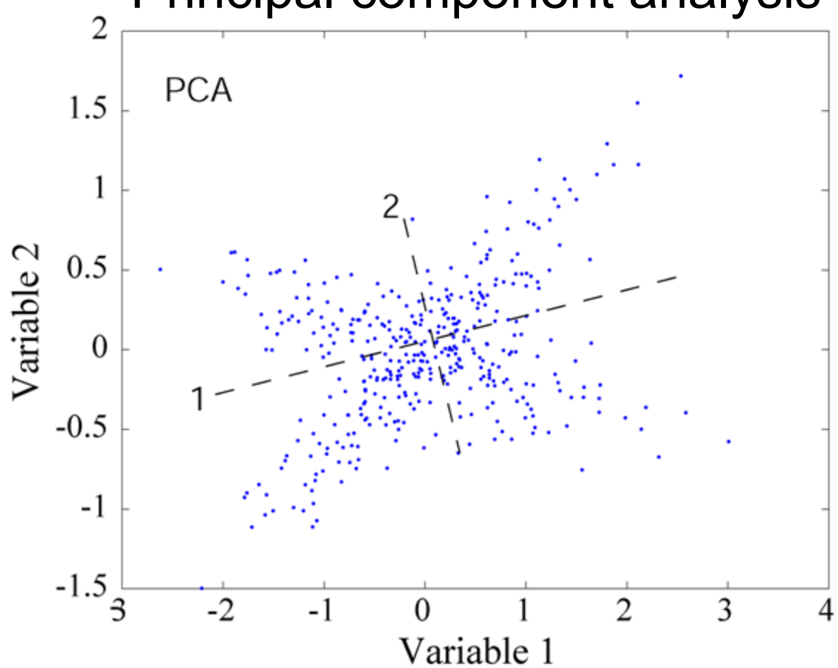
A unifying Information-theoretic framework for ICA

- Pearlmutter & Parra showed that InfoMax, ML estimation are equivalent.
- Lee et al. (1999) showed negentropy has the equivalent property to InfoMax.
- Girolami & Fyfe showed nonlinear PCA can be viewed from

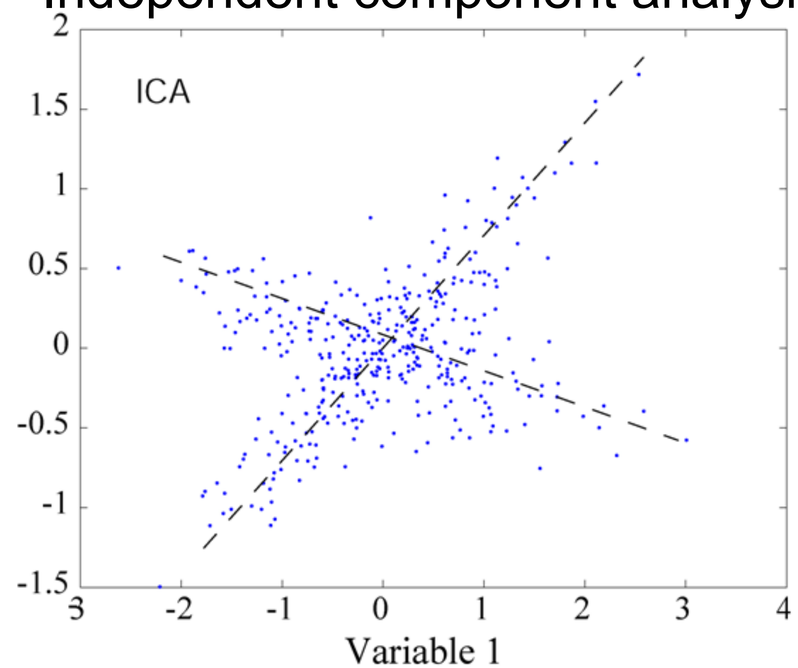
ICA and PCA

ICA is a method to recover a version, of the original sources by multiplying the data by a unmixing matrix,

Principal component analysis



Independent component analysis



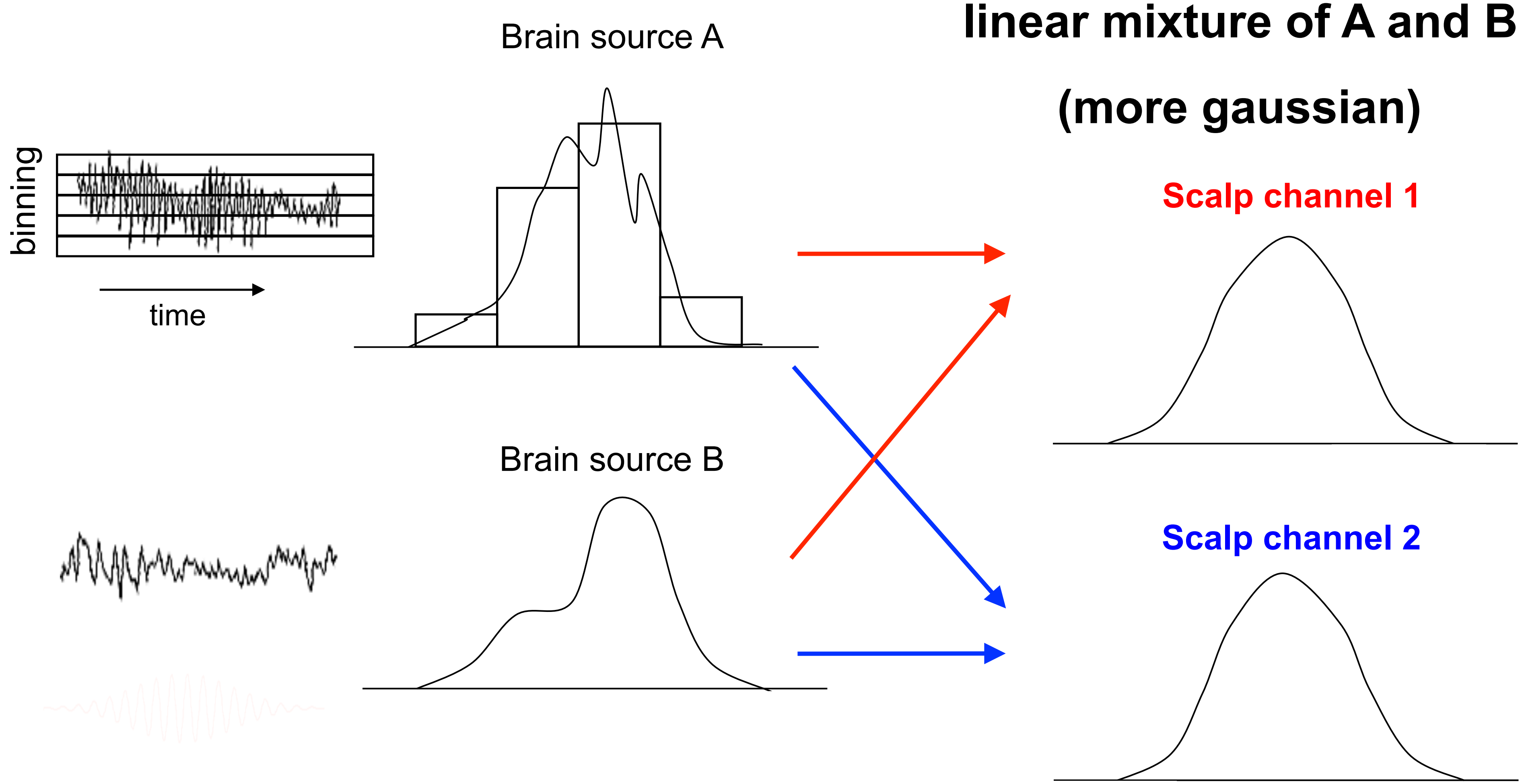
While PCA simply decorrelates the outputs (using an orthogonal mixing matrix), ICA attempts to make the outputs statistically independent, while placing no constraints on the minxing matrix.

Central limit theorem

Scalp channels =

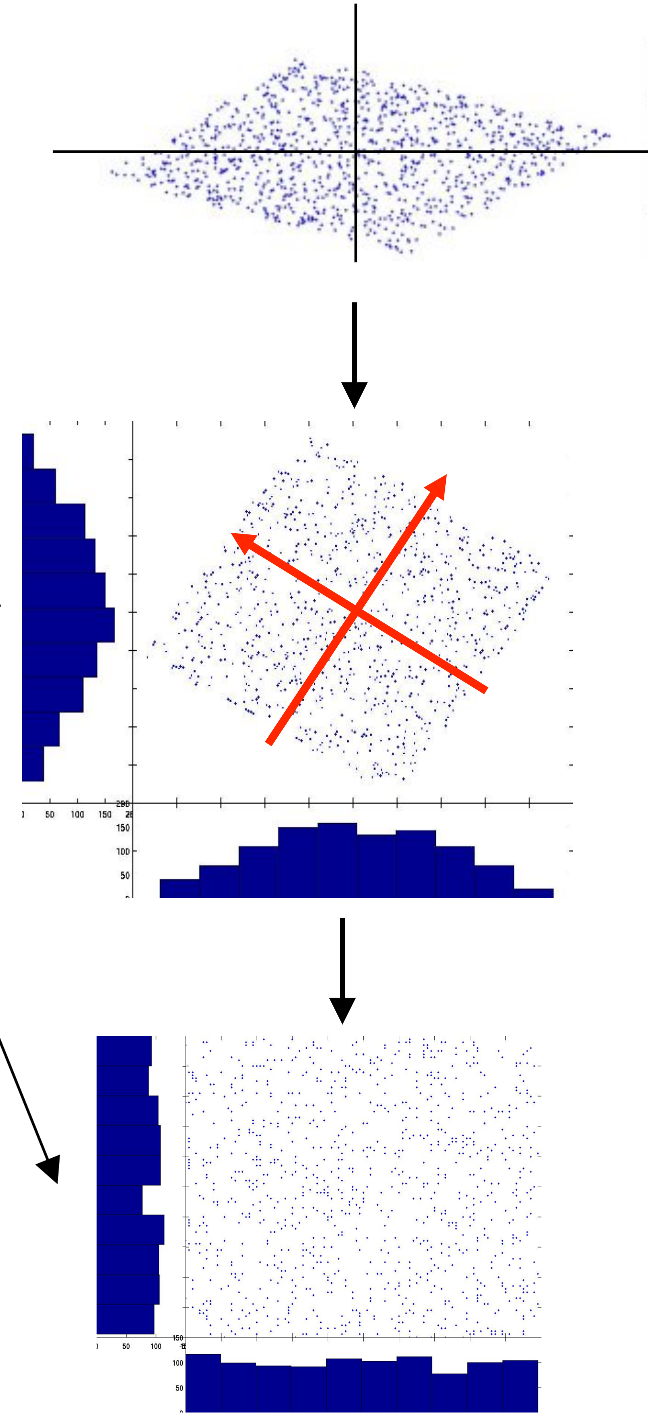
linear mixture of A and B

(more gaussian)



ICA Training Process

Central limit theorem



- Remove the mean
  $x = x - \langle x \rangle$
- 'Sphere' the data by diagonalizing its covariance matrix,

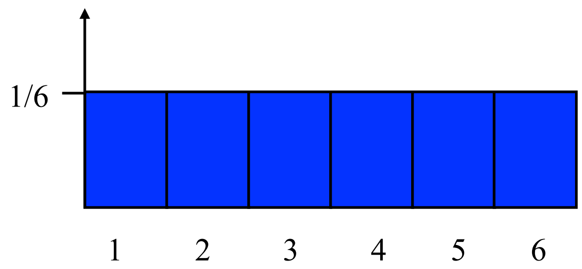
  $x = \langle xx^T \rangle^{-1/2}(x - \langle x \rangle).$
- Update W according to

$$\Delta W \propto \frac{\partial H(\mathbf{y})}{\partial W} W^T W$$

Entropy

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log_b p(x).$$

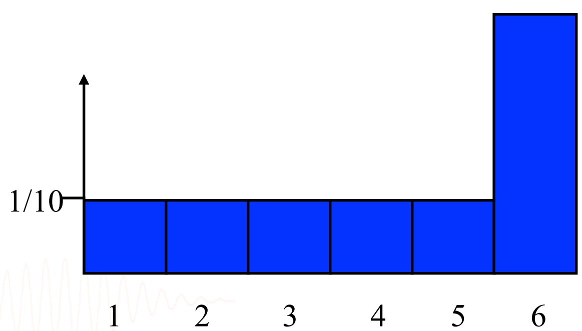
Dice: 1/6



$$H = 6\left(-\frac{1}{6}\log_2\left(\frac{1}{6}\right)\right) = 2.58$$

Fake dice (make a 6 half of the time): entropy 2.16 (base 2)

less random



$$H = 5\left(-\frac{1}{10}\log_2\left(\frac{1}{10}\right)\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right) = 2.16$$

Entropy

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log_b p(x).$$

Joint entropy

$$H(X,Y) = -\sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} p(x,y) \log_b p(x,y);$$

Mutual Information

$$H(y_1, y_2) = H(y_1) + H(y_2) - I(y_1, y_2).$$

Shannon in his landmark 1948 paper ``A Mathematical Theory of Communication."

From http://planetmath.org/encyclopedia/ShannonsTheoremEntropy.html

Contingency table for stress and emotionality

	STRE						
	1	**2**	**3**	**4**	**5**	**6**	**Total**
EMOT= **1**	19	4					23
2	11	63	64	3	1		142
3	2	16	18	20	2	2	60
4	1	4	1	9	6	2	23
5			1	2	4	3	10
6				1	1	1	3
Total	33	87	84	35	13	8	

From http://tecfa.unige.ch/~lemay/thesis/THX-Doctorat/node149.html

Contingency frequencies for stress and emotionality

	STRE					
	1	**2**	**3**	**4**	**5**	**6**
EMOT= **1**	0.07	0.02				
2	0.04	0.24	0.25	0.01		
3	0.01	0.06	0.07	0.08	0.01	0.01
4		0.02		0.03	0.02	0.01
5				0.01	0.02	0.01
6						

Joint entropy 3.46; exercise: compute mutual information

$$H(X,Y) = -\sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} p(x,y)\log_b p(x,y)$$

ICA learning rule

How to make the outputs statistical independent?

Minimize their redundancy or mutual information.

Consider the joint entropy of two components,

$$H(y_1, y_2) = H(y_1) + H(y_2) - I(y_1, y_2).$$

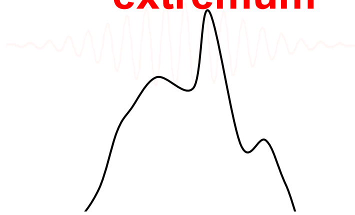
Maximizing $H(y_1, y_2) \Longrightarrow$ minimizing $I(y_1, y_2)$.

The learning rule:

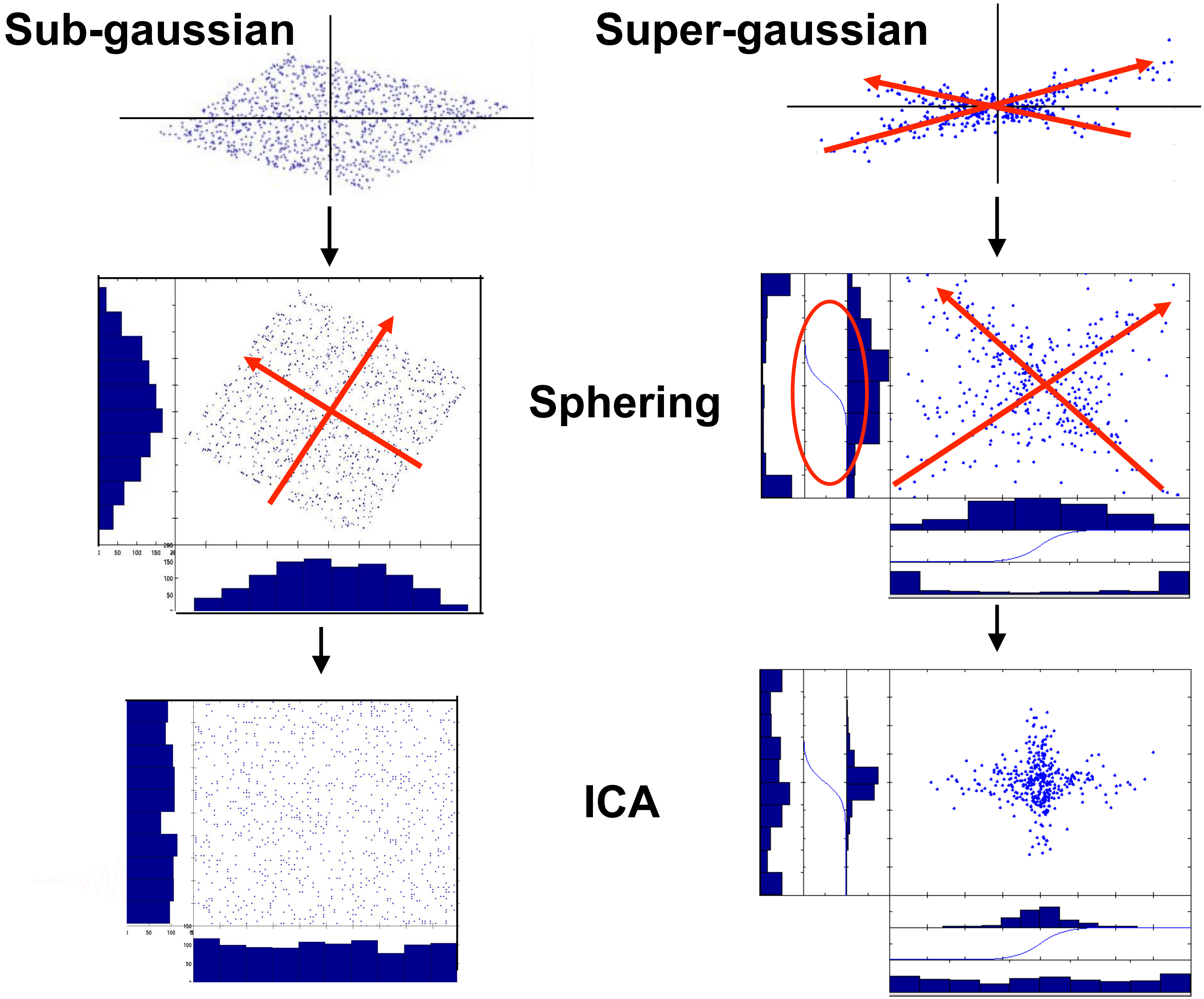
$$\Delta \mathbf{W} \propto \frac{\partial H(\mathbf{y})}{\partial \mathbf{W}} \underbrace{\mathbf{W}^T \mathbf{W}}$$

=0 if the two variables are independent

Entropy extremum

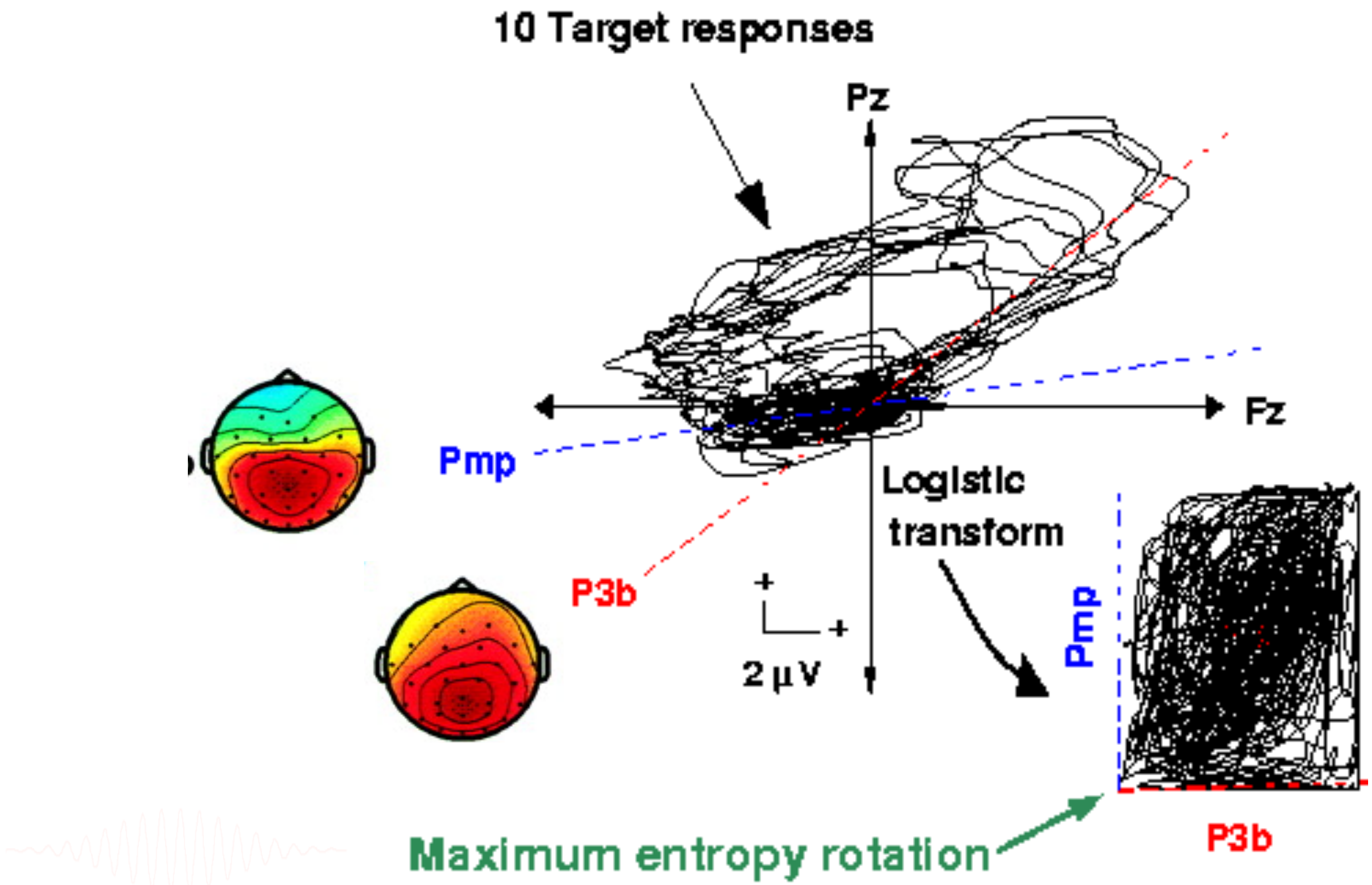


Natural gradient (Amari)



Sub-gaussian
Super-gaussian
Sphering
ICA

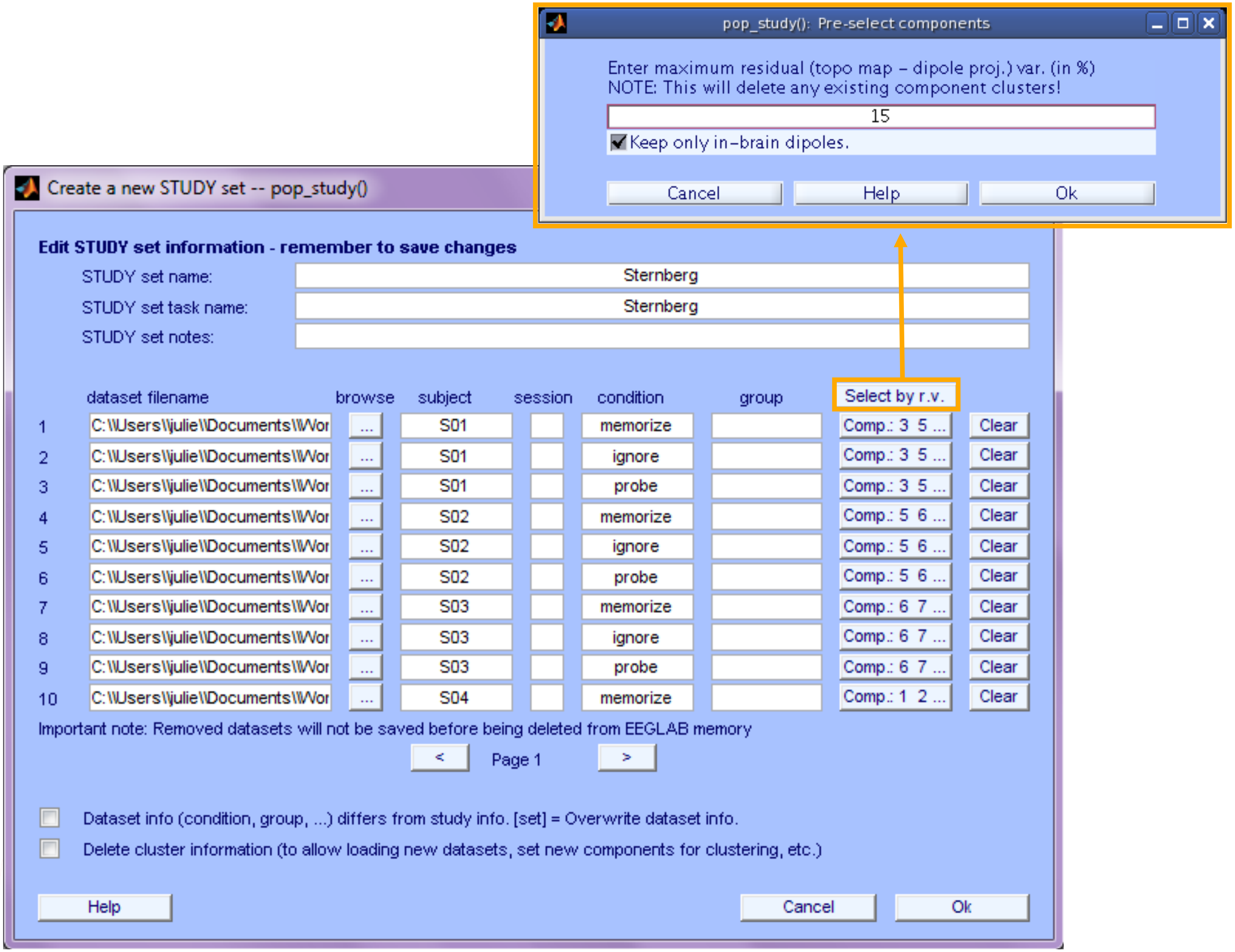
Independent components of EEG/ERP



Steps of clustering

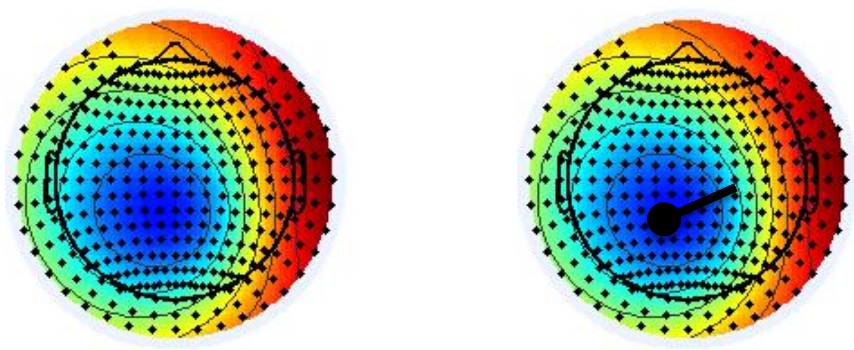
- Select ICA components for clustering
- Precompute measures of interest
- Cluster measures
- Plot clusters and edit them if necessary

Edit dataset info



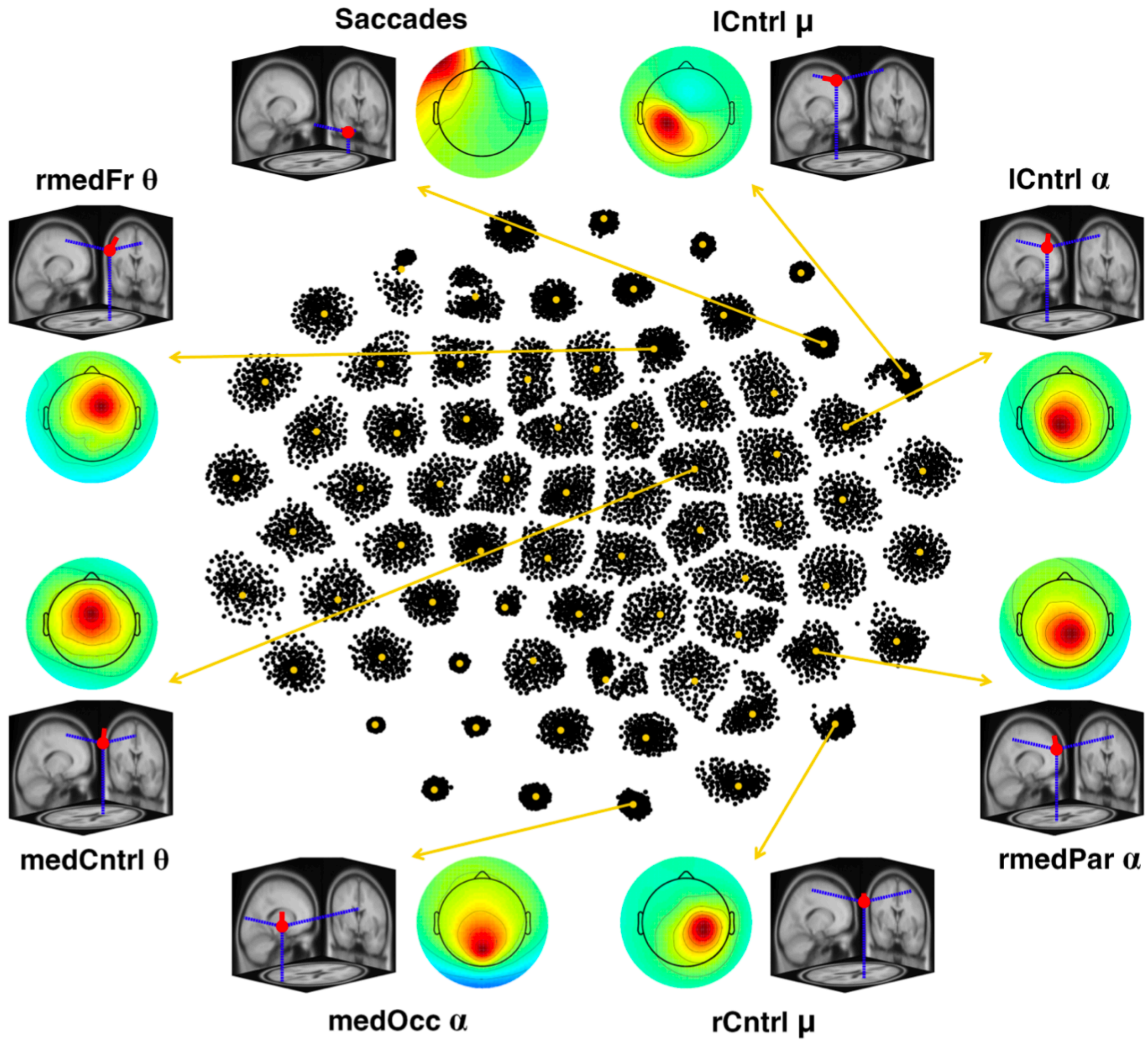
Computing residual variance (%)

Actual Dipole projection



$$r = \Sigma(x_i - \tilde{x}_i)^2 / \Sigma x_i^2$$

Clustering results example

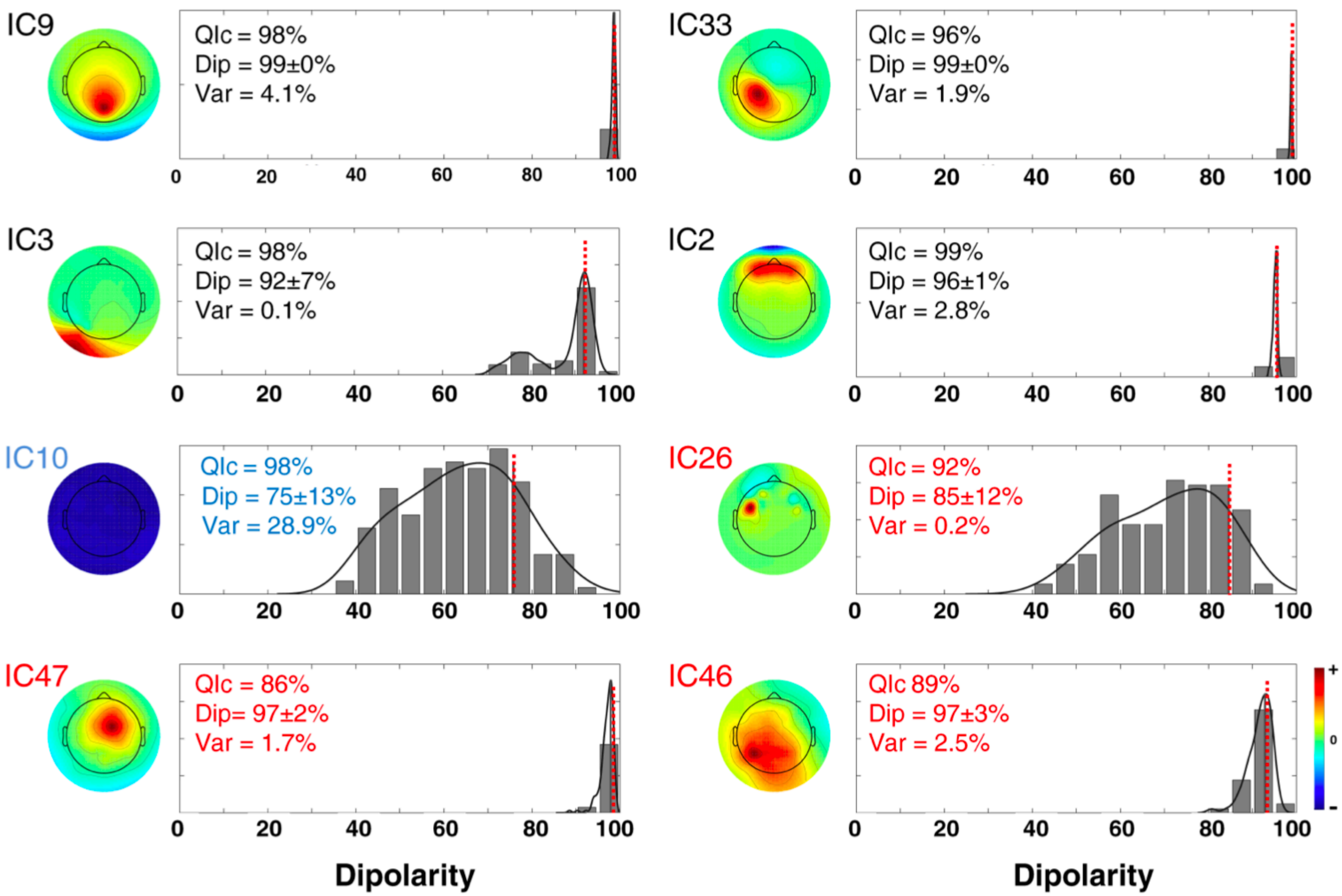


RELICA: A method for estimating the reliability of independent components

Fiorenzo Artoni [a,*], Danilo Menicucci [b], Arnaud Delorme [c,e,f], Scott Makeig [c], Silvestro Micera [a,d]

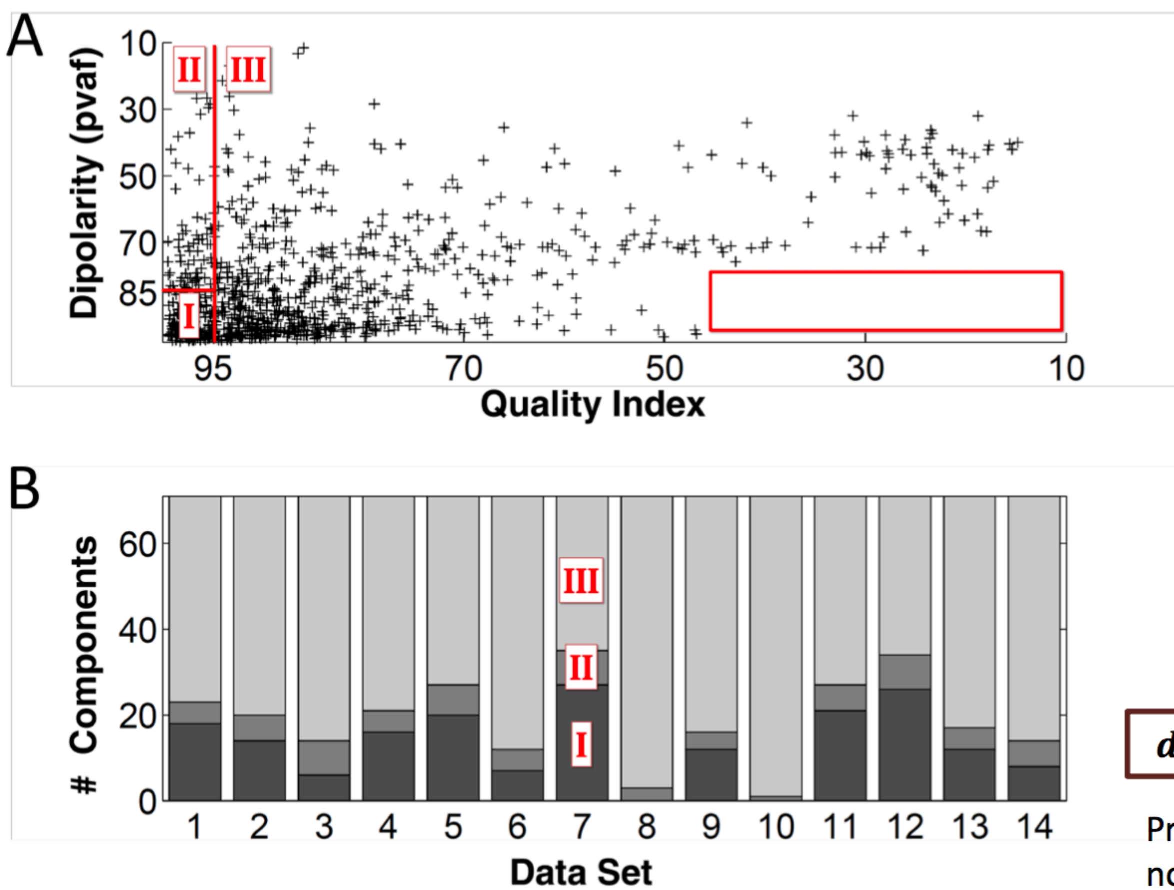
Within-cluster reliability

The distribution of dipolarity within the cluster helps assessing the **quality** and characteristics of Independent Components



Reliability criteria and the rv<15%

First justification why we should select an **r.v <15%** for components to include in further analyses: there is a forbidden region underlined in red, that indicates the absence of



CLASS I

Quality Index and Dipolarity above Retention threshold: **Good**

CLASS II

Quality Index above threshold, dipolarity below: **artifact** or mixing of multiple processes

CLASS III

Quality Index below retention threshold

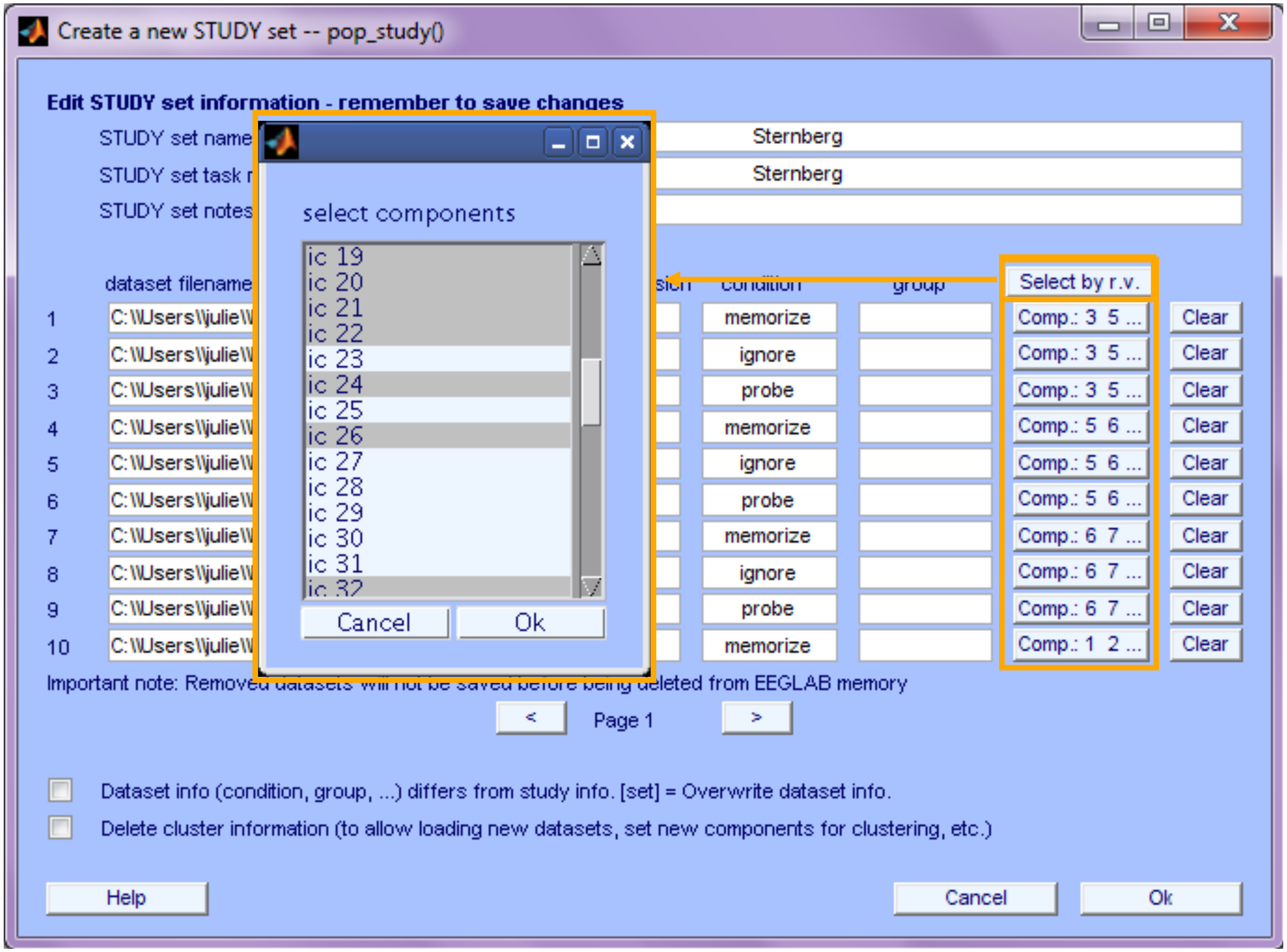
$dip \pm std > th$

Probable Inseparable noise: variance explained useful or **multiple subject confirmation**

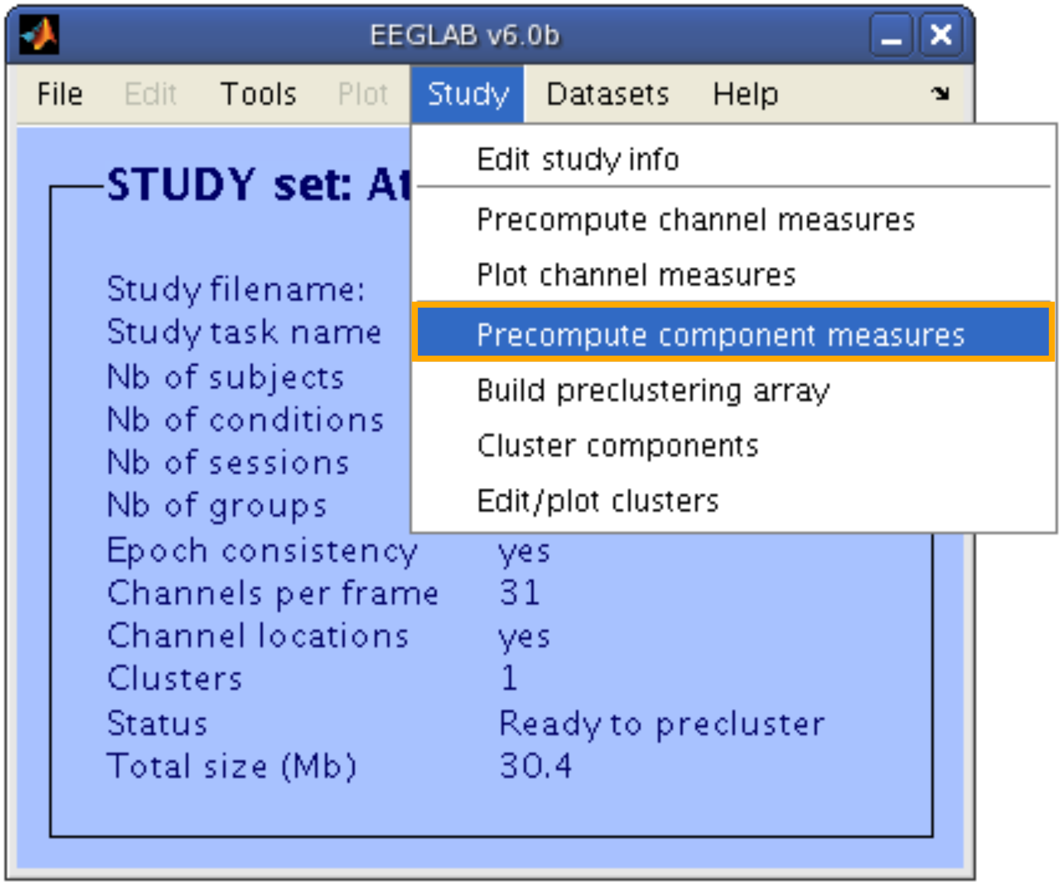
$dip \pm std < th$

Discard

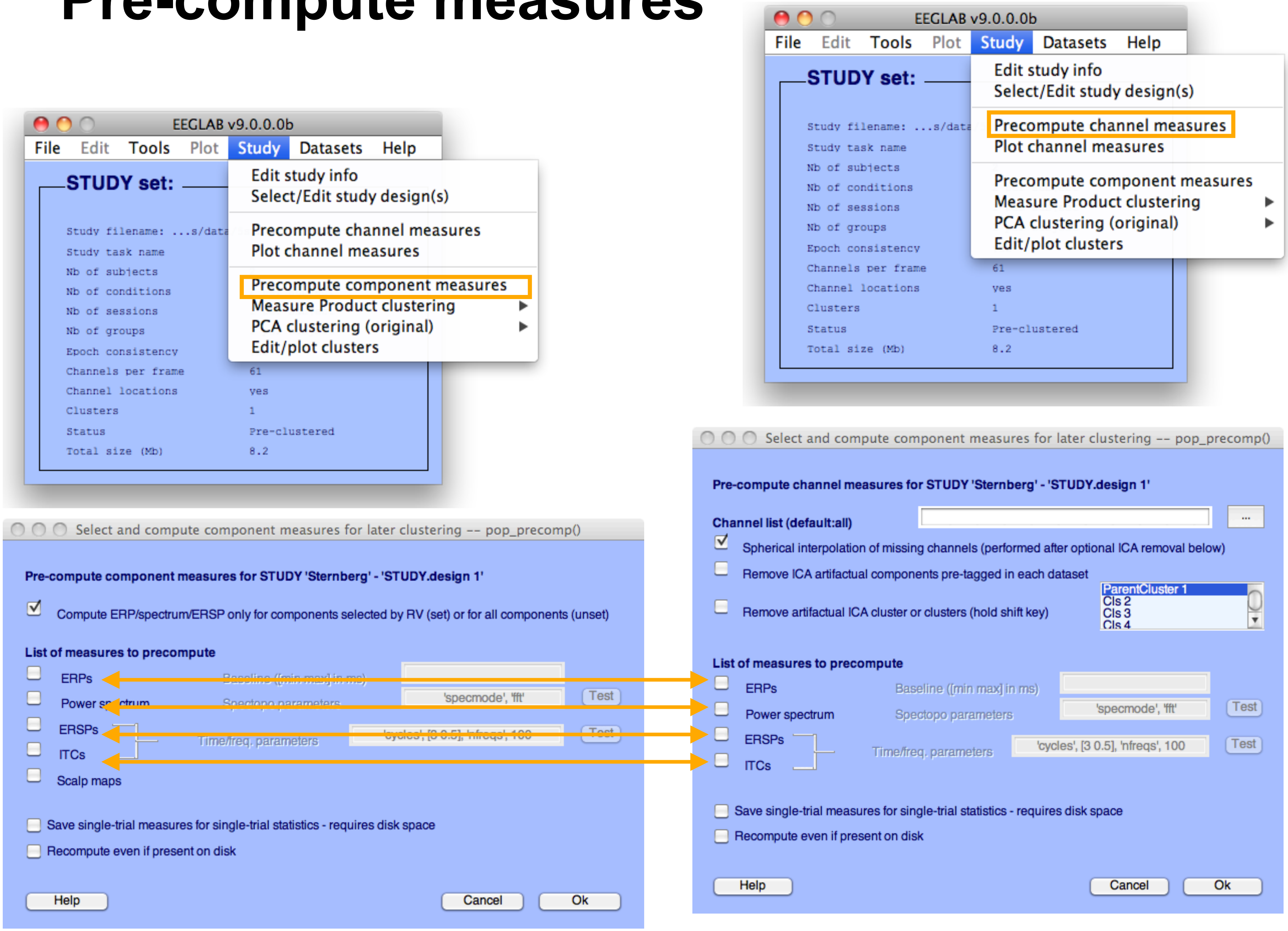
ICs to cluster



Precompute data measures

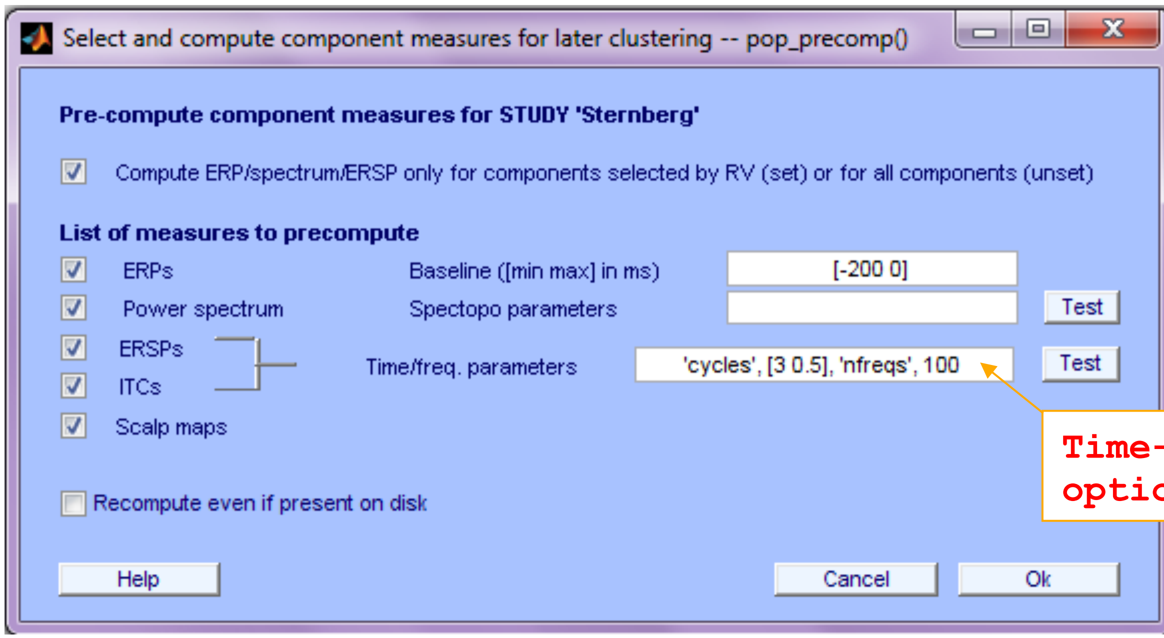


Pre-compute measures



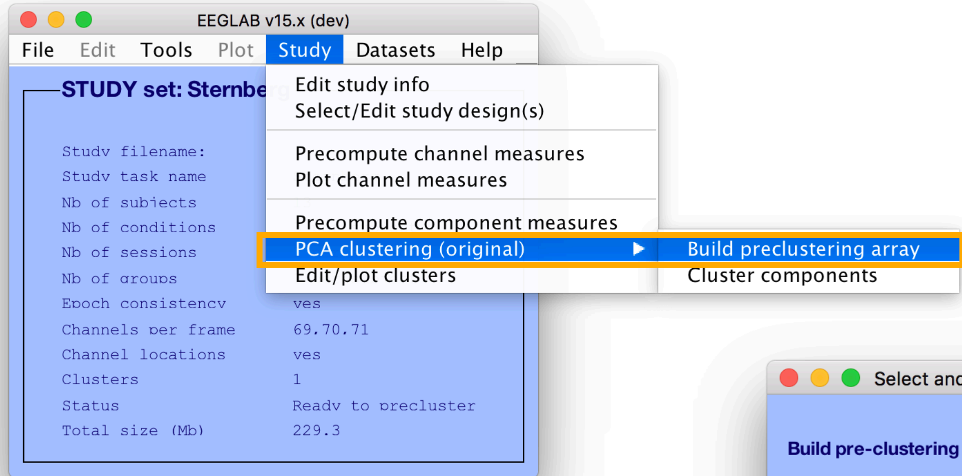
Precompute data measures

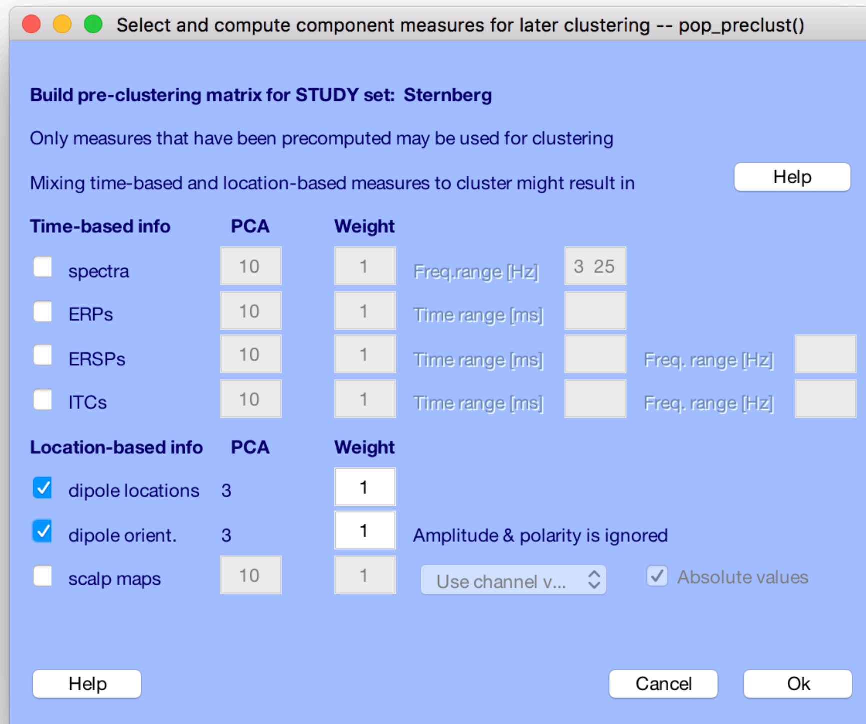
TIP: Compute all measures so you can test different combinations for clustering



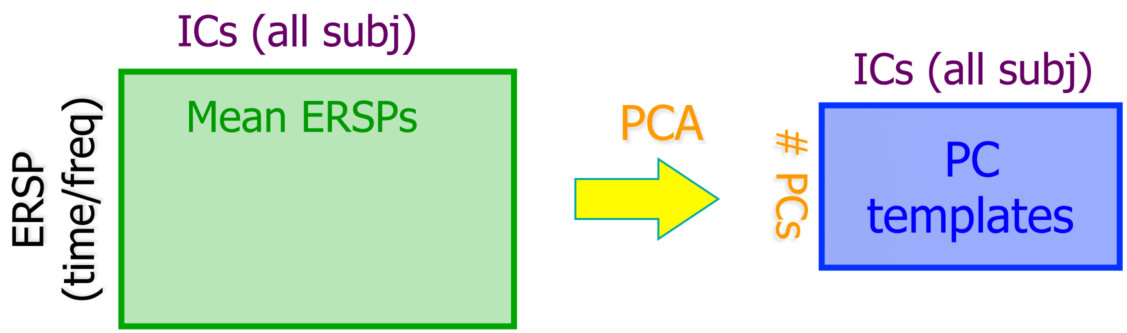
Time-frequency options

Cluster components

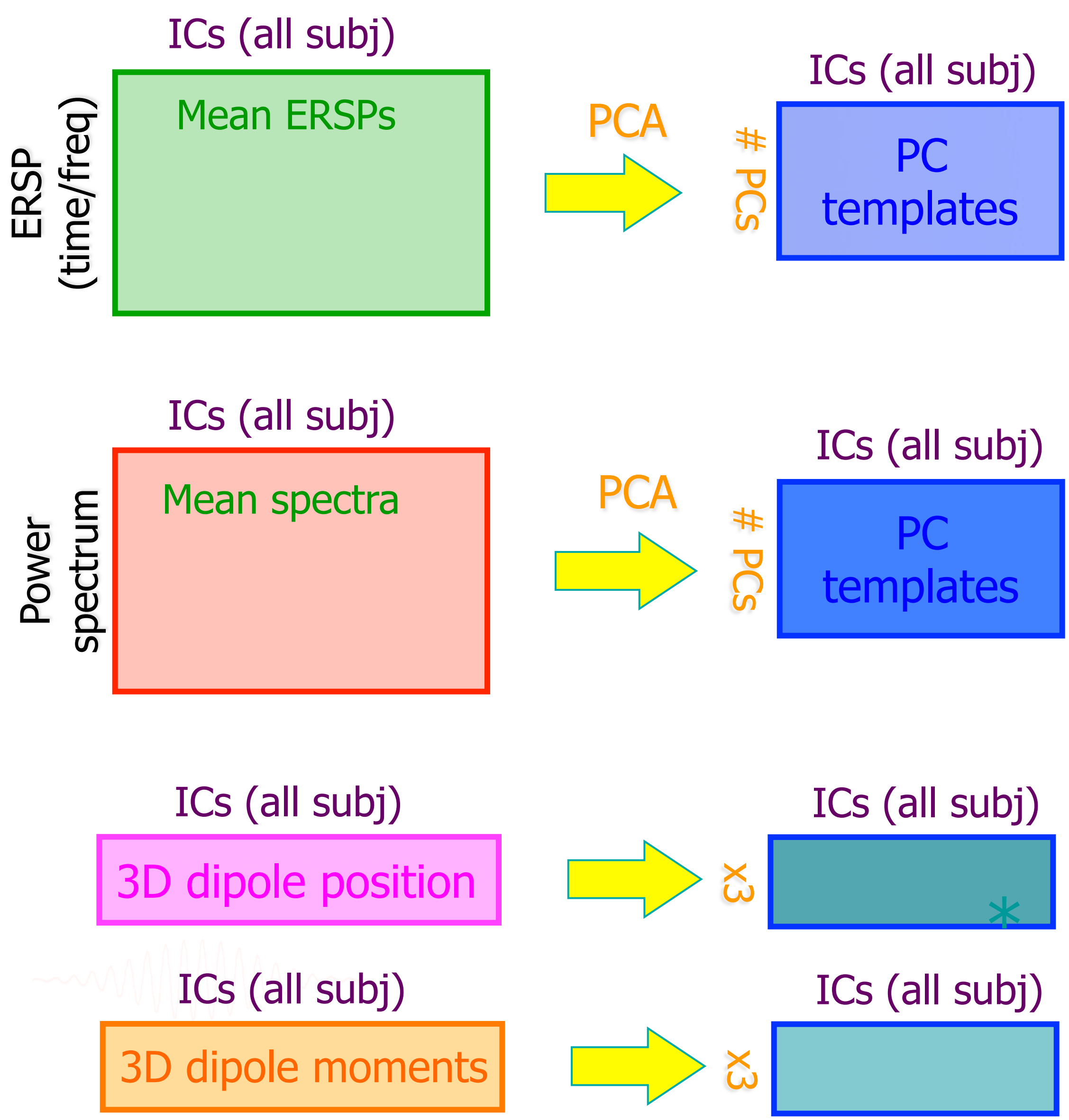




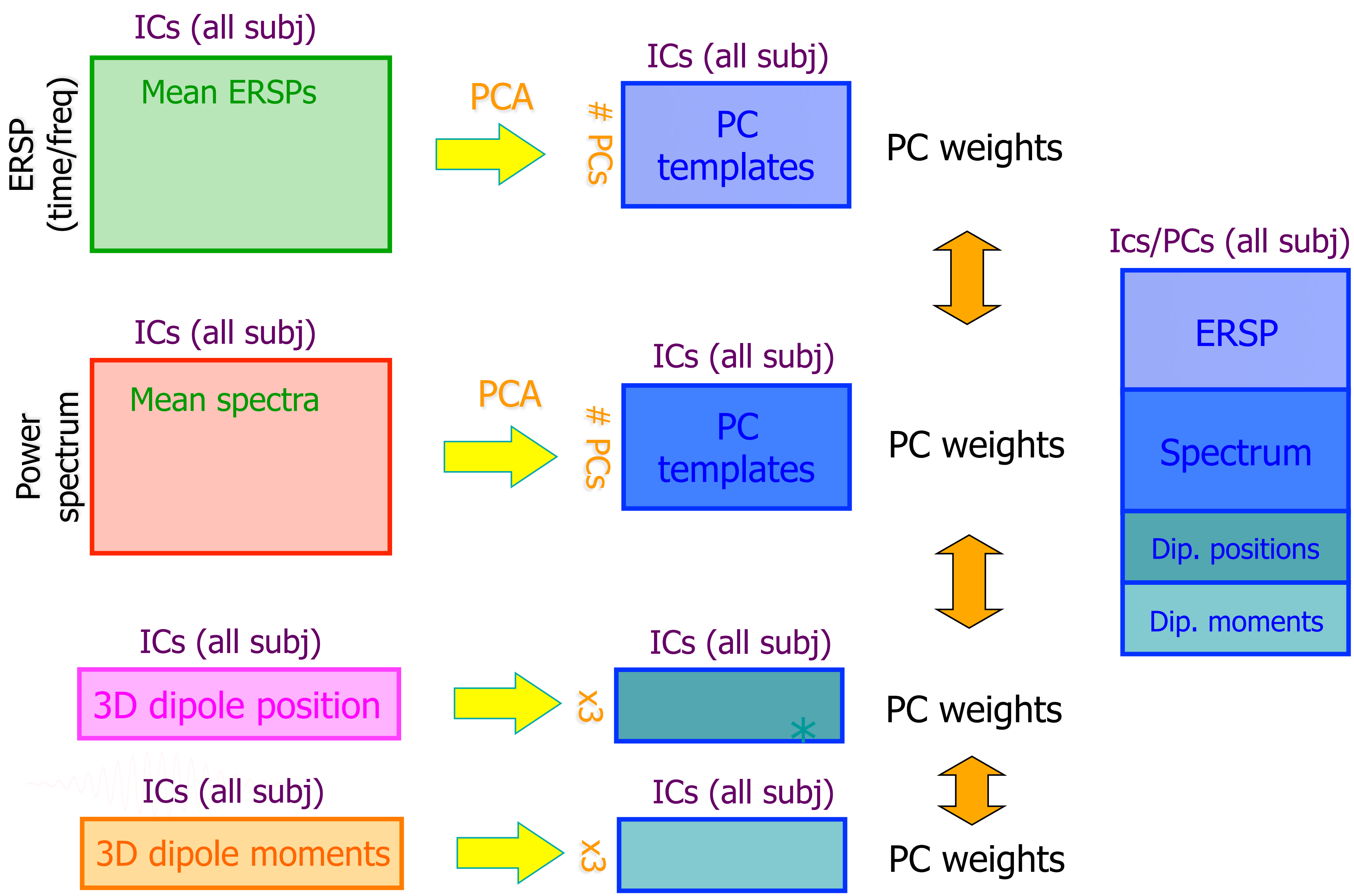
Precluster schematic



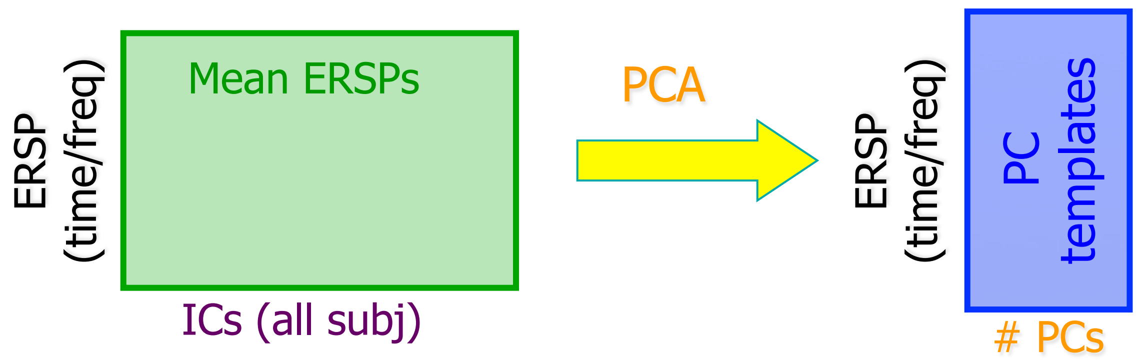
Precluster schematic

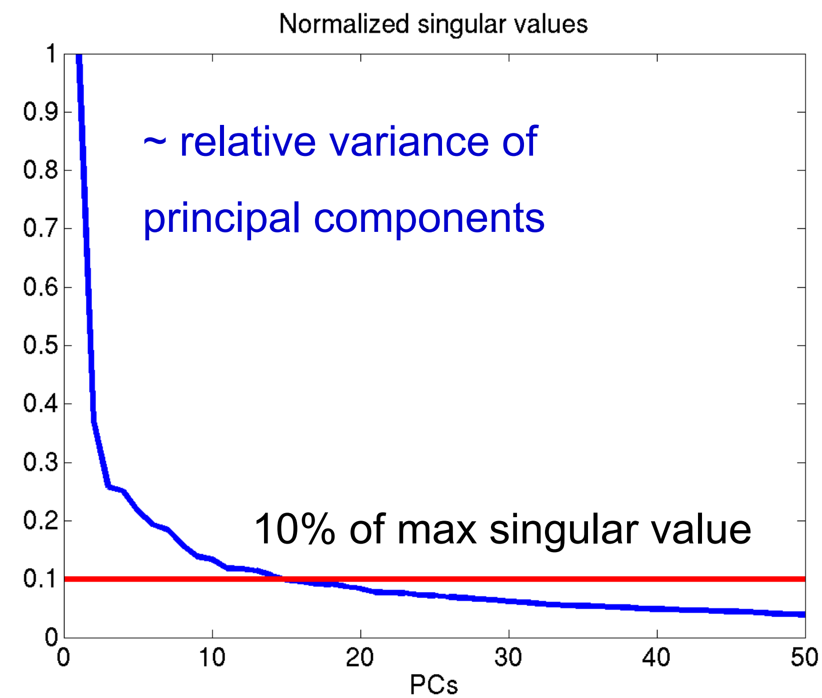


Precluster schematic

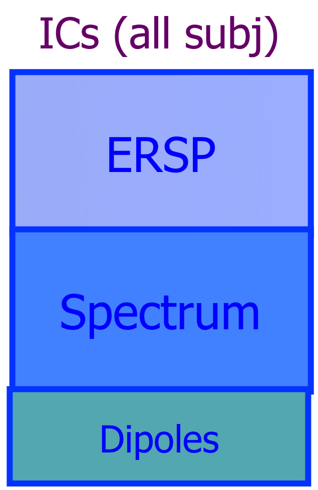


Precluster: Use singular values from PCA

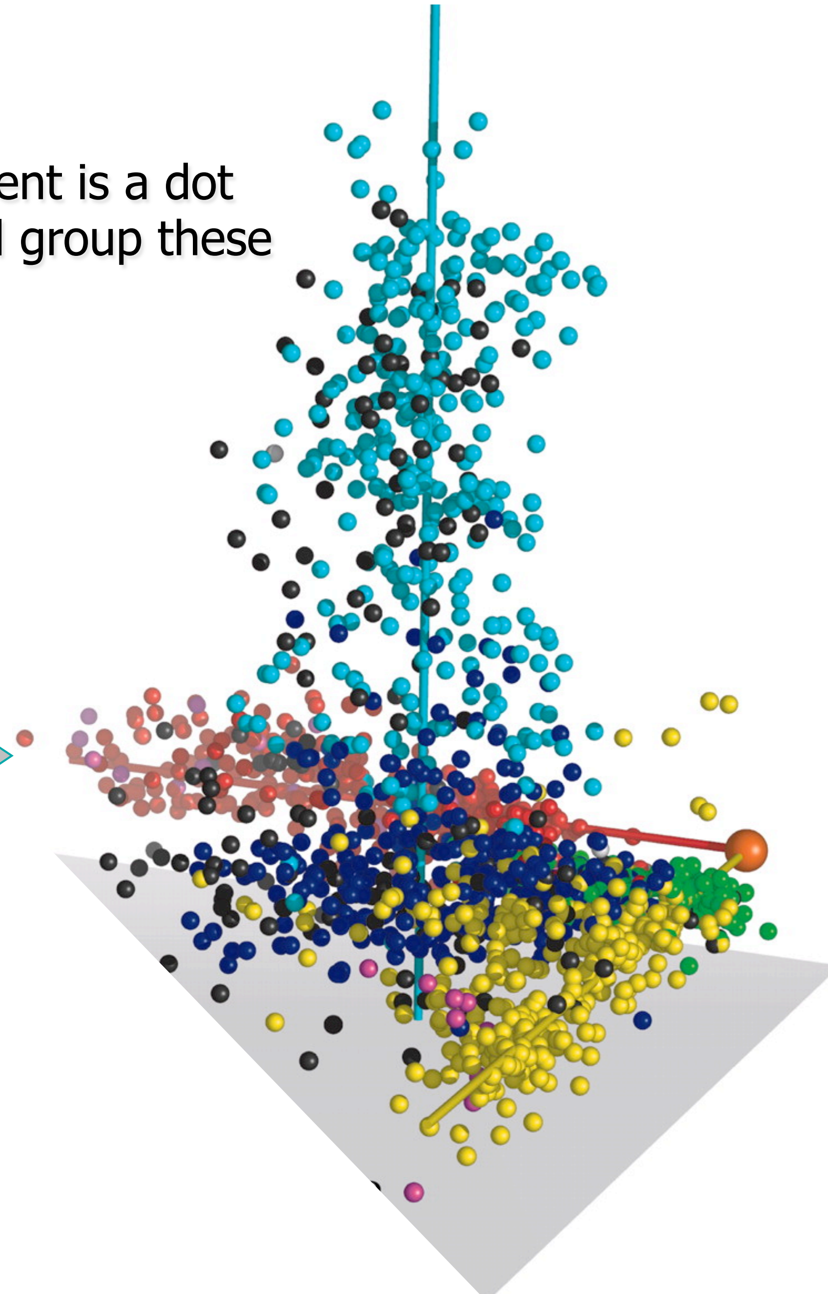




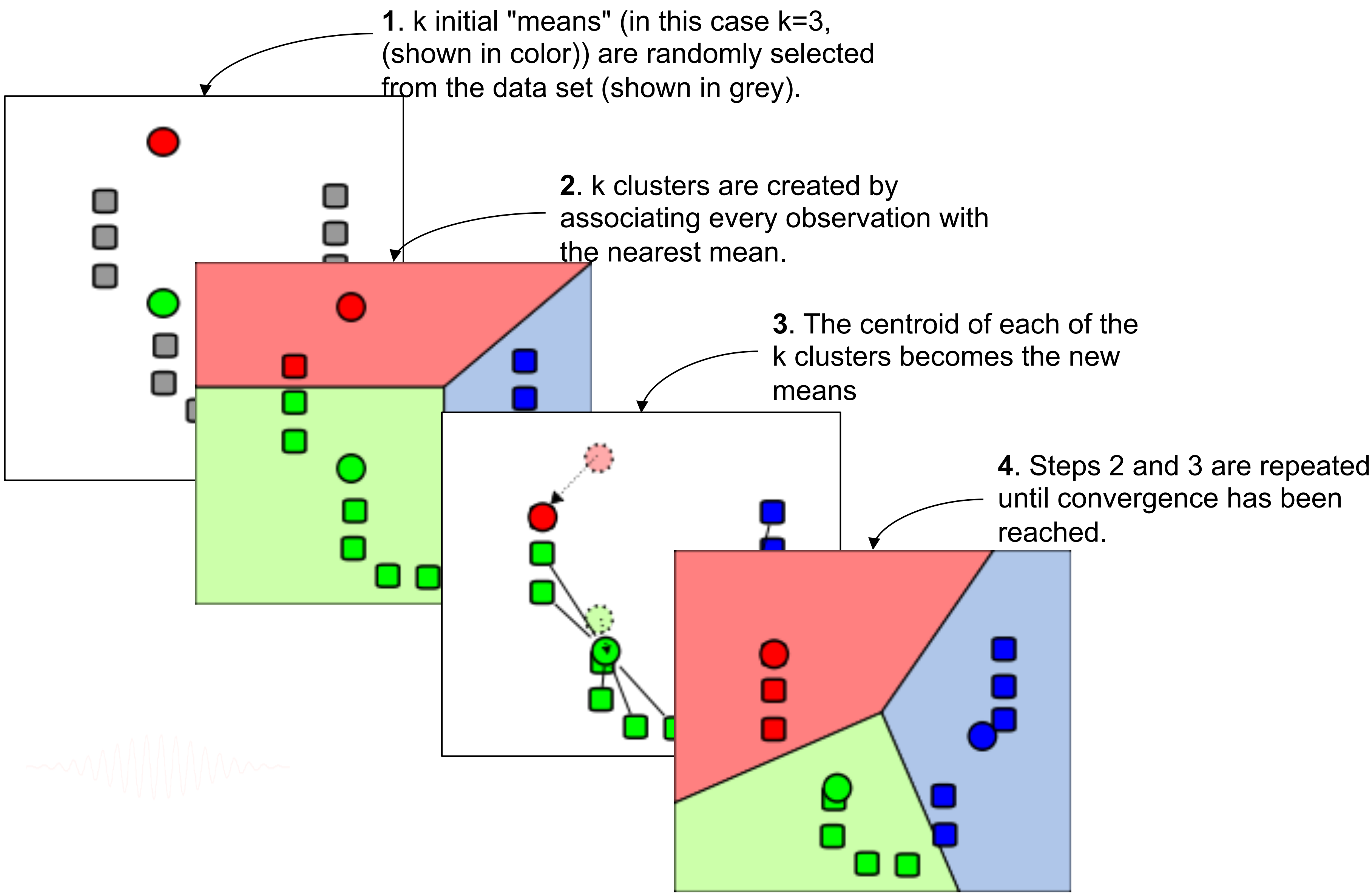
Precluster schematic



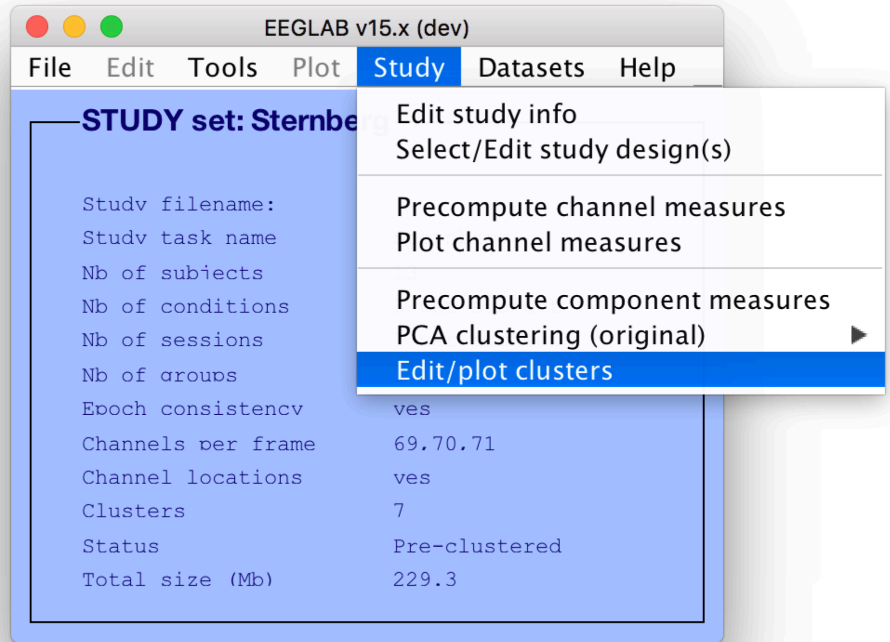
Each component is a dot
Clustering will group these dots

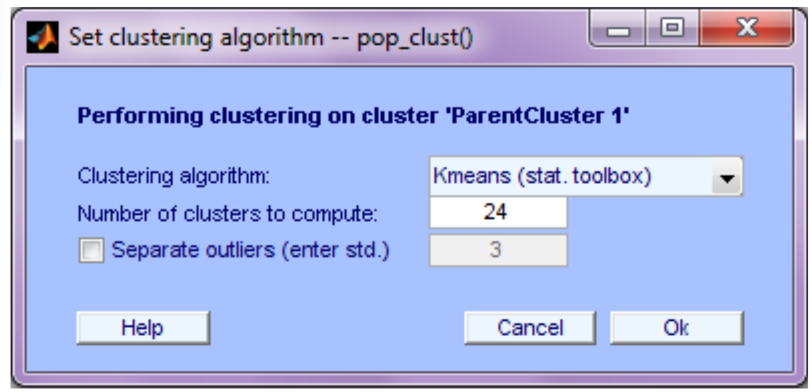


Classical KMean



Cluster components





Choosing data measures

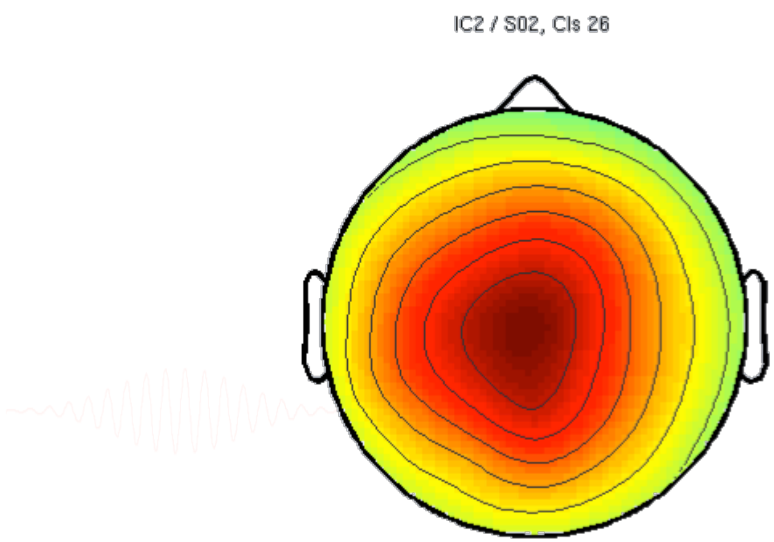
What measure(s) should you use?

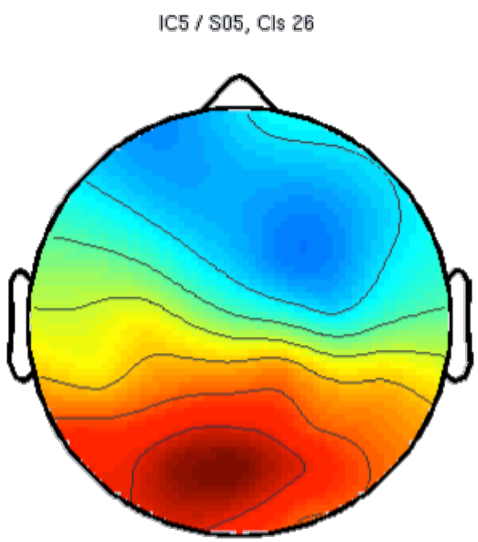
It depends on your final cluster criteria…

- If for example, your priority is dipole location, then cluster only based on dipole location…

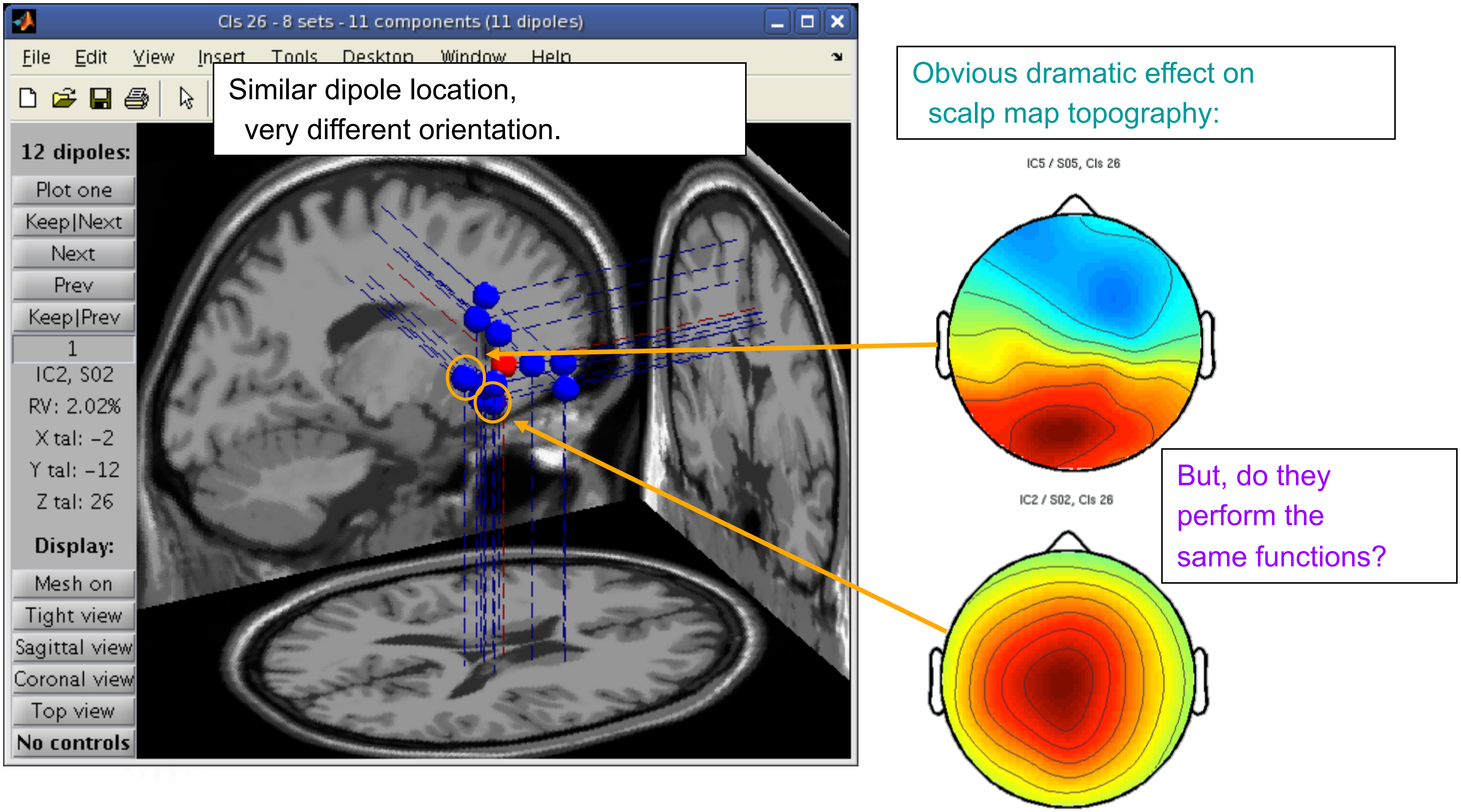
But consider:

- What is the difference between these two components?

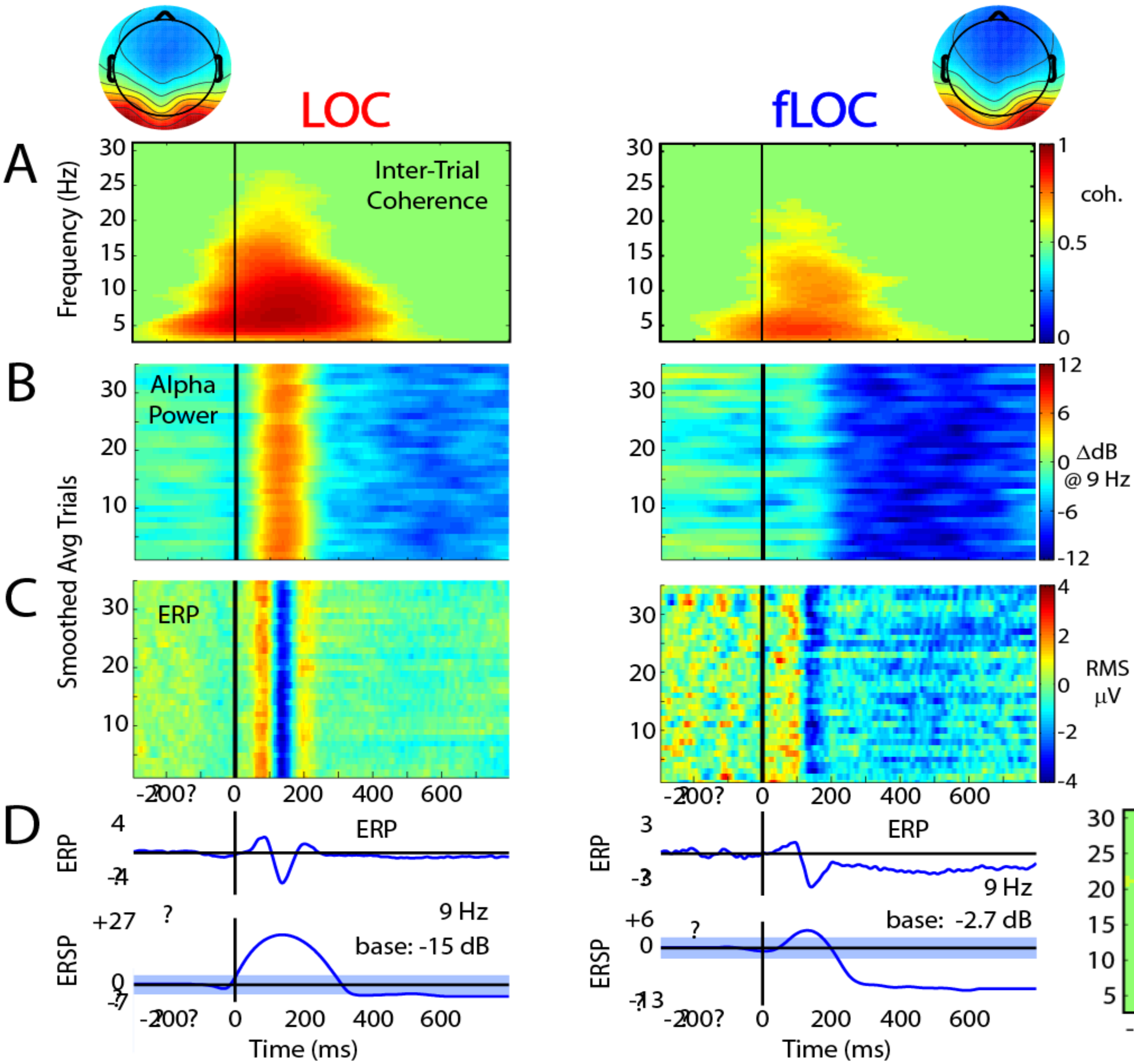




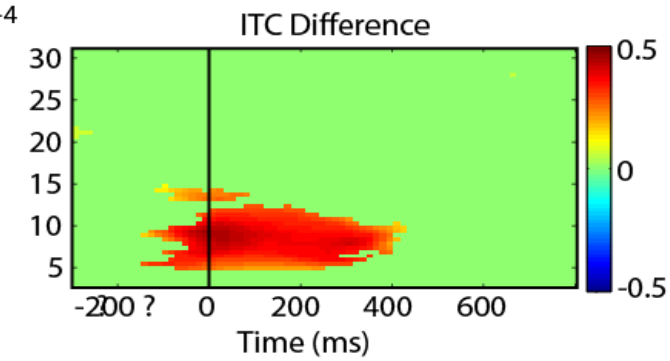
Choosing data measures



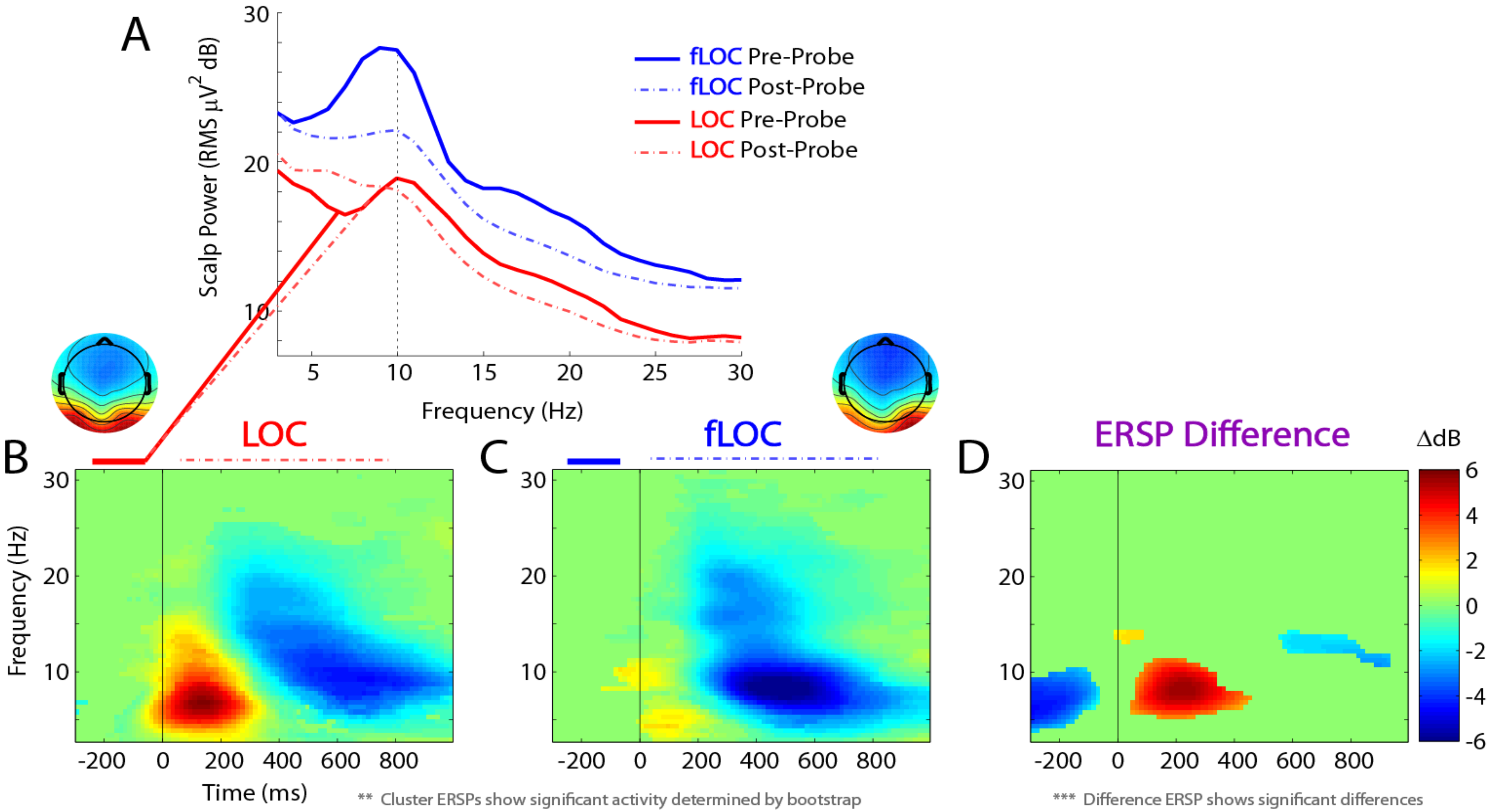
Subject differences?



Significant ITC differences (by bootstrap) between the LOC and fLOC clusters immediately follow Probe presentation (5-11 Hz).



Subject differences?

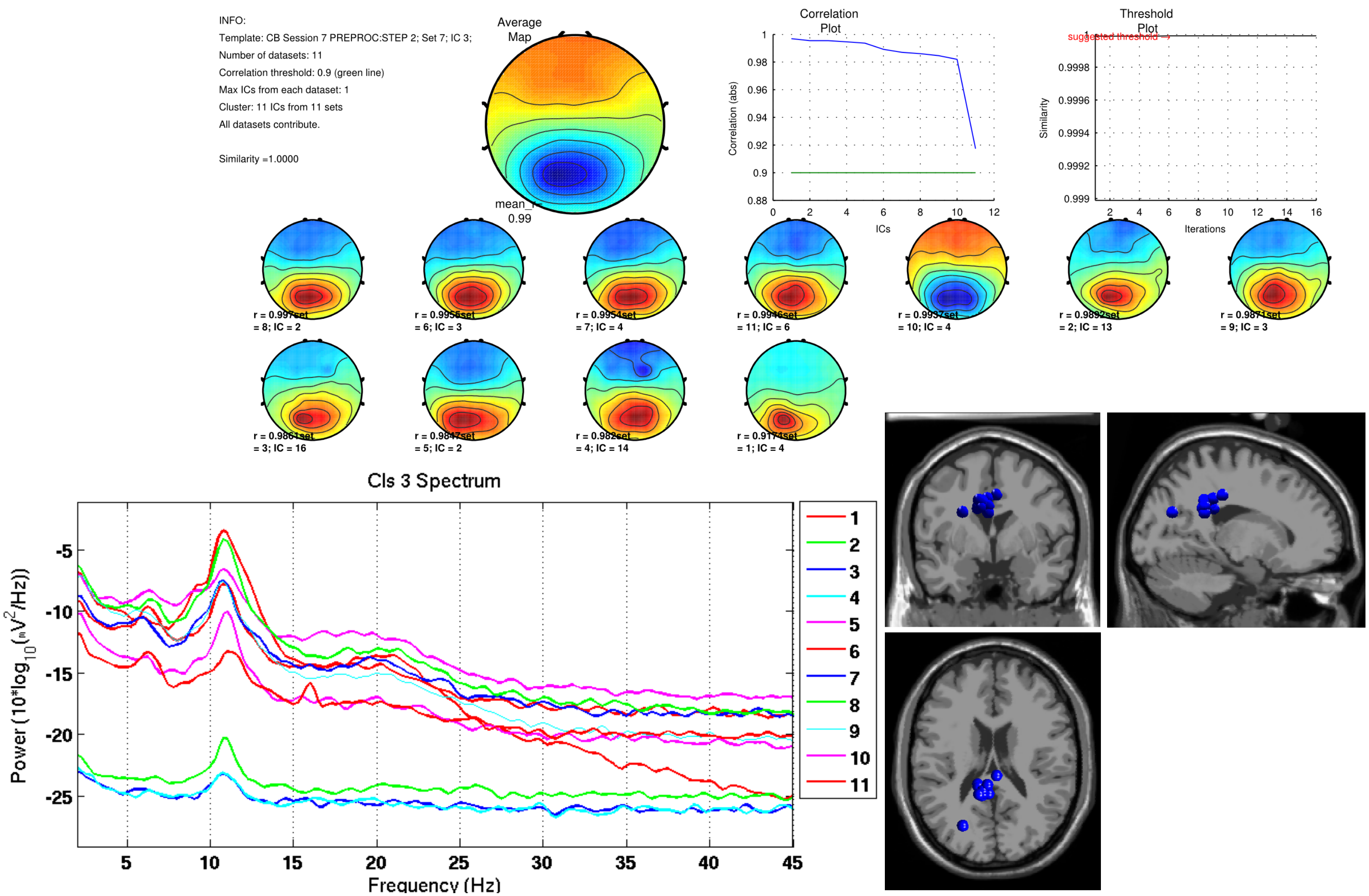


** Cluster ERSPs show significant activity determined by bootstrap statistics within subject and binomial probability between subjects ($p < 0.01$)

*** Difference ERSP shows significant differences between the two clusters by bootstrap statistics ($p < 0.001$)

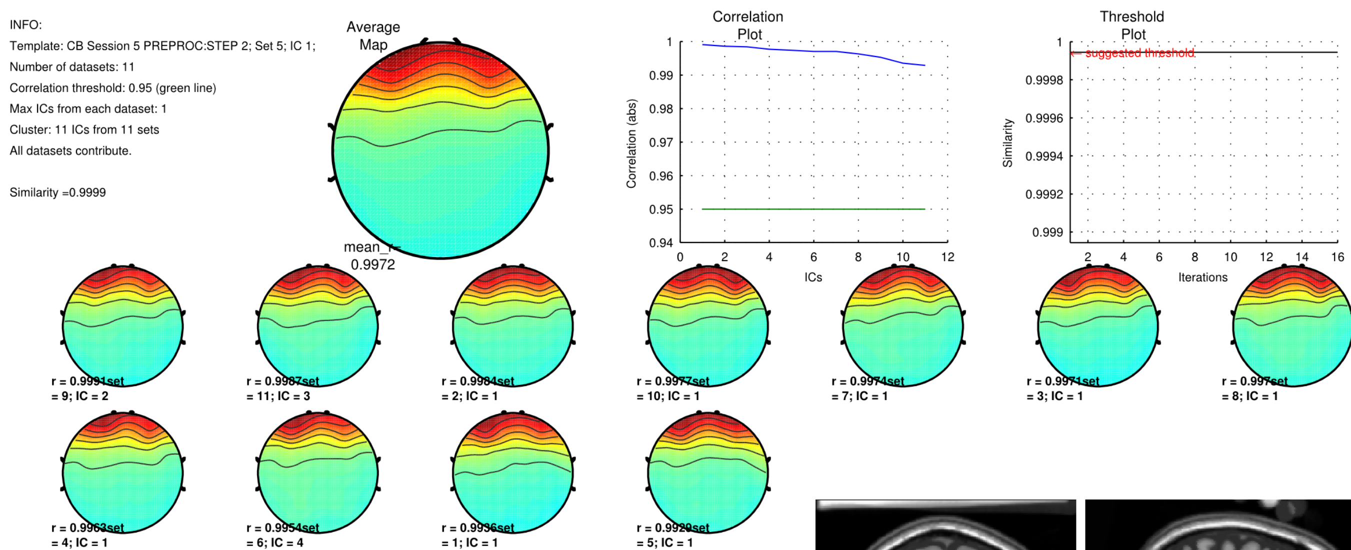
Results (Cluster 1)

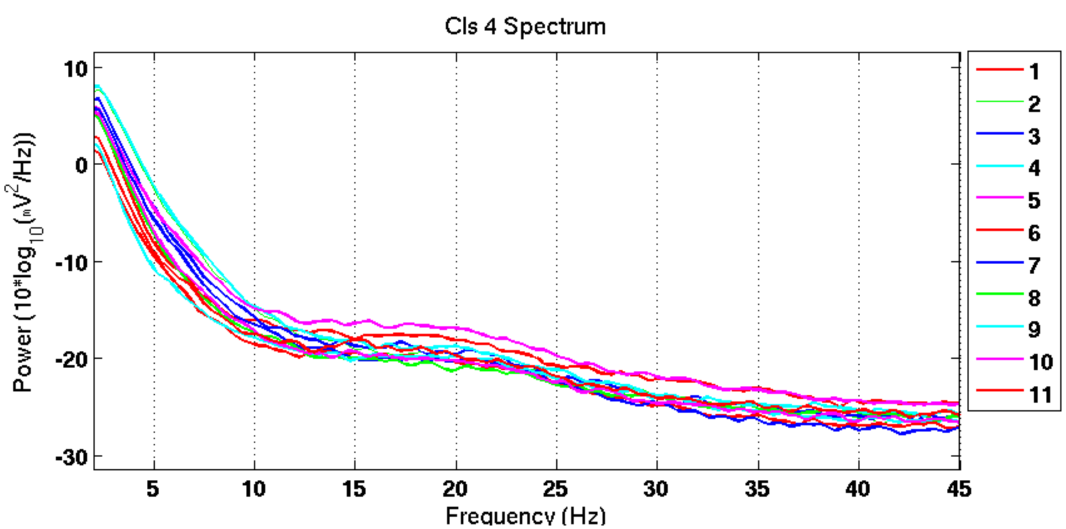
100 % Sessions contribute

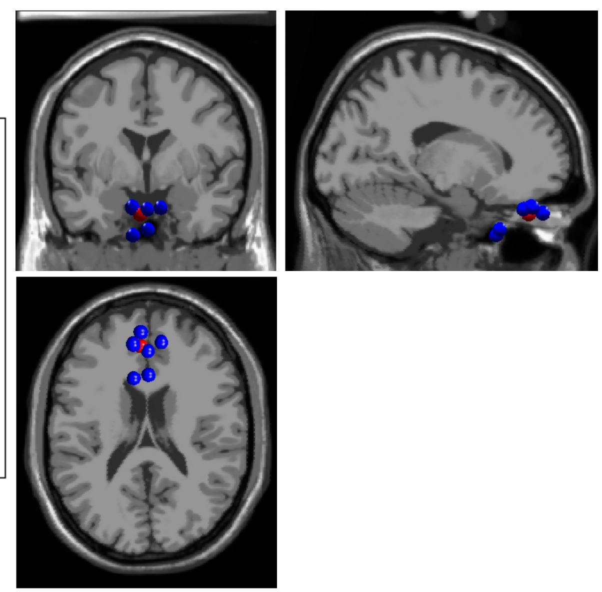


Results (Cluster 2)

100 % Sessions contribute

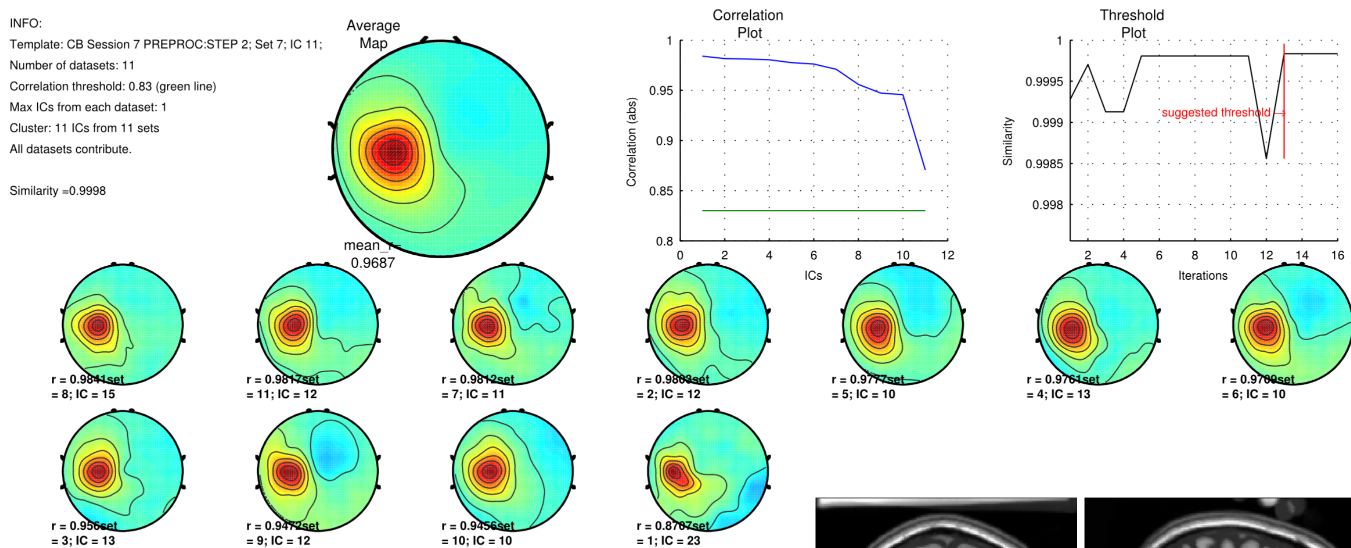


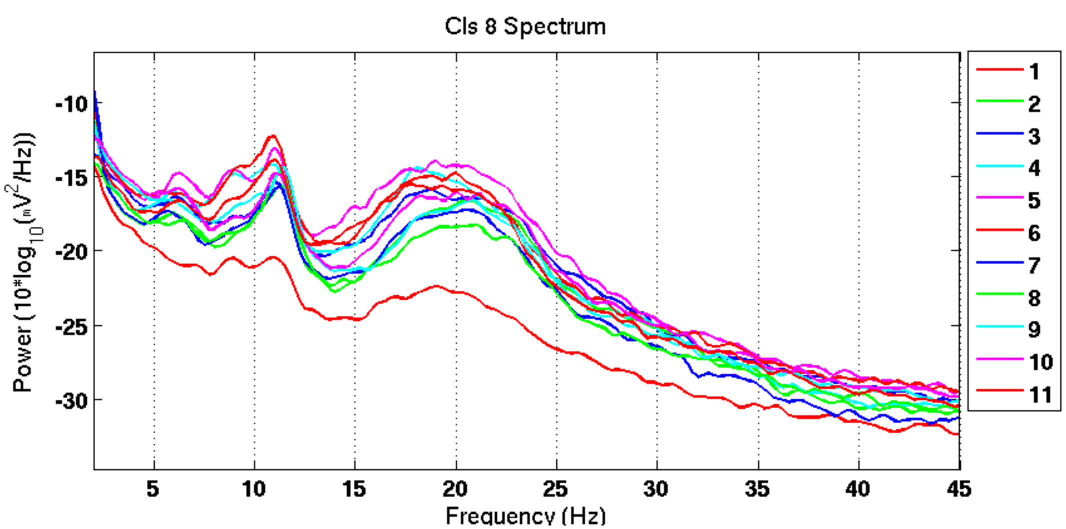


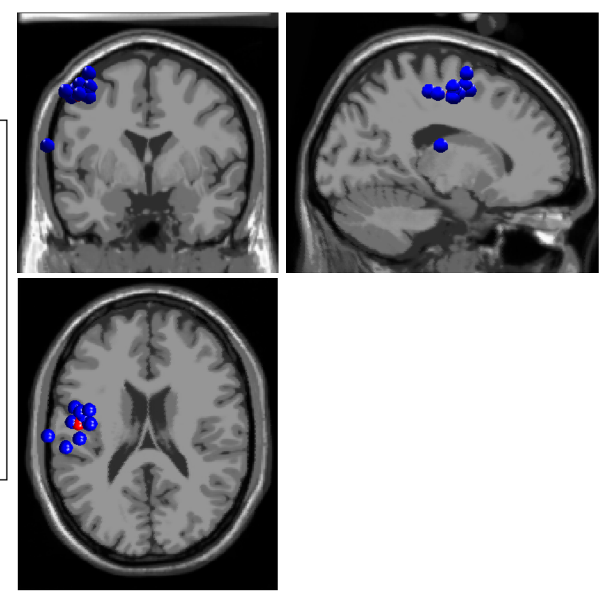


Results (Cluster 8)

100 % Sessions contribute

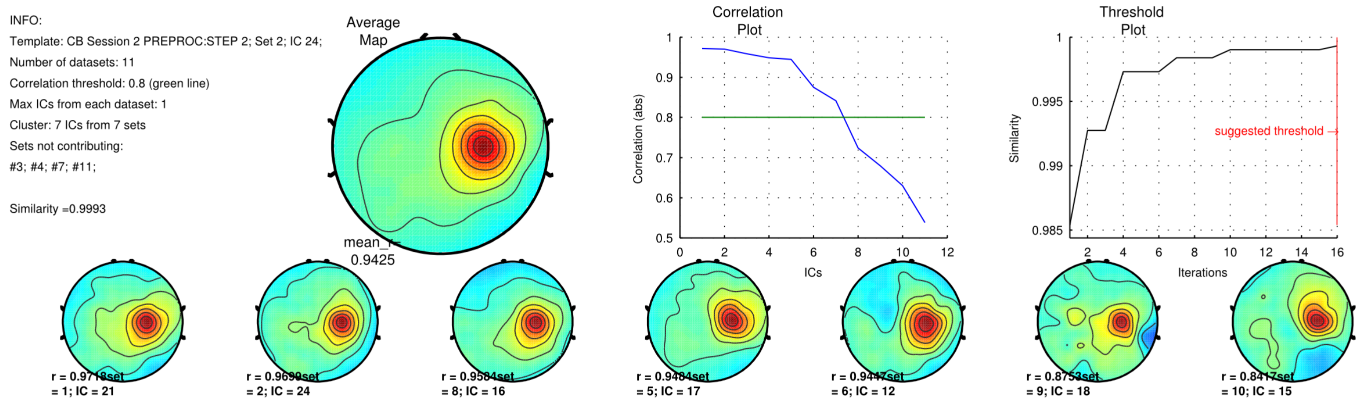


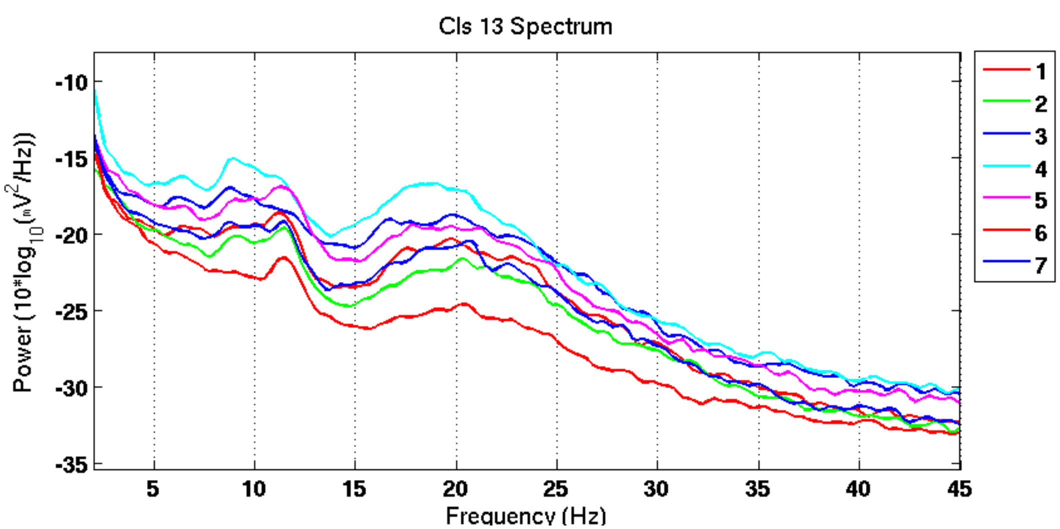


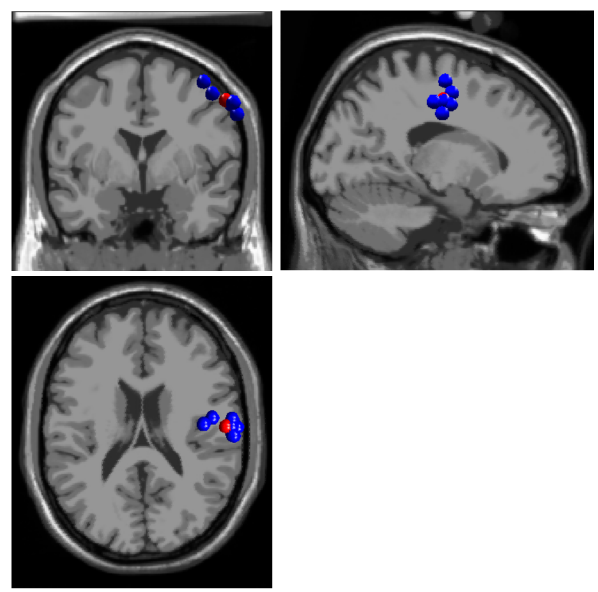


Results (Cluster 13)

63.64% Sessions contribute

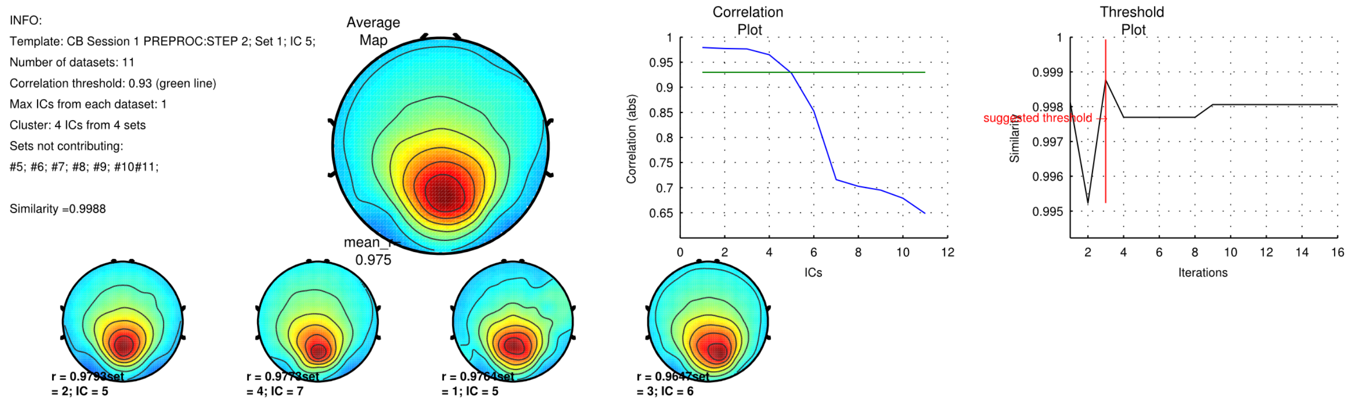


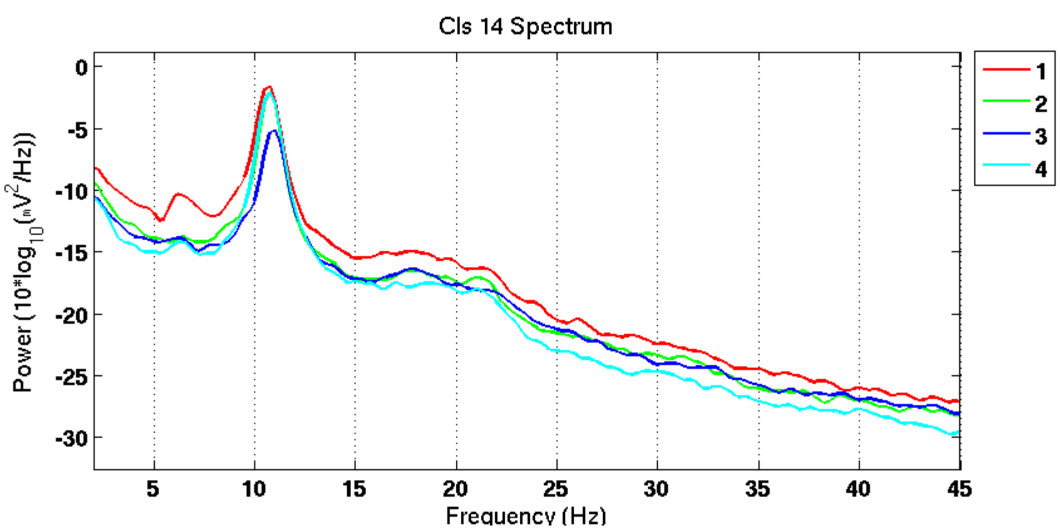


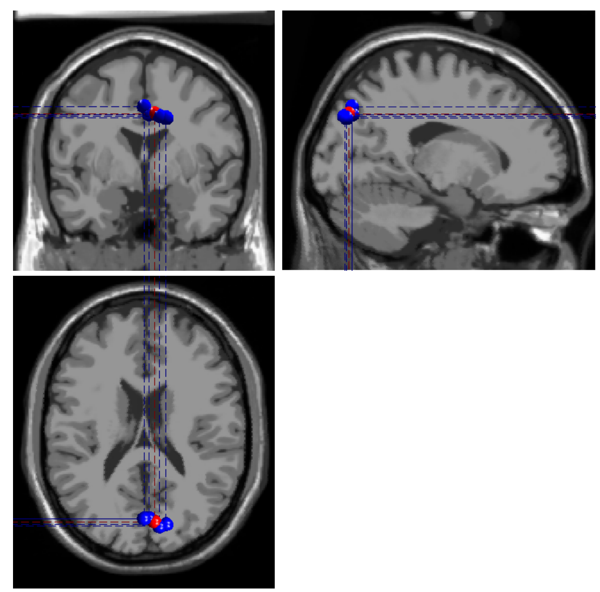


Results (Cluster 14)

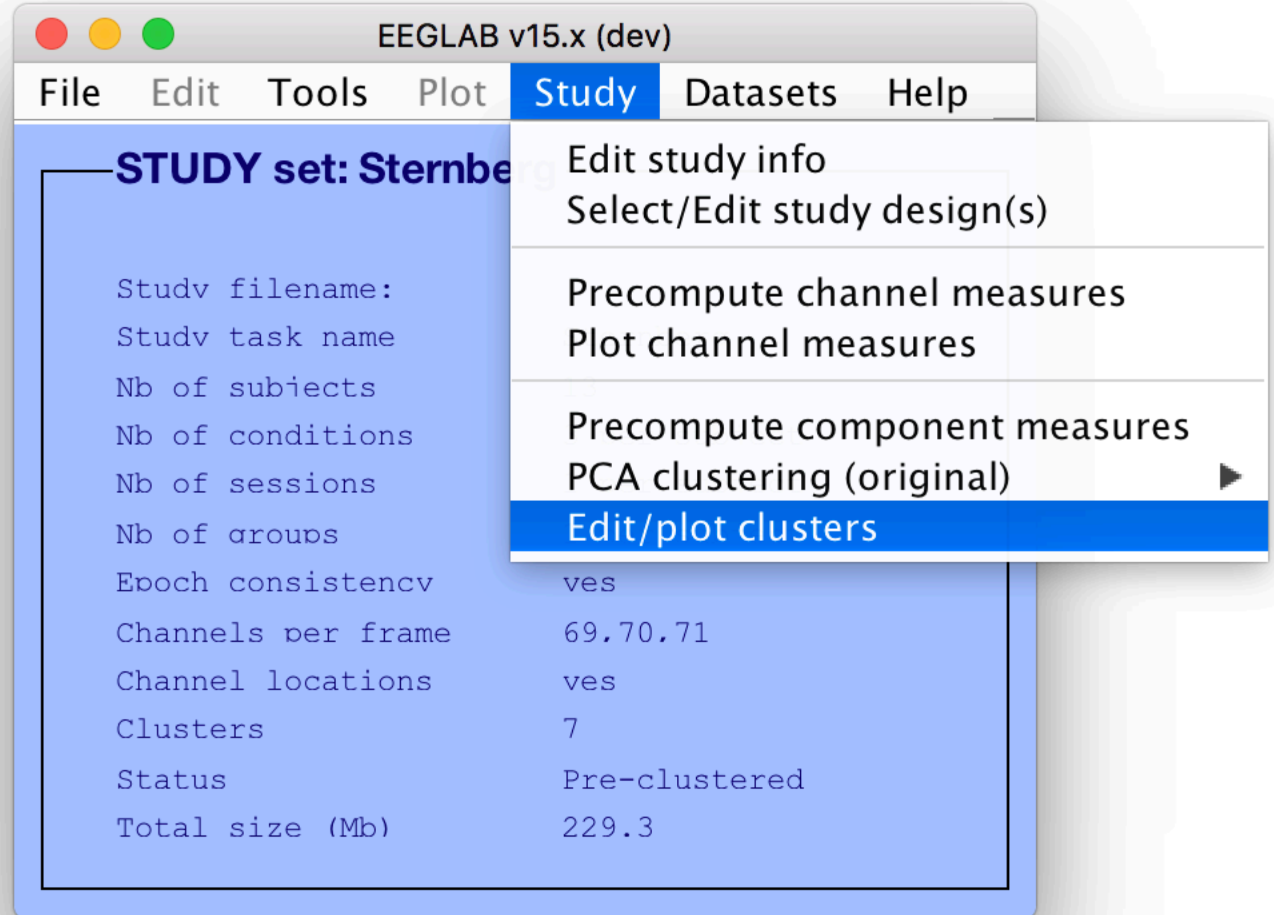
36.36% Sessions contribute



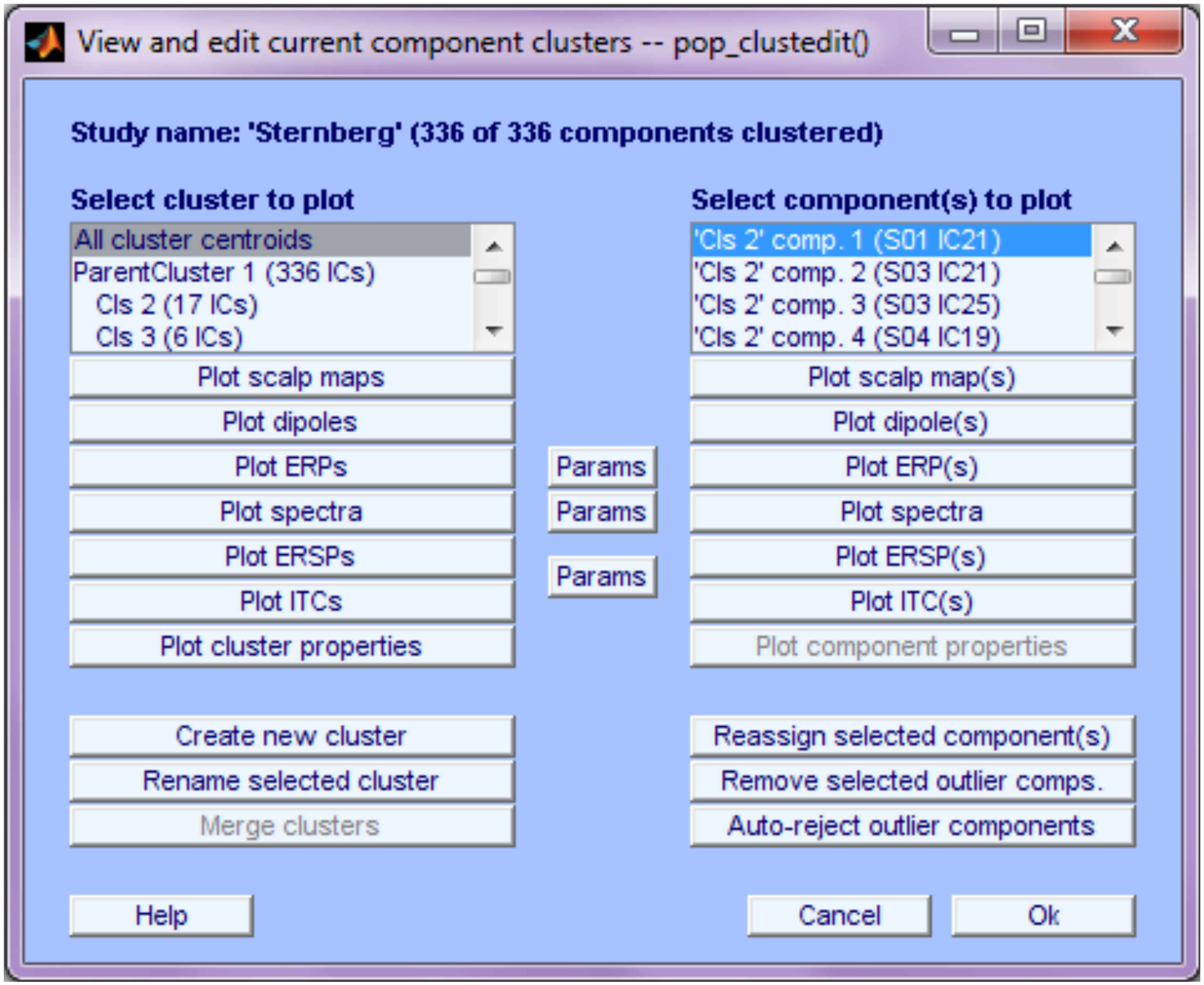




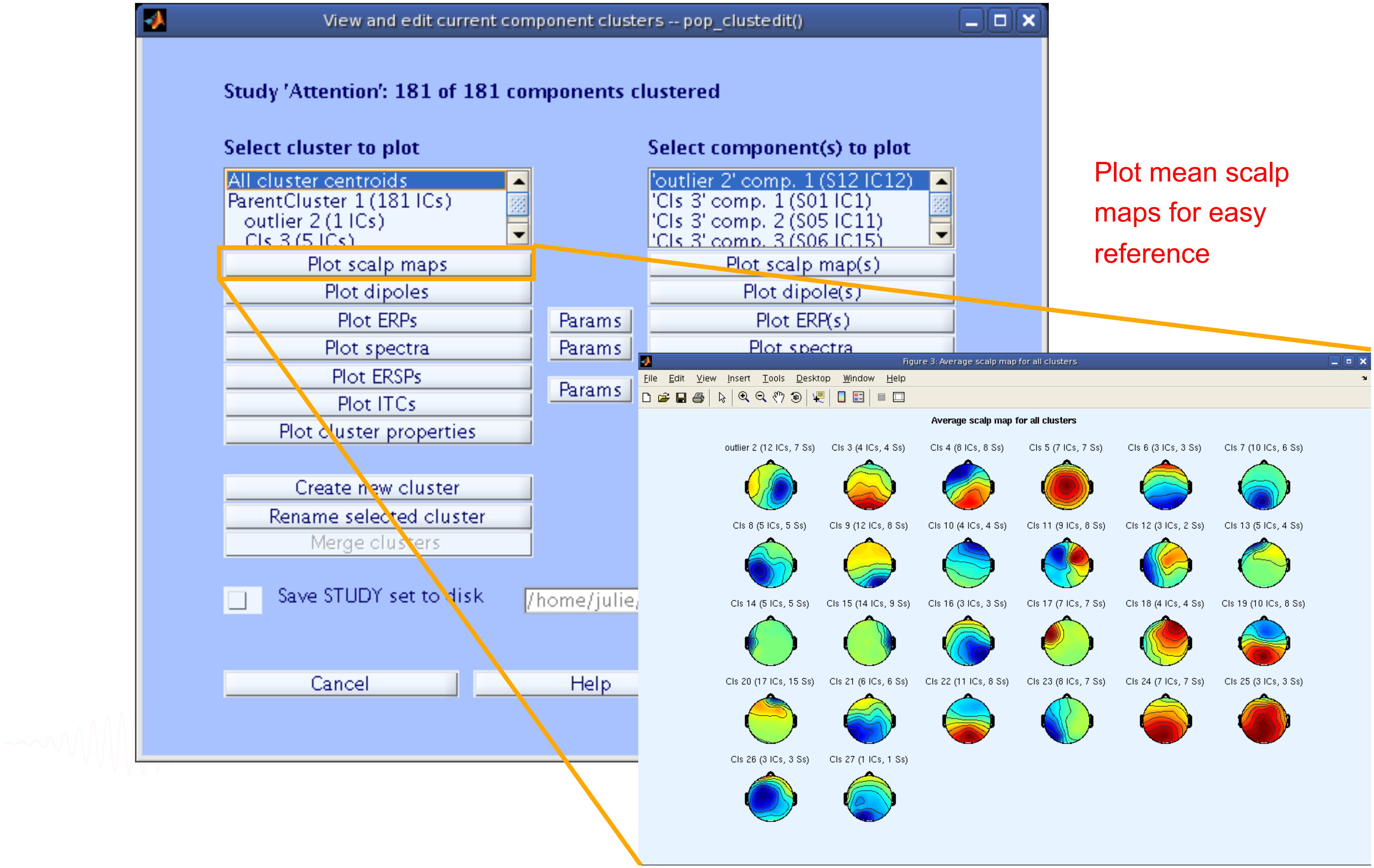
View and edit clusters



Plot/edit clusters

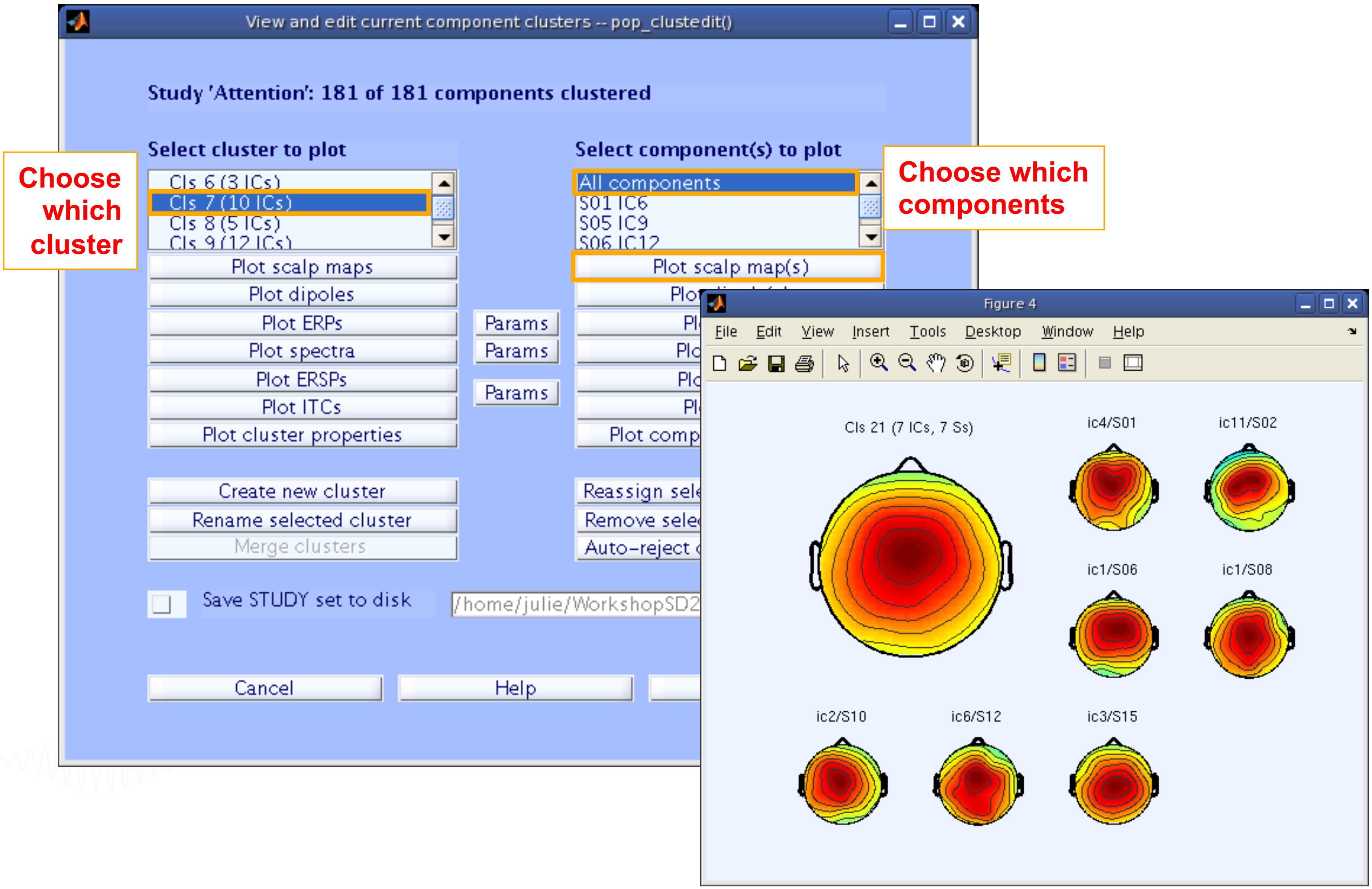


Plot cluster data

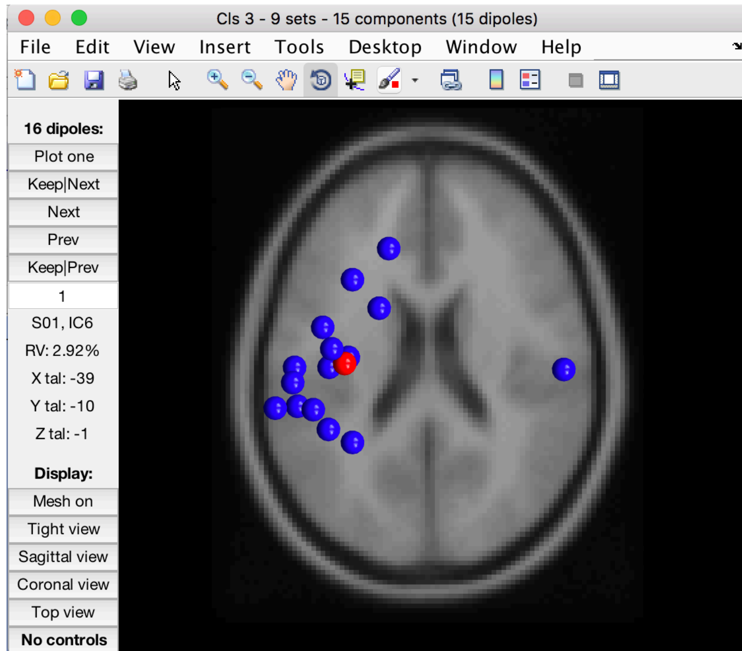


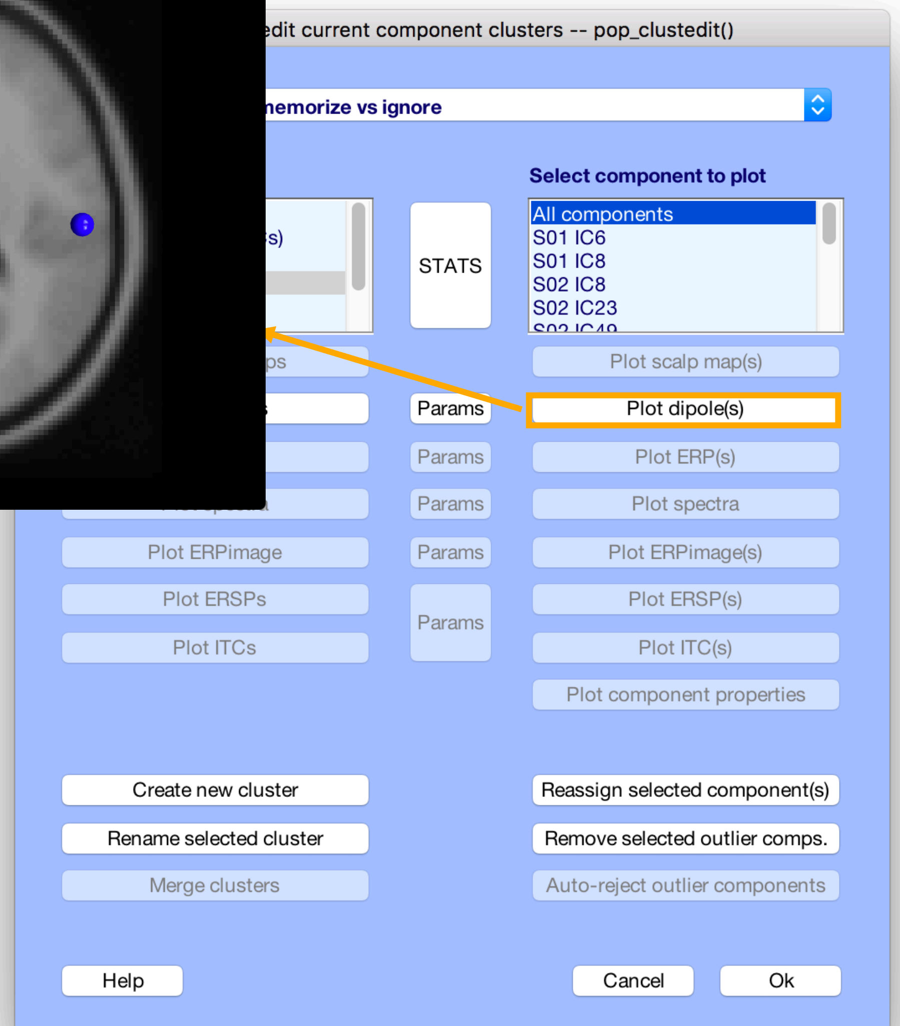
Plot mean scalp maps for easy reference

Plot cluster data



Plot cluster data



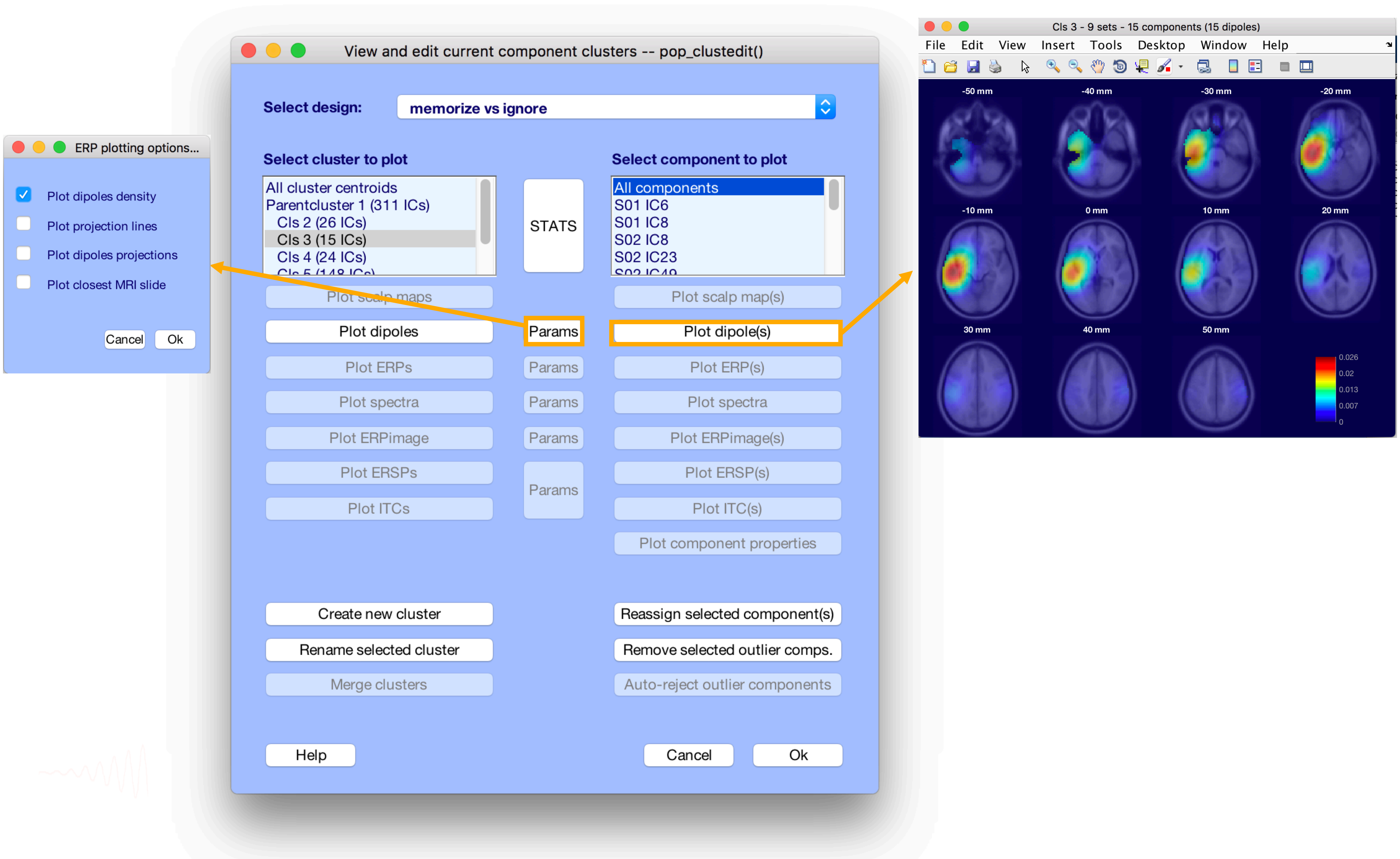


MAKE SURE THAT SIFT PATH

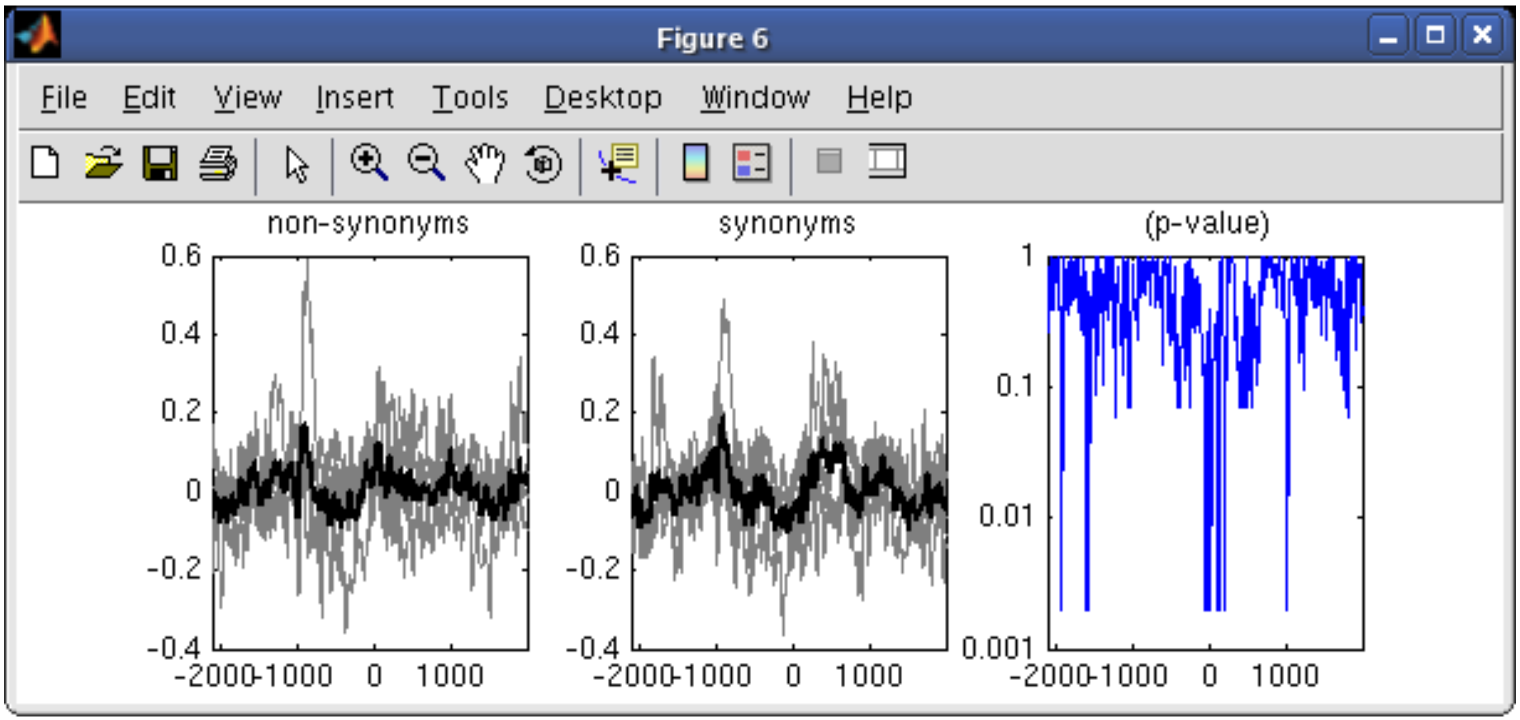
…/eeglab/plugins/sift_1.5/utils/

IS AT THE BOTTOM (EDIT PATH)

Plot cluster data



Plot cluster ERP



Exercise

- Load the STUDY stern.study
- Precompute **spectrum, ERP** and **scalp maps** for components
- Precluster and cluster components using **dipole locations** and **dipole moments** (affinity clustering)
- Look at your cluster. Identify frontal midline theta cluster and occipital alpha cluster
- Remove outliers if any
- Plot significant difference (parametric statistics) for one component cluster spectrum between the two conditions ignore vs memorize