Independent Component Analysis and Its Applications

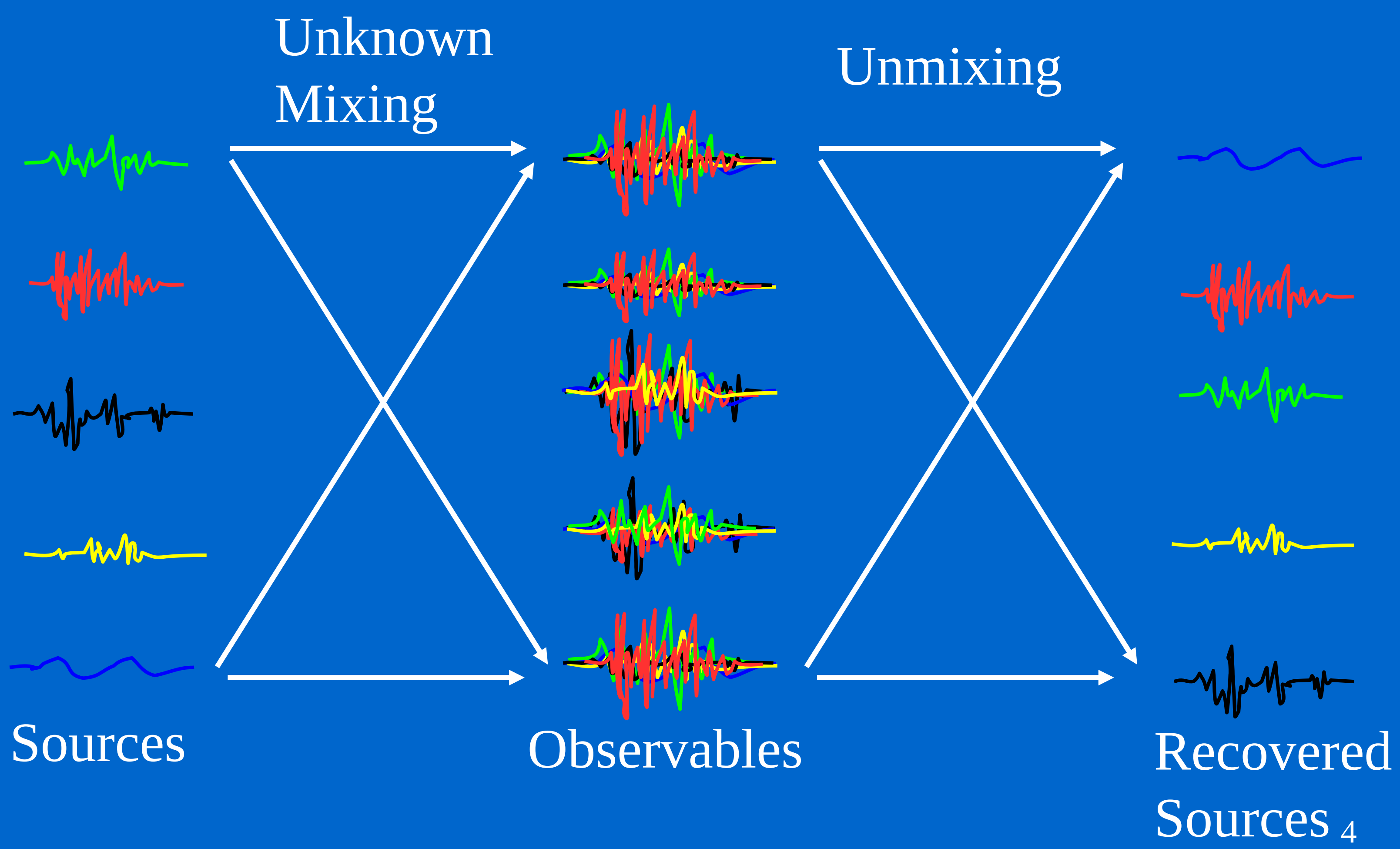
Tzyy-Ping Jung

Swartz Center for Computational Neuroscience
Institute for Neural Computation
University of California, San Diego
&
Department of Computer Science
National Chiao-Tung University, Hsinchu, Taiwan

Outline

- Blind Source Separation:
  - Solving the "cocktail party problem"
- Applications
  - Speech separation and clarity
  - EEG/ERP
  - fMRI
  - Image processing

Blind Source Separation



Brief History of ICA

- Herault & Jutten ("Space or time adaptive signal processing by neural network models“, *Neural Nets for Computing Meeting*, Snowbird, Utah, 1986): Seminal paper, neural network
- Comon (1994): Approximation of MI by 4th order statistics
- Bell & Sejnowski (1995): Information Maximization
- Amari et al. (1996): Natural Gradient Learning
- Cardoso (1996): JADE
- Hyvärinen & Oja: Nonlinear PCA, FastICA

- Applications of ICA to biomedical signals
  - EEG/ERP analysis (Makeig, Bell, Jung & Sejnowski, 1996; Jung et al., 1997; Makeig et al., 1997; Jung et al., 2001)
  - fMRI analysis (McKeown, Jung et al. 1998, Jung et al., 2001)
  - ECG analysis (Cardoso 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)

- Pearlmutter & Parra showed InfoMax, ML estimation are equivalent.
- Lee et al. showed negentropy has the equivalent property to InfoMax.
- Girolami & Fyfe showed nonlinear PCA can be viewed from information-theoretic principle.
- A unifying Information-theoretic framework for ICA (Lee et al. 1999)

Independent Component Analysis

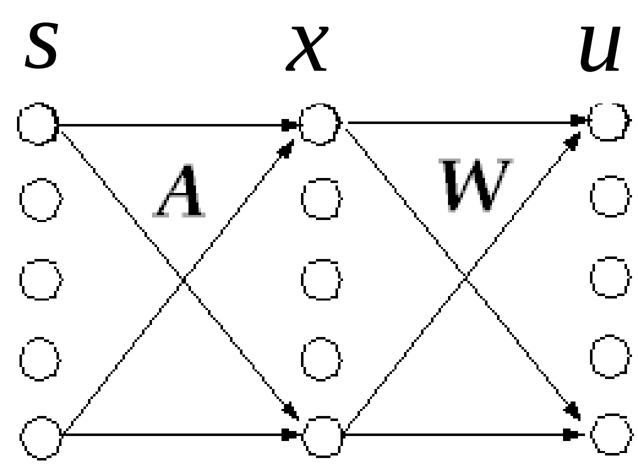
ICA is a method to recover a version, of the original sources by multiplying the data by a unmixing matrix,

$$\mathbf{u} = \mathbf{W}\mathbf{x},$$

where x is our observed signals, a linear mixtures of sources,

$$\mathbf{x} = \mathbf{A}\mathbf{s}.$$

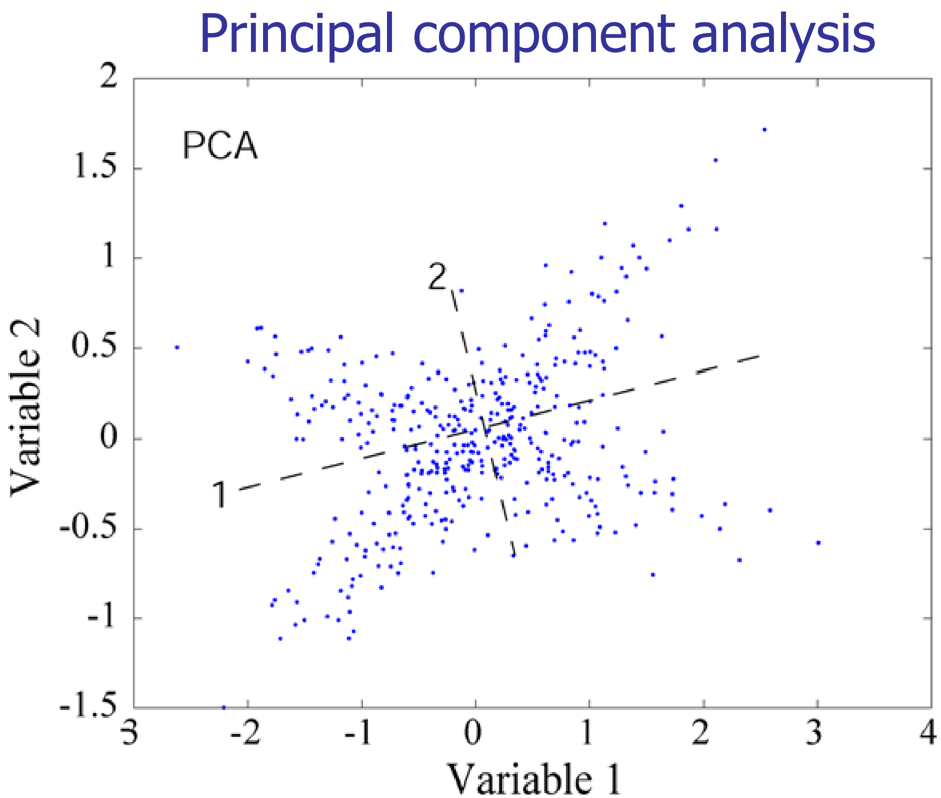
While PCA simply decorrelates the outputs (using an orthogonal matrix **W**), ICA attempts to make the outputs **statistically independent**, while placing no constraints on the matrix **W**



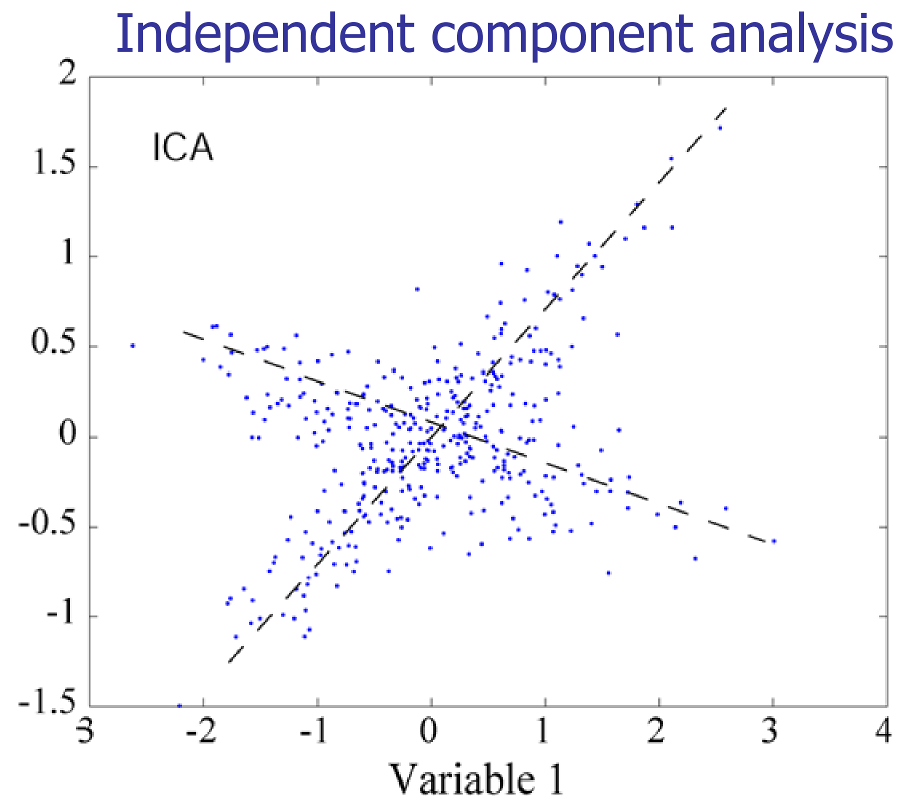
WA after learning:

$$\begin{bmatrix} \boxed{-4.09} & 0.13 & 0.09 & -0.07 & -0.01 \\ 0.07 & \boxed{-2.92} & 0.00 & 0.02 & -0.06 \\ 0.02 & -0.02 & -0.06 & -0.08 & \boxed{-2.20} \\ 0.02 & 0.03 & 0.00 & \boxed{1.97} & 0.02 \\ -0.07 & 0.14 & \boxed{-3.50} & -0.01 & 0.04 \end{bmatrix}$$

ICA vs PCA



Principal component analysis
PCA
1
2
Variable 1
Variable 2
-3
-2
-1
0
1
2
3
4
2
1.5
1
0.5
0
-0.5
-1
-1.5



Independent component analysis
ICA
Variable 1
-3
-2
-1
0
1
2
3
4
2
1.5
1
0.5
0
-0.5
-1
-1.5

Statistical Independence

Statistical Independence:

$$f_{\mathbf{s}}(\mathbf{s}) = \prod_{i=1}^{N} f_{s_i}(s_i)$$

Or the mutual information:

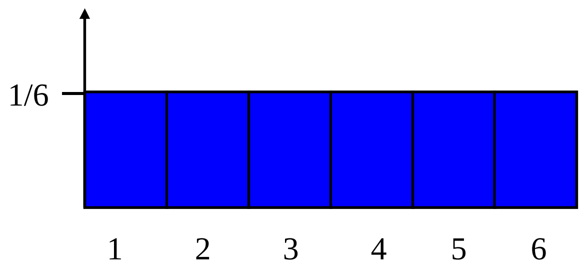
$$I(s_i, s_j) = E\left[\ln \frac{f_{\mathbf{s}}(\mathbf{s})}{\prod_{i=1}^{N} f_{s_i}(s_i)}\right] = 0, for\ \forall i \neq j$$

The problem of blind separation is to find $\mathbf{W}$ such that the linear transformation $\mathbf{u} = \mathbf{W}\mathbf{x} = \mathbf{W}\mathbf{A}\mathbf{s}$ reestablishes the condition of statistical independence.

Entropy

$$H(X) = -\sum_{x \in X} p(x) \log(p(x))$$

Dice: 1/6



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

ICA learning rule

How to make the outputs statistically independent?

Minimize their redundancy or mutual information.

Entropy: $$H(X) = -\sum_{x \in X} p(x)\log(p(x))$$

Joint entropy $$H(X,Y) = -\sum_{(x,y)\in X \times Y} p(x,y)\log(p(x,y))$$

Mutual Information $I(y_1,y_2) = H(y_1) + H(y_2) - H(y_1,y_2)$

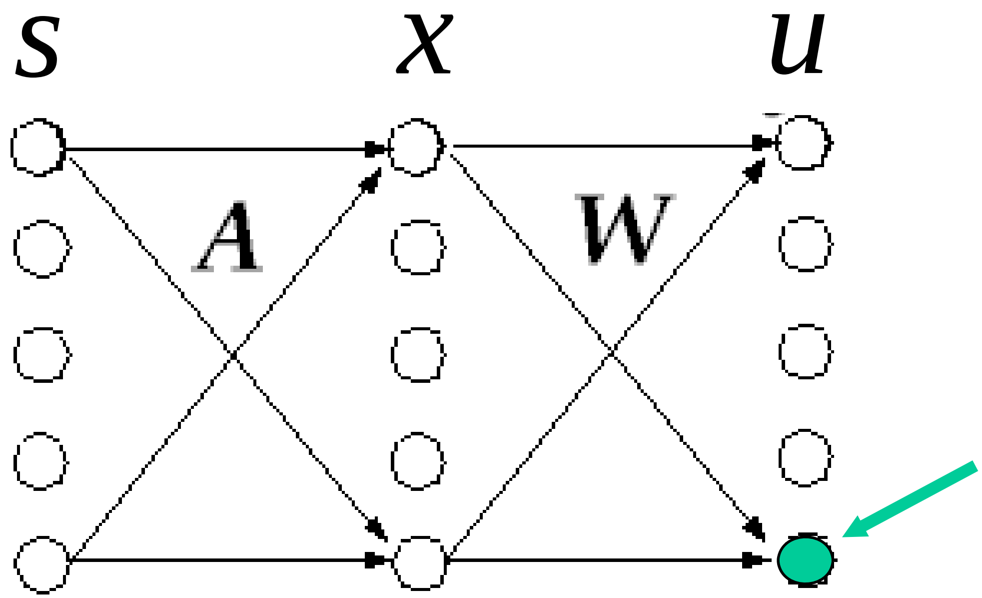
Minimizing $I(y_1,y_2)$ → **Maximizing $H(y_1,y_2)$**

=0 if the two variables are independent

ICA learning rule

$$\Delta W = \frac{\partial H(y_1, y_2, \ldots)}{\partial W} \underbrace{W^T W}_{\text{Natural gradient (Amari)}}$$

Independent Component Analysis

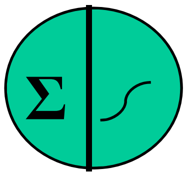


ICA is a method to recover a version, of the original sources by multiplying the data by a unmixing matrix,

$$\mathbf{u} = \mathbf{W}\mathbf{x}$$

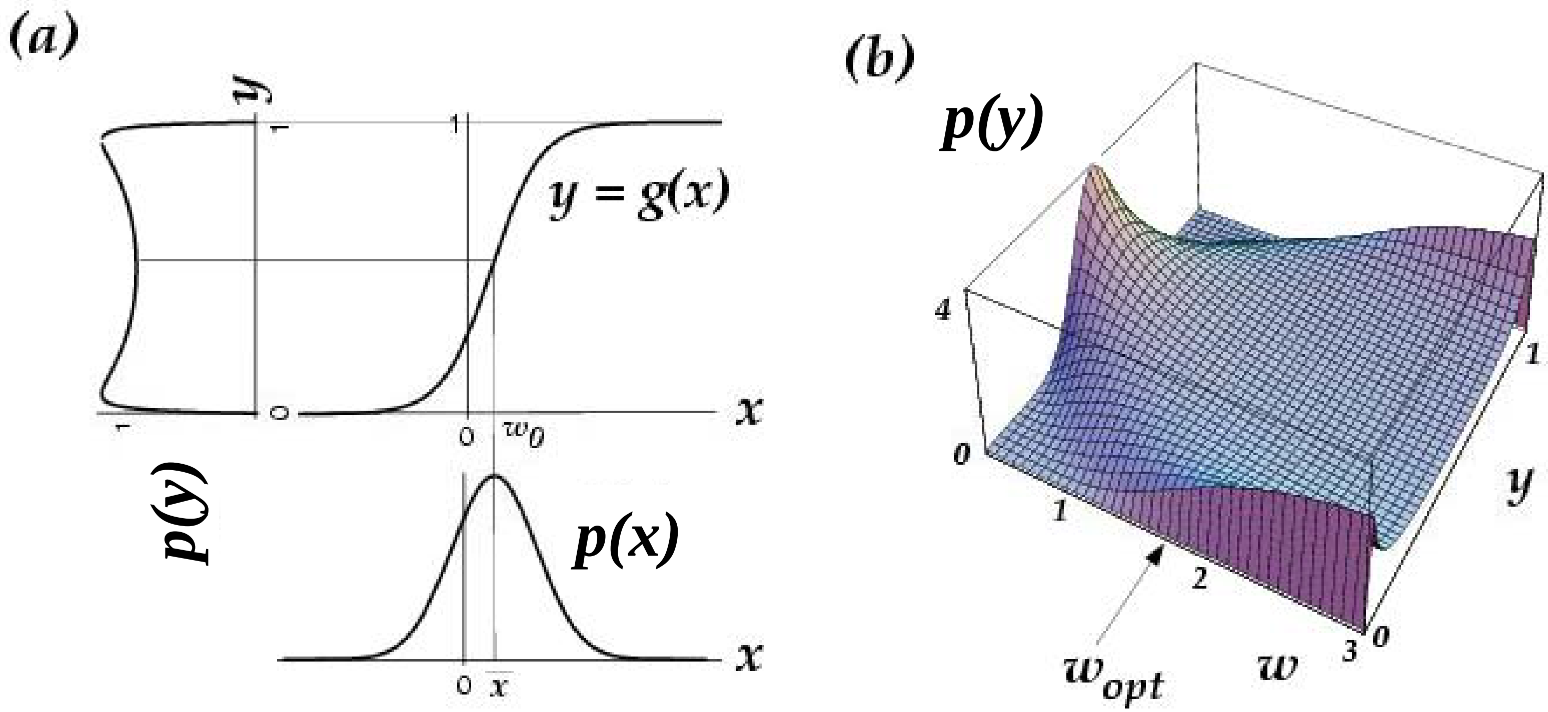
InfoMax (Bell & Sejnowski, 1995)

To make the u_i independent, we need to operate on non-linear transformed output variables, $\mathbf{y} = g(\mathbf{u})$, such as



$$\mathbf{y} = \frac{1}{1 + e^{-\mathbf{u}}}, \qquad \mathbf{u} = \mathbf{W}\mathbf{x} + \mathbf{w}_0$$

The non-linear function provides all the higher-order statistics necessary to establish independence.



From Bell & Sejnowski *Neural Compu. 1995.*

ICA learning rule

The learning rule:

$$\Delta \mathbf{W} \propto \frac{\partial H(\mathbf{y})}{\partial \mathbf{W}} \mathbf{W}^T \mathbf{W} = \left[\mathbf{I} + \phi \mathbf{u}^T \right] \mathbf{W},$$

where $\phi_i = (\partial / \partial u_i) \ln(\partial y_i / \partial u_i)$.

For super-Gaussian,

$$\phi_i = 1 - 2y_i (for \quad logistic \quad nolinearity).$$

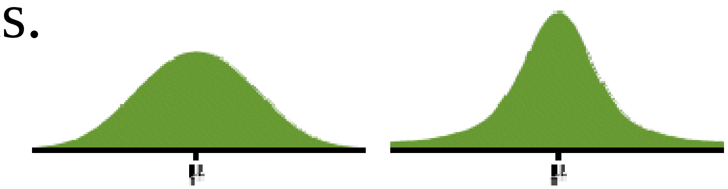
For sub- and/or super-Gaussian,

$$\phi_i = \begin{cases} +\tanh(u_i) - u_i & kurtosis < 0 \\ -\tanh(u_i) - u_i & kurtosis > 0 \end{cases}$$

Kurtosis, Super- and Sub-Gaussian

Kurtosis: a measure of how peaked or flat of a probability distribution is.

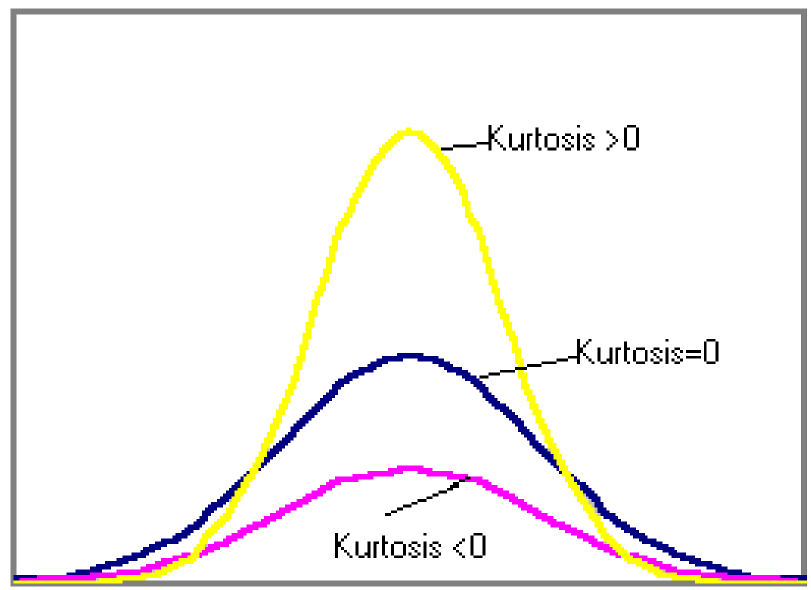
$$kurt(X) = \frac{E\left[(X-\mu)^4\right]}{\sigma^4} - 3$$



Gaussian Dis. Kurtosis = 0

Super-Gaussian: kurtosis > 0

Sub-Gaussian: kurtosis < 0



- Remove the mean

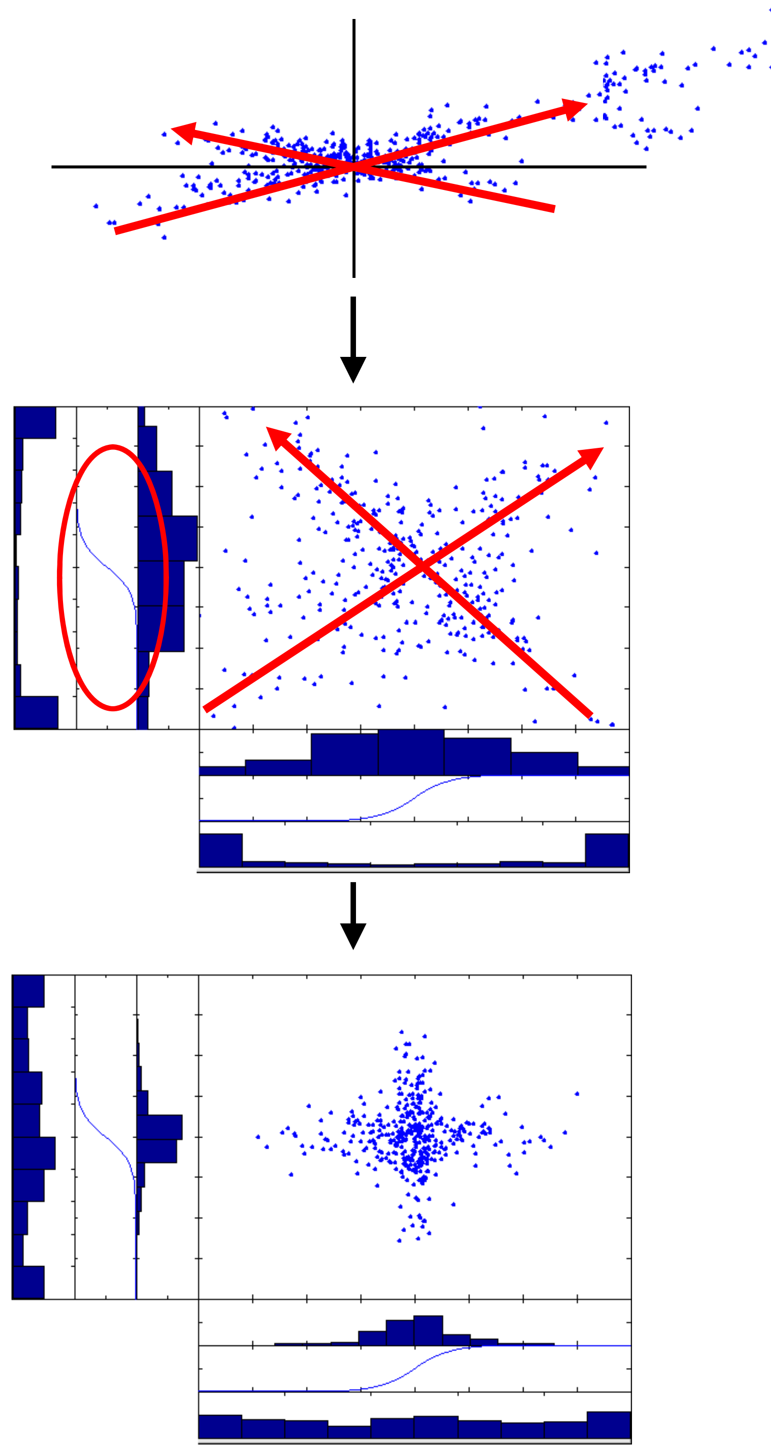
  $x = x - \langle x \rangle$.

- ‘Sphere’ the data by diagonalizing its covariance matrix,

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

- Update W according to

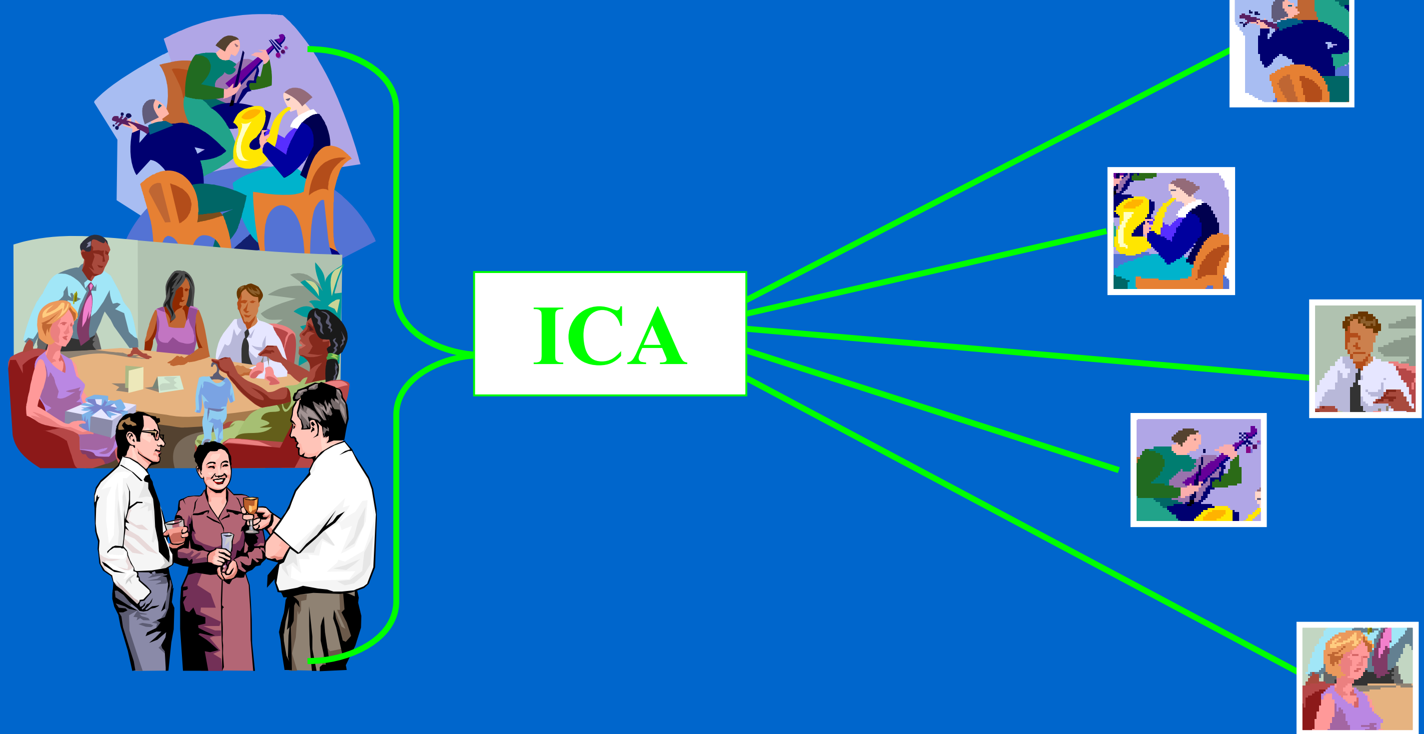
$$\Delta \mathbf{W} \propto \frac{\partial H(\mathbf{y})}{\partial \mathbf{W}} \mathbf{W}^T \mathbf{W} = \left[\mathbf{I} + \phi \mathbf{u}^T\right] \mathbf{W},$$



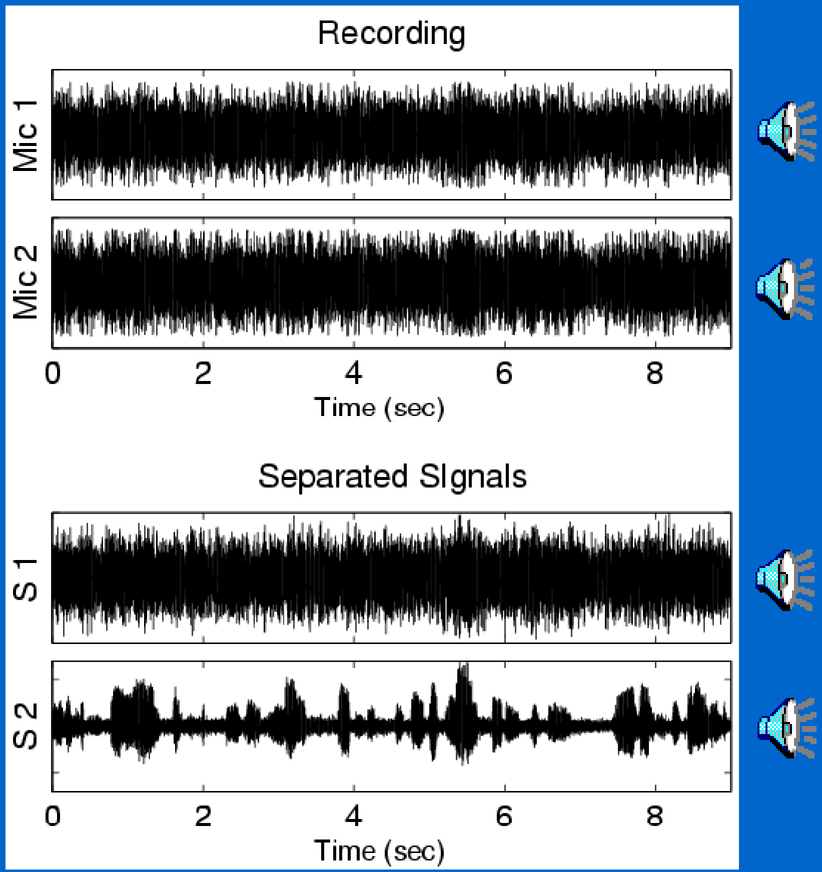
ICA Applications

- Speech enhancement (noisy speech recognition)
- Biomedical signal processing (**EEG**, ERP, **fMRI**, MEG)
- Image processing

Example: Speech Separation



Speech Enhancement & Recognition

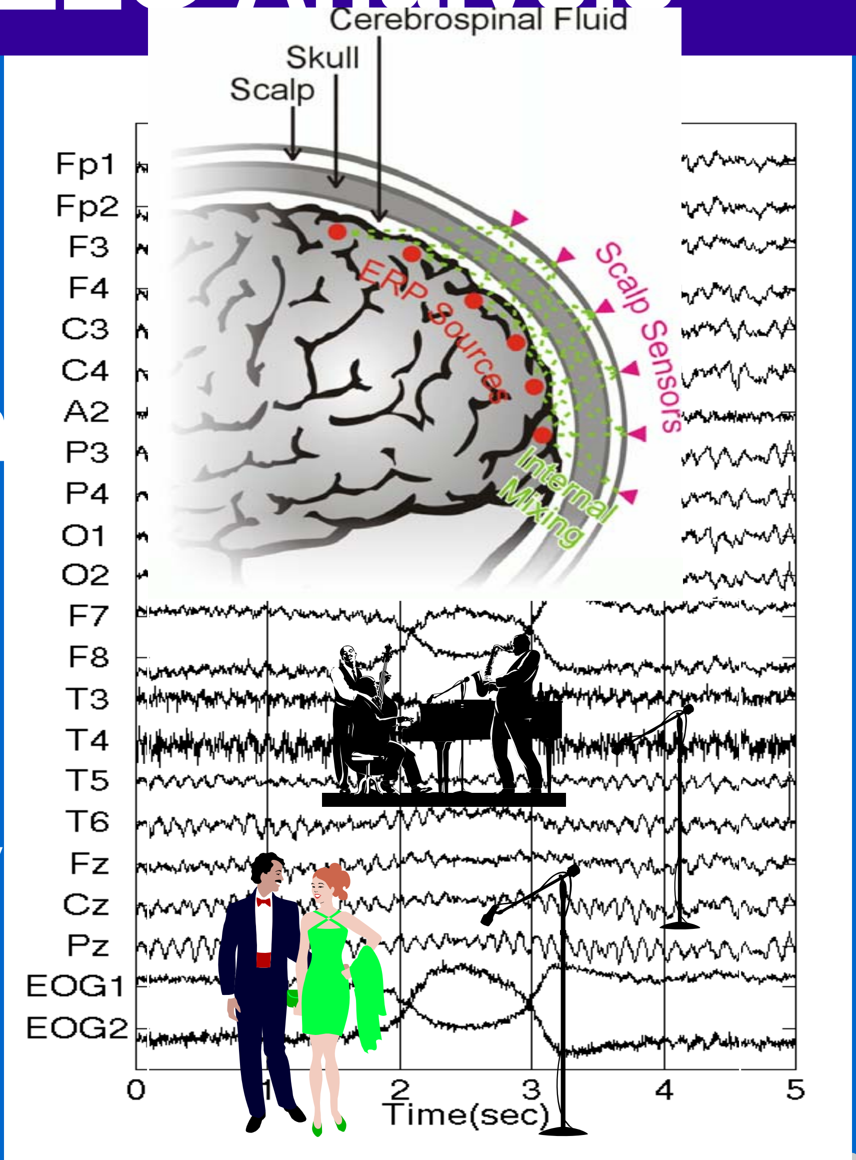


ICA Applications

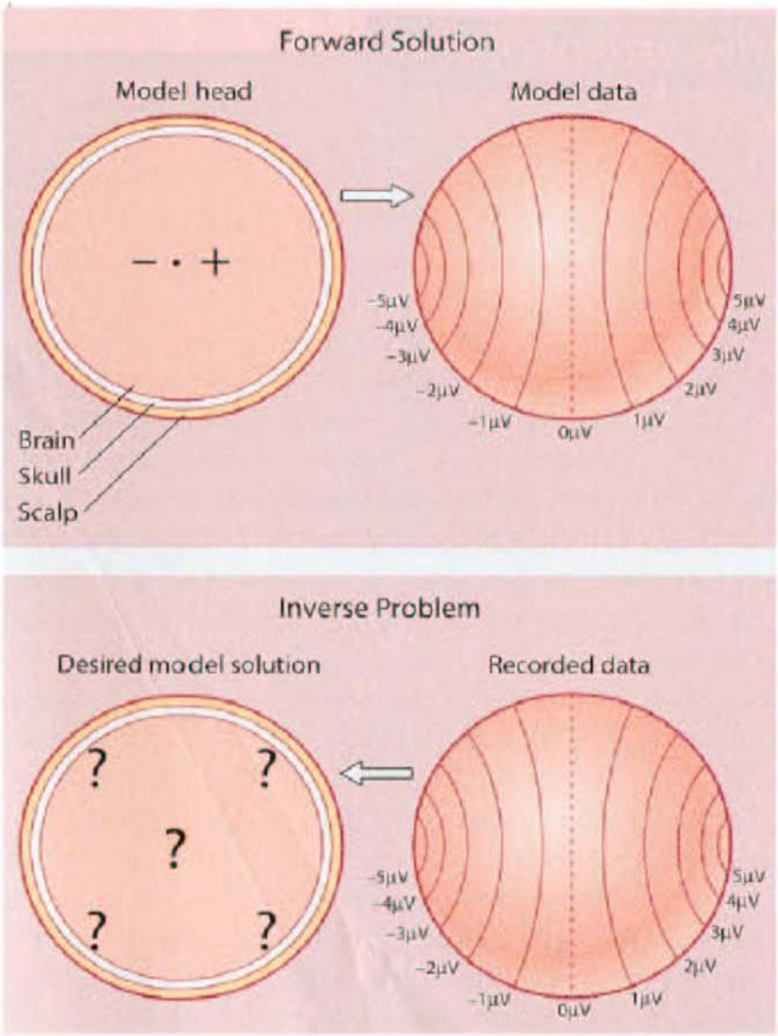
- Speech enhancement (noisy speech recognition)
- Biomedical signal processing (**EEG**, ERP, **fMRI**, MEG)
- Image processing

Challenges of EEG Analysis

- Pervasive artifacts
- EEG recordings are mixtures of all brain activities arising from different networks
- Response variability
- Inverse problem
- etc

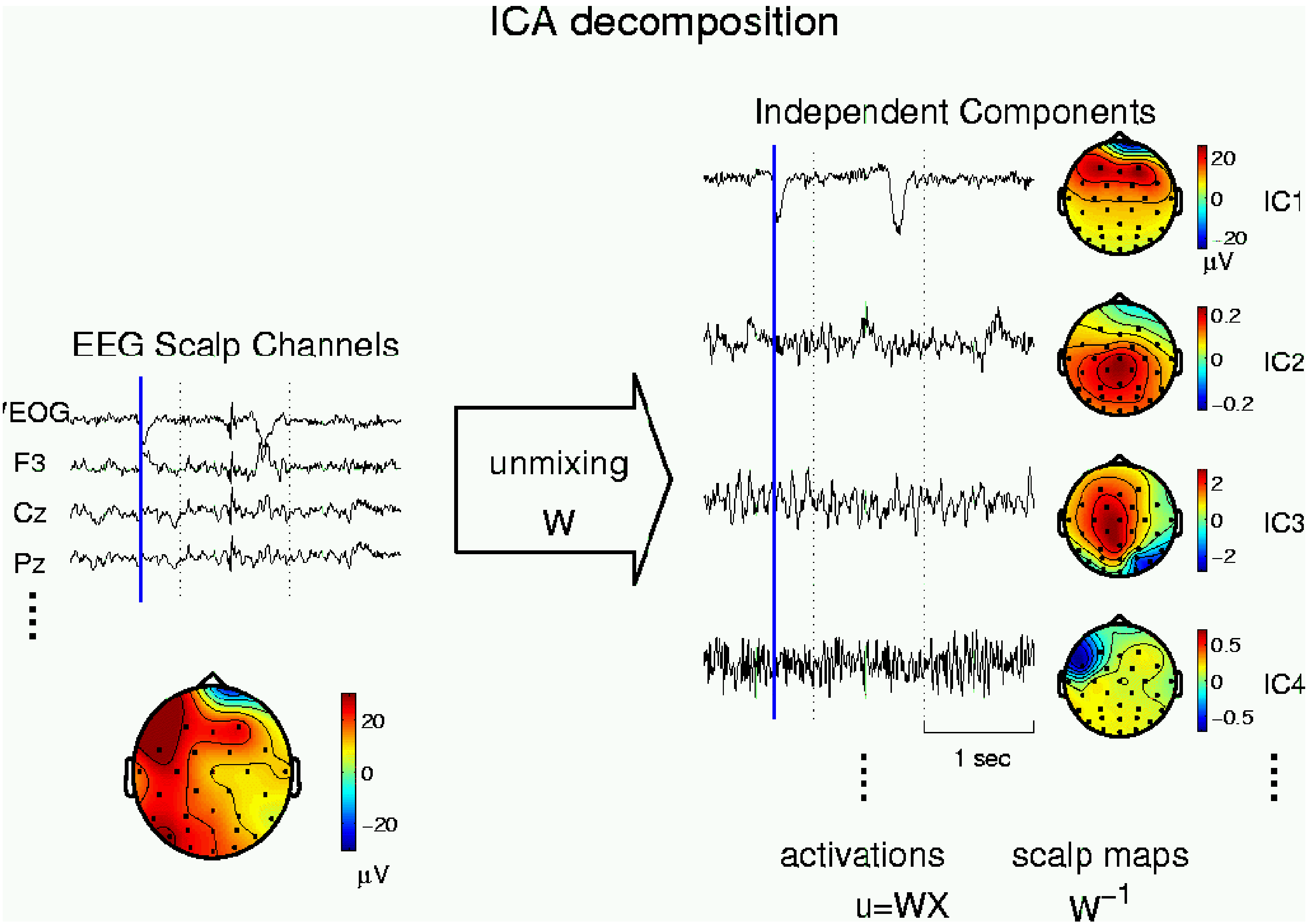


Inverse solution is not unique



A single pattern of neural activity will produce a unique scalp map

BUT …A single scalp map could have been produced by an infinite number of patterns of neural activity

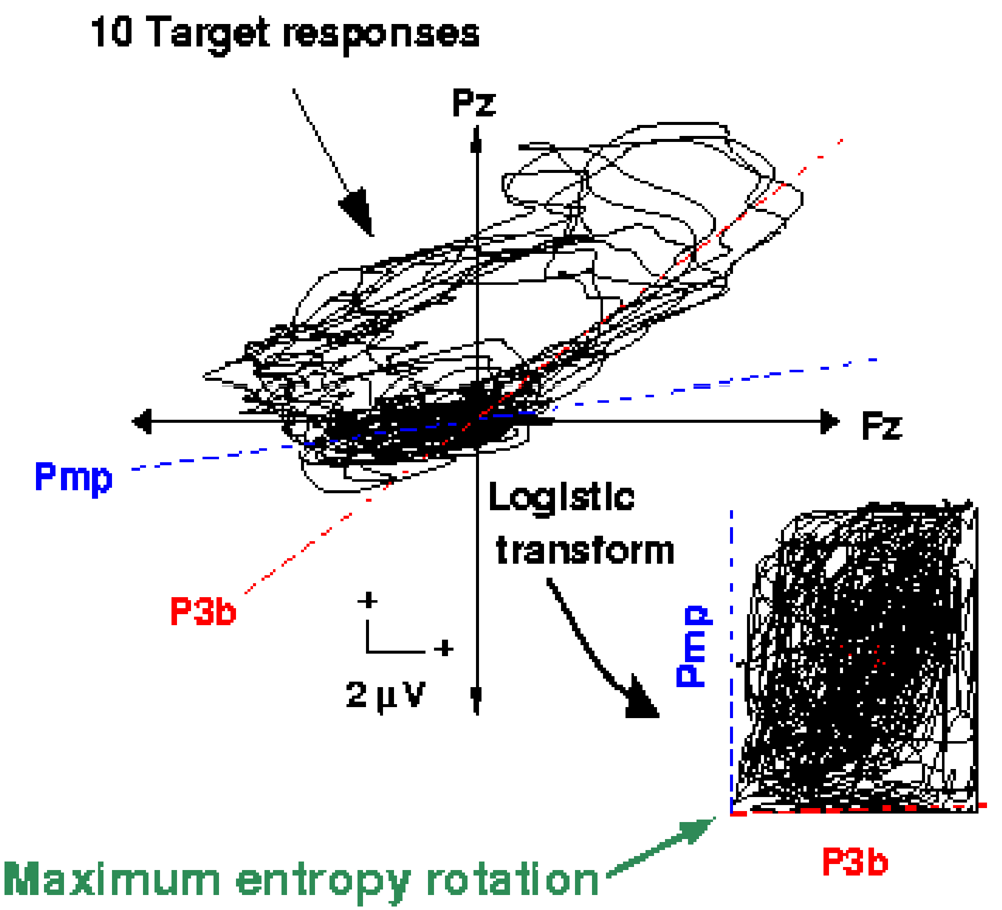


From Jung et al., *Clinical Neurophysiology*, 2000.

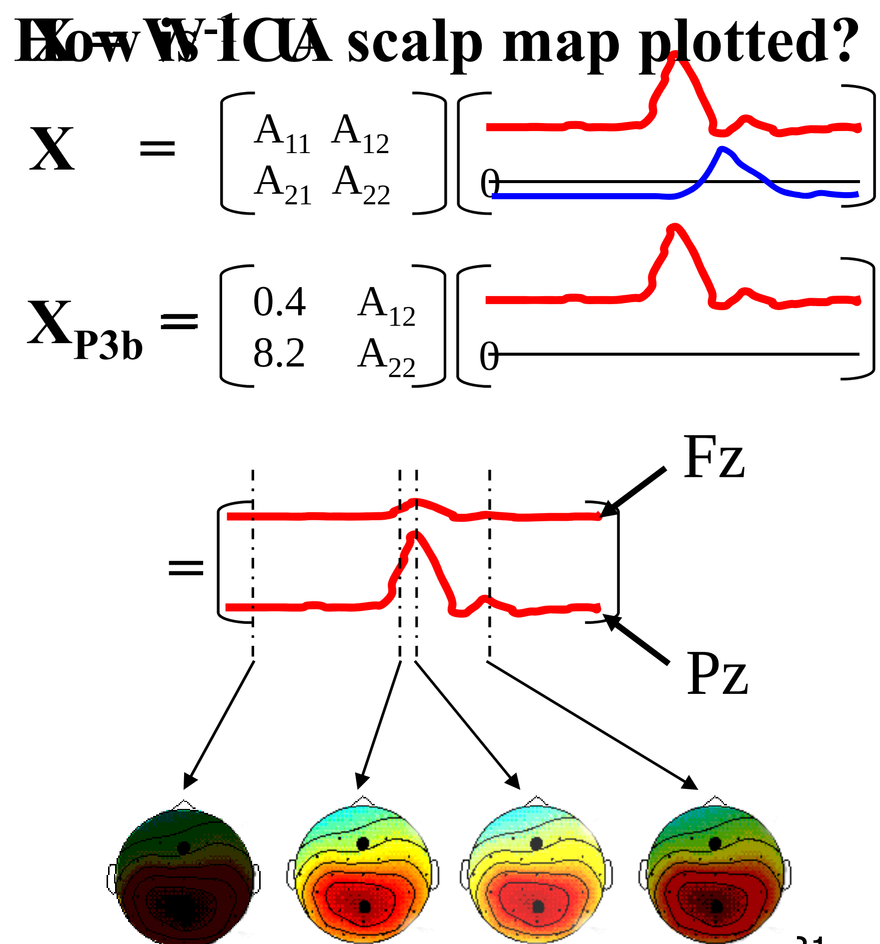
ICA/EEG Assumptions

- Mixing is linear at electrodes
- Propagation delays are negligible
- Component time courses are independent
- Number of components $\leq$ number of channels.

Independent components of EEG/ERP



From Makeig et al., *JNS,* 1999.



From Jung & Makeig, 2009.

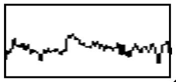
Frequently Asked Questions

- What is temporal and spatial ICA?

For EEG, we are looking at temporally independent brain activities arising from different brain networks.

For fMRI, the independence is considered over voxels because of brain modularity. i.e., Simplistically, "Different places do different things."

Temporal ICA



ICA activity U

$$\begin{array}{cl} \begin{bmatrix} u_{1,1} & u_{1,2} & u_{1,3} & u_{1,4} & \\ u_{2,1} & u_{2,2} & u_{2,3} & u_{2,4} & \cdots \\ u_{3,1} & u_{3,2} & u_{3,3} & u_{3,4} & \end{bmatrix} & \begin{array}{l} \leftarrow \text{Comp. 1} \\ \leftarrow \text{Comp. 2} \\ \leftarrow \text{Comp. 3} \end{array} \end{array}$$

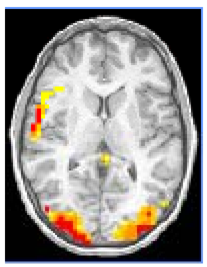
$$\mathbf{X}=\mathbf{W}^{-1}\mathbf{U}$$

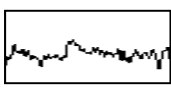
$$\begin{bmatrix} w_{1,1}^{-1} & w_{1,2}^{-1} & w_{1,3}^{-1} \\ w_{2,1}^{-1} & w_{2,2}^{-1} & w_{2,3}^{-1} \\ w_{3,1}^{-1} & w_{3,2}^{-1} & w_{3,3}^{-1} \end{bmatrix} \longrightarrow \begin{array}{cl} \begin{bmatrix} u_{1,1}w_{1,1}^{-1}+u_{2,1}w_{1,2}^{-1}+u_{3,1}w_{1,3}^{-1} & u_{1,2}w_{1,1}^{-1}+u_{2,2}w_{1,2}^{-1}+u_{3,2}w_{1,3}^{-1} & \\ u_{1,1}w_{2,1}^{-1}+u_{2,1}w_{2,2}^{-1}+u_{3,1}w_{2,3}^{-1} & u_{1,2}w_{2,1}^{-1}+u_{2,2}w_{2,2}^{-1}+u_{3,2}w_{2,3}^{-1} & \cdots \\ u_{1,1}w_{3,1}^{-1}+u_{2,1}w_{3,2}^{-1}+u_{3,1}w_{3,3}^{-1} & u_{1,2}w_{3,1}^{-1}+u_{2,2}w_{3,2}^{-1}+u_{3,2}w_{3,3}^{-1} & \end{bmatrix} & \begin{array}{l} \leftarrow \text{Chan 1} \\ \leftarrow \text{Chan 2} \\ \leftarrow \text{Chan 3} \end{array} \end{array}$$

Inverse weight matrix W^{-1}

Data X (EEG/MEG time series)

Spatial ICA





ICA activity U

$$\begin{array}{cl} \begin{bmatrix} u_{1,1} & u_{1,2} & u_{1,3} & u_{1,4} & \\ u_{2,1} & u_{2,2} & u_{2,3} & u_{2,4} & \cdots \\ u_{3,1} & u_{3,2} & u_{3,3} & u_{3,4} & \end{bmatrix} & \begin{array}{l} \leftarrow \text{Comp. 1} \\ \leftarrow \text{Comp. 2} \\ \leftarrow \text{Comp. 3} \end{array} \end{array}$$

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

$$\begin{bmatrix} w_{1,1}^{-1} & w_{1,2}^{-1} & w_{1,3}^{-1} \\ w_{2,1}^{-1} & w_{2,2}^{-1} & w_{2,3}^{-1} \\ w_{3,1}^{-1} & w_{3,2}^{-1} & w_{3,3}^{-1} \end{bmatrix} \longrightarrow \begin{array}{cl} \begin{bmatrix} u_{1,1}w_{1,1}^{-1}+u_{2,1}w_{1,2}^{-1}+u_{3,1}w_{1,3}^{-1} & u_{1,2}w_{1,1}^{-1}+u_{2,2}w_{1,2}^{-1}+u_{3,2}w_{1,3}^{-1} & \\ u_{1,1}w_{2,1}^{-1}+u_{2,1}w_{2,2}^{-1}+u_{3,1}w_{2,3}^{-1} & u_{1,2}w_{2,1}^{-1}+u_{2,2}w_{2,2}^{-1}+u_{3,2}w_{2,3}^{-1} & \cdots \\ u_{1,1}w_{3,1}^{-1}+u_{2,1}w_{3,2}^{-1}+u_{3,1}w_{3,3}^{-1} & u_{1,2}w_{3,1}^{-1}+u_{2,2}w_{3,2}^{-1}+u_{3,2}w_{3,3}^{-1} & \end{bmatrix} & \begin{array}{l} \leftarrow \text{Time 1} \\ \leftarrow \text{Time 2} \\ \leftarrow \text{Time 3} \end{array} \end{array}$$

Inverse weight matrix W^{-1}

Data X (EEG/MEG voxel activities)

Frequently Asked Questions (cont.)

- How much data is enough data?

  There is no fixed limit to the number of points needed for a "good" ICA solution - and in fact no fixed way to judge whether an ICA solution is "good" or not.

Frequently Asked Questions (cont.)

- Pre-ICA procedures
  - Check the rank of the data (if not full rank, use PCA)
  - ‘Messy' channels or epochs should be removed
  - Ultra-low frequency activity should be removed, including the DC offset (a.k.a. remove baseline')
- Check ICA solution prior to further analysis
  - Review component scalp maps and check their 'dipolarity'
  - If component maps are 'messy', remove ‘messy’ epochs/channels and try again...

Frequently Asked Questions (cont.)

- How should the activations be scaled?

$$U=WX, \quad X=W^{-1}*U$$

The strength of source activity is distributed between the columns of W^{-1} and the rows of U.

- Ordering of ICs

Not well-defined and intuitive.

- Can ICA separate 'correlated' source activities?

Practical Issues with ICA of EEG/ERP

1. Apply ICA to averaged ERPs
   - How many time points are needed for training?

     Suggestion: At least several times number of variables in the unmixing matrix.
   - Which EEG processes may express their independence in the ERP training data?

     Suggestion: Decompose the concatenated collection of ERP averages in respond to the experimental stimulus and task conditions.
   - ICA decomposition of averaged ERPs must be interpreted with caution.

Practical Issues with ICA of EEG/ERP

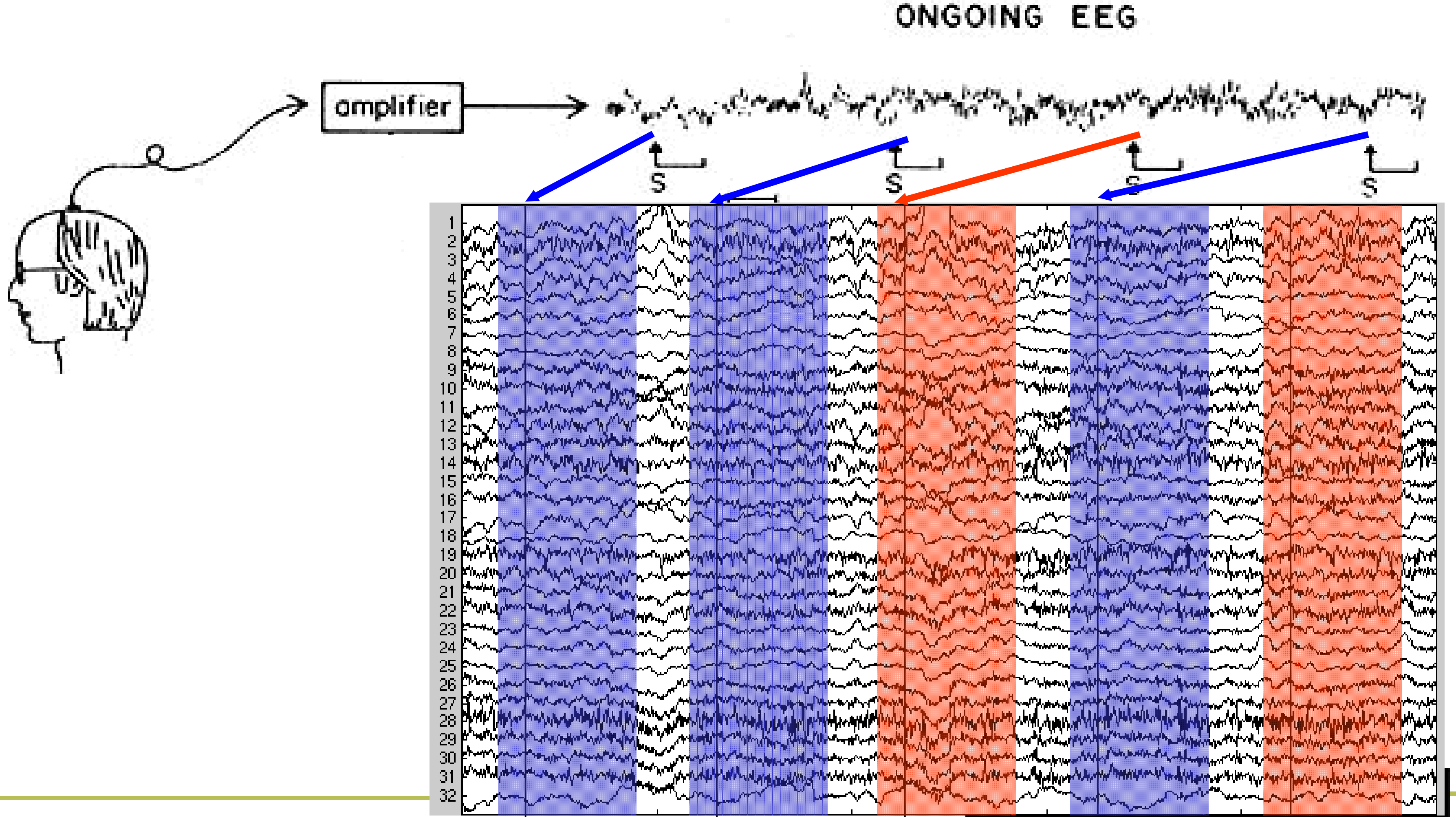
2. Apply ICA to continuous EEG data
   - Are components spatially stationary through time?

     Suggestion: Perform separate decompositions of subsets of the recorded data, each consisting of periods during which the sources may be stationary.

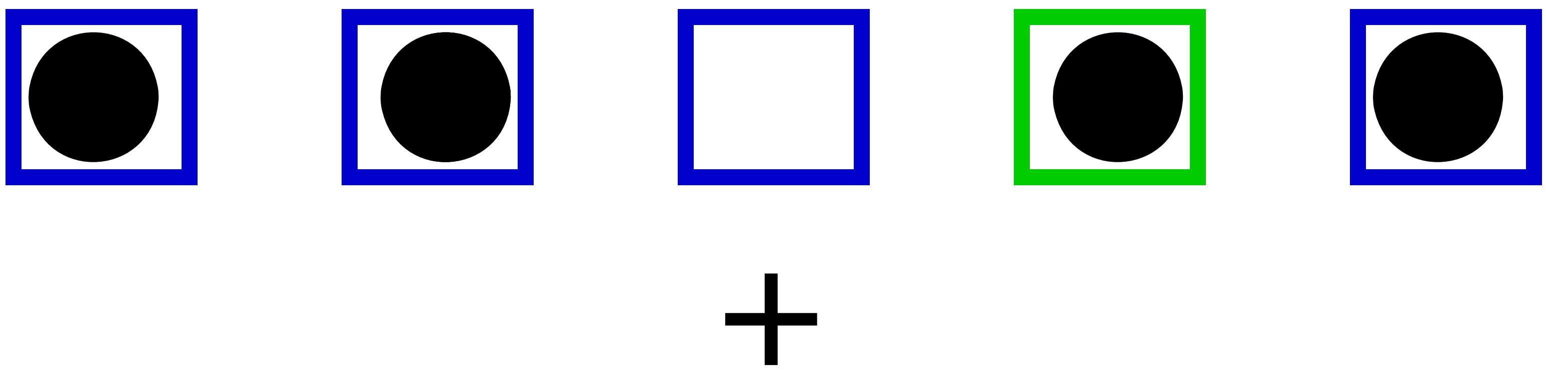
     Or, you can use a mixture of ICA model.

Practical Issues with ICA of EEG/ERP

3. Apply ICA to unaveraged event-related EEG



Experiment

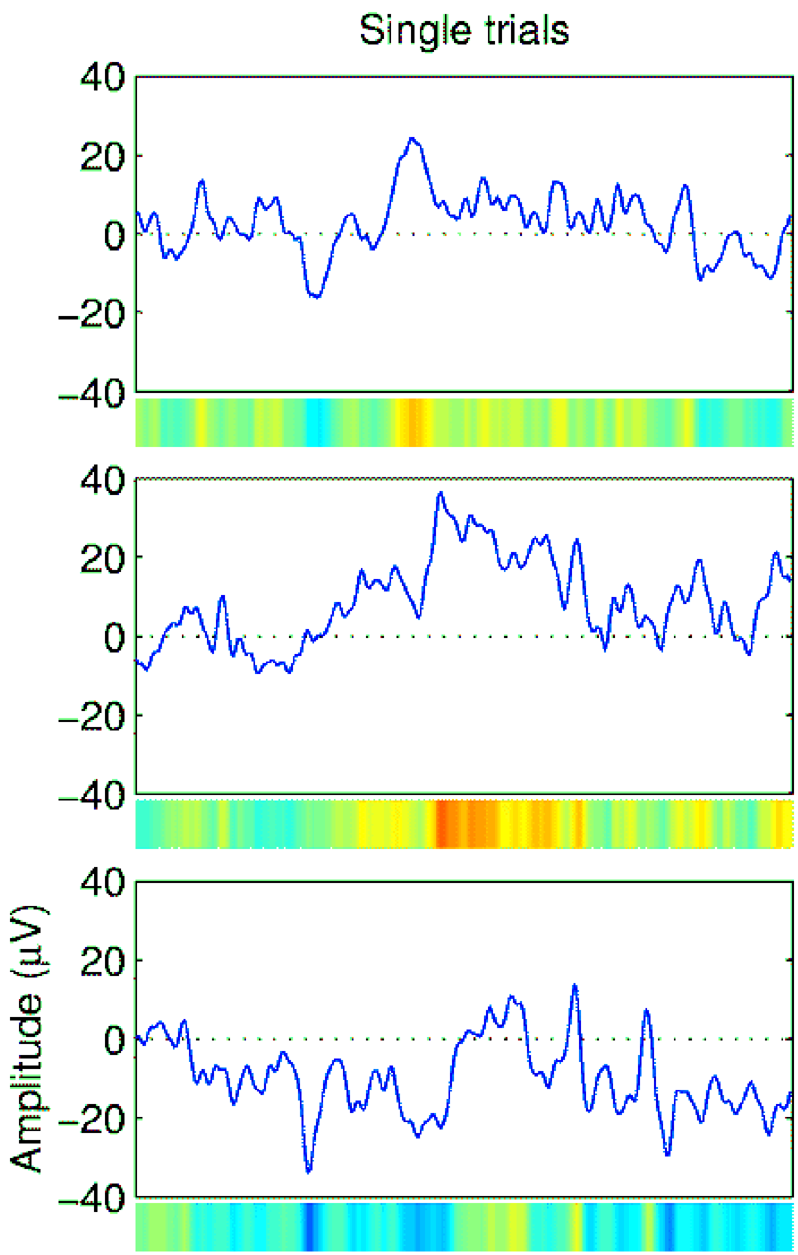


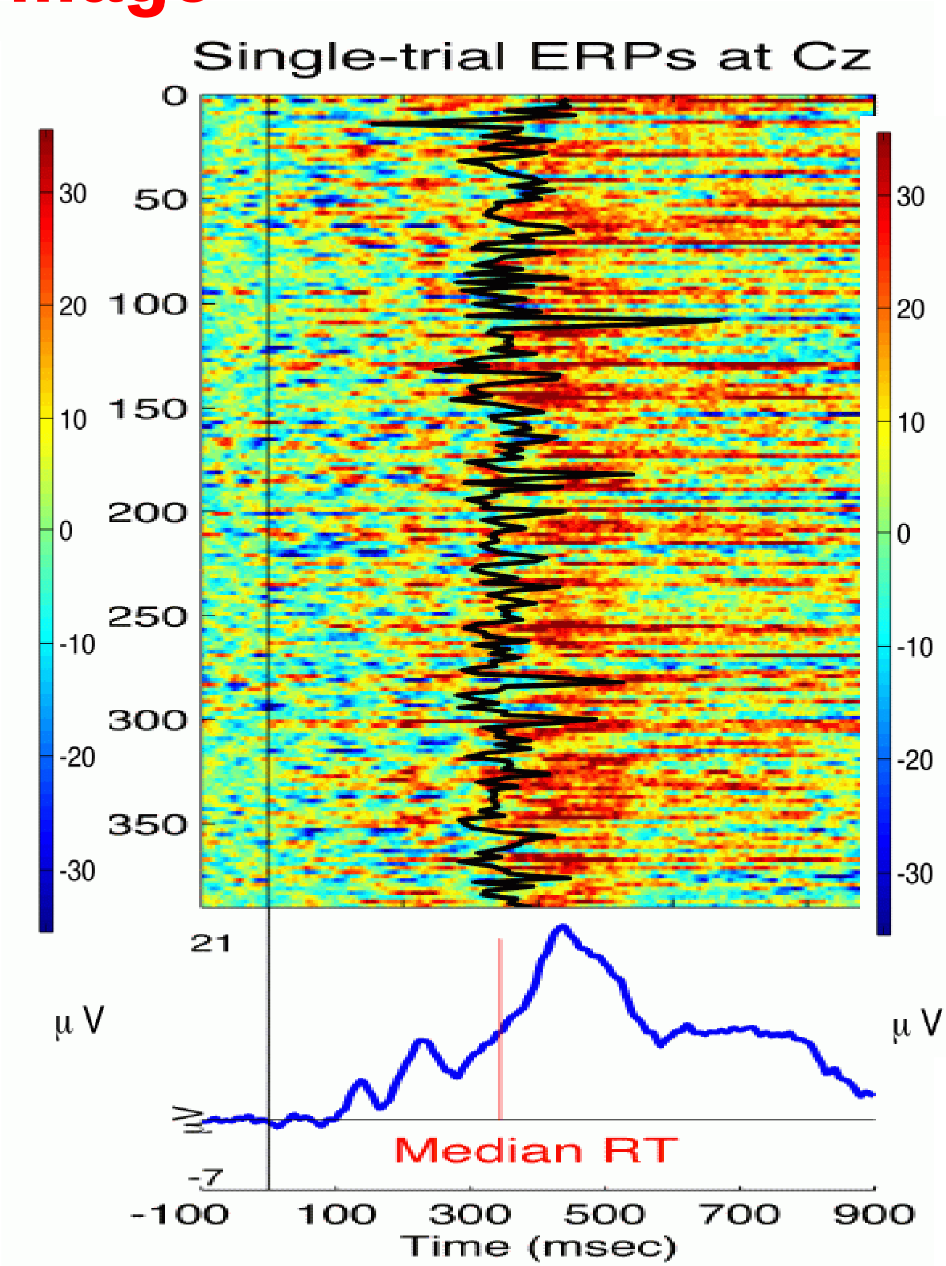
Task: Fixate **cross** while covertly attending to **green box**. Press button when **circle** is flashed in green box.

Subject: 28 normal control, 14 autistic and 8 cerebellar lesion subjects.

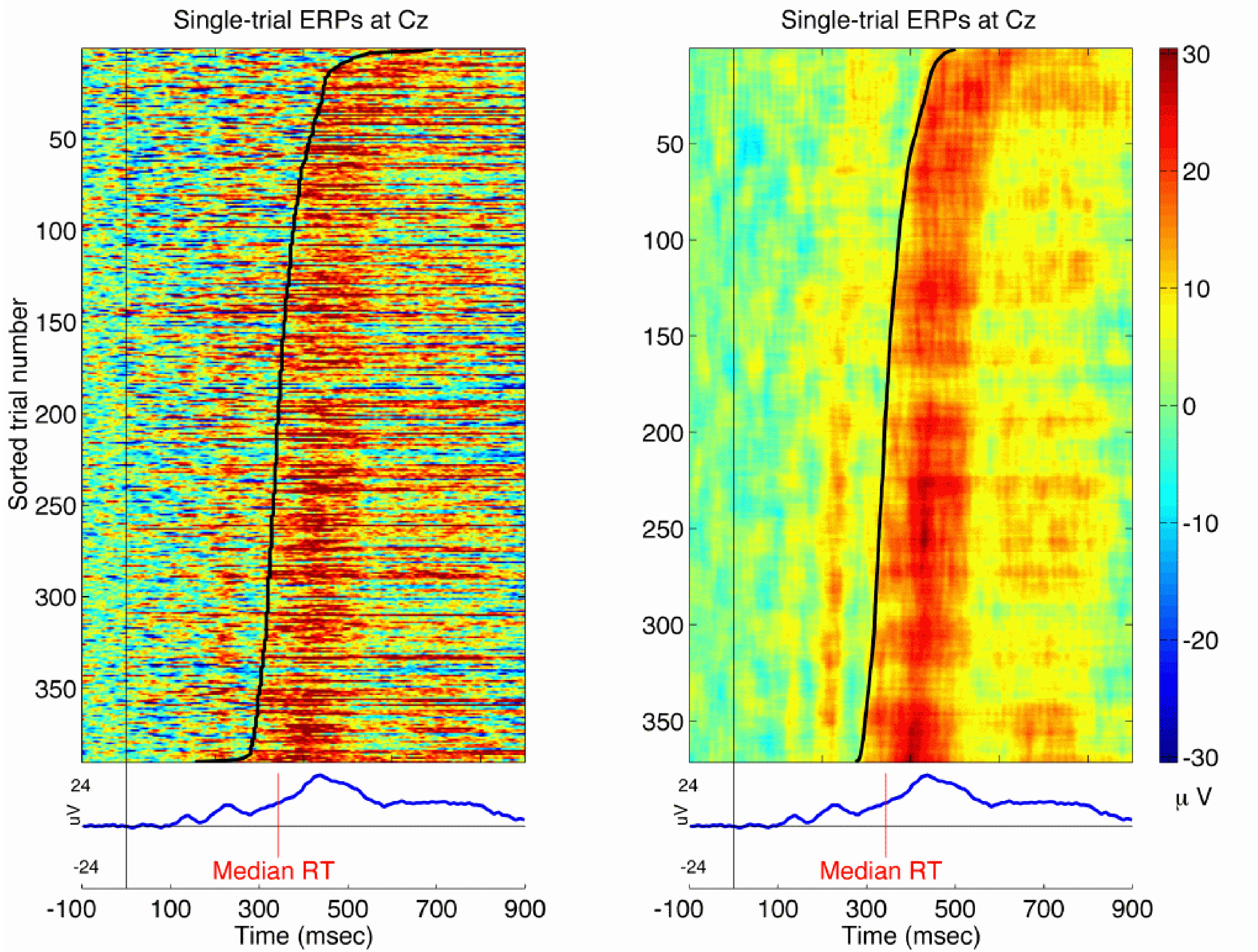
Session: 30 72-s task blocks, including 120 **targets** and 480 nontargets in each of the 5 locations.

ERP Image





From Jung et al., *NIPS*, 1999.



From Jung et al., *NIPS*, 1999.

Analysis of Single-trial ERPs

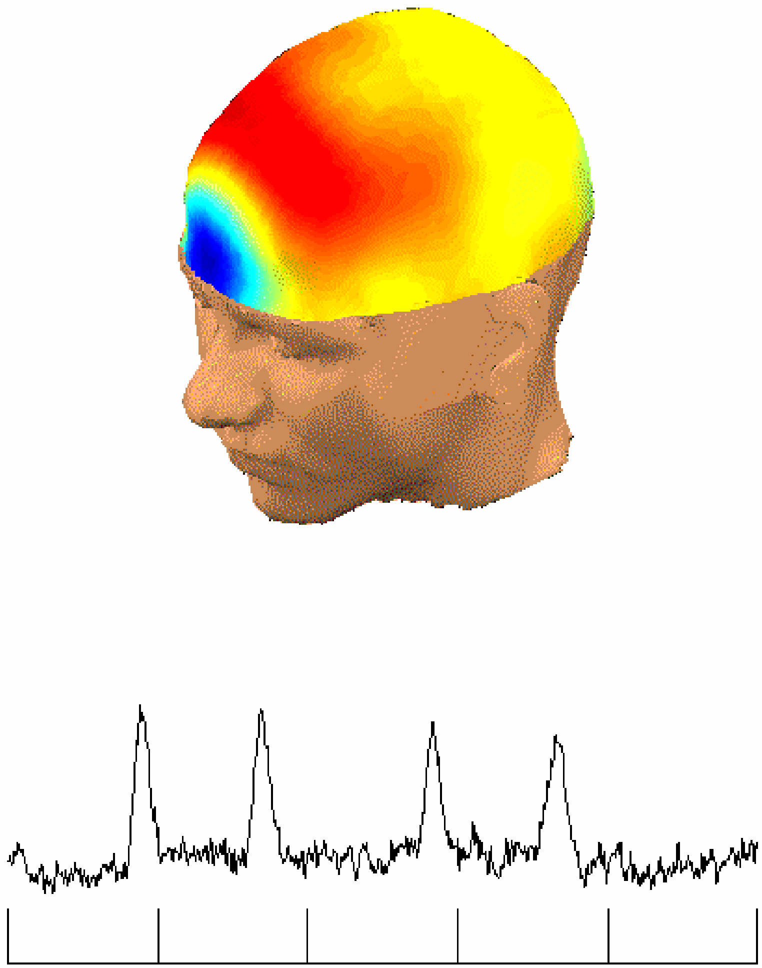
ICA applied to ~600 (single-subject, 31-channel, 1-s) concatenated single-trial response epochs timelocked to detected **target stimuli**

↓

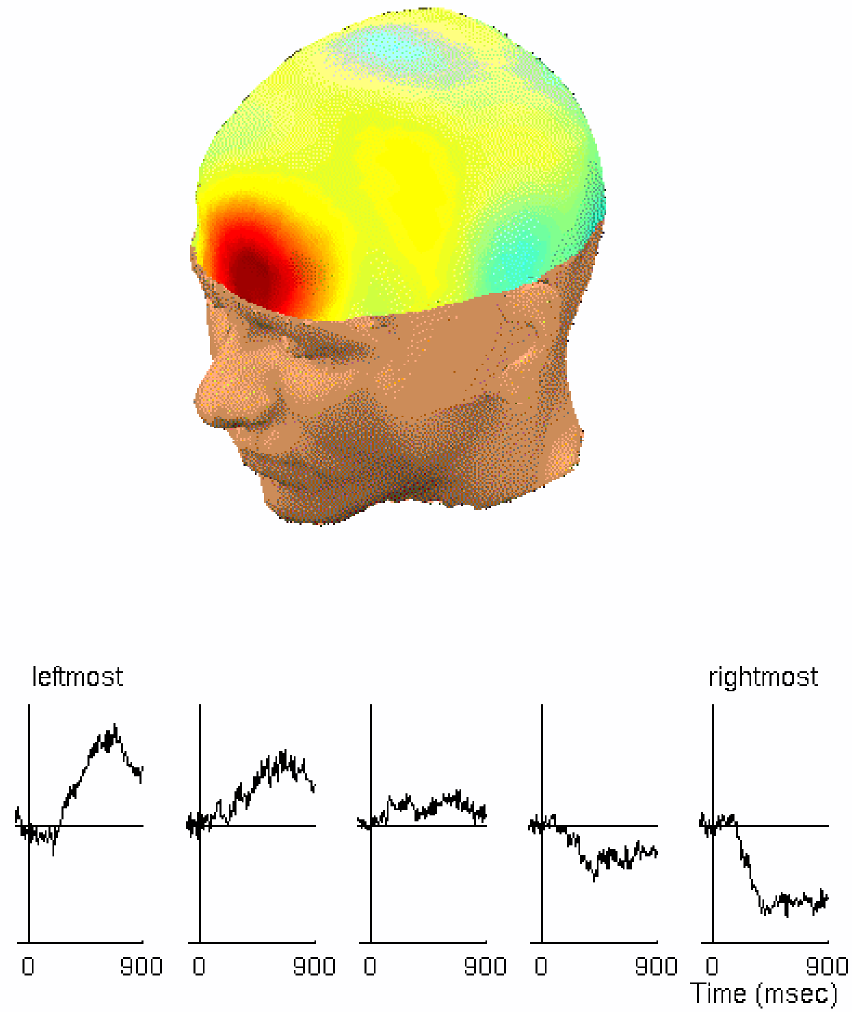
31 independent components having:

- fixed spatial projections to the scalp
- temporally independent time courses of activation

Component 1

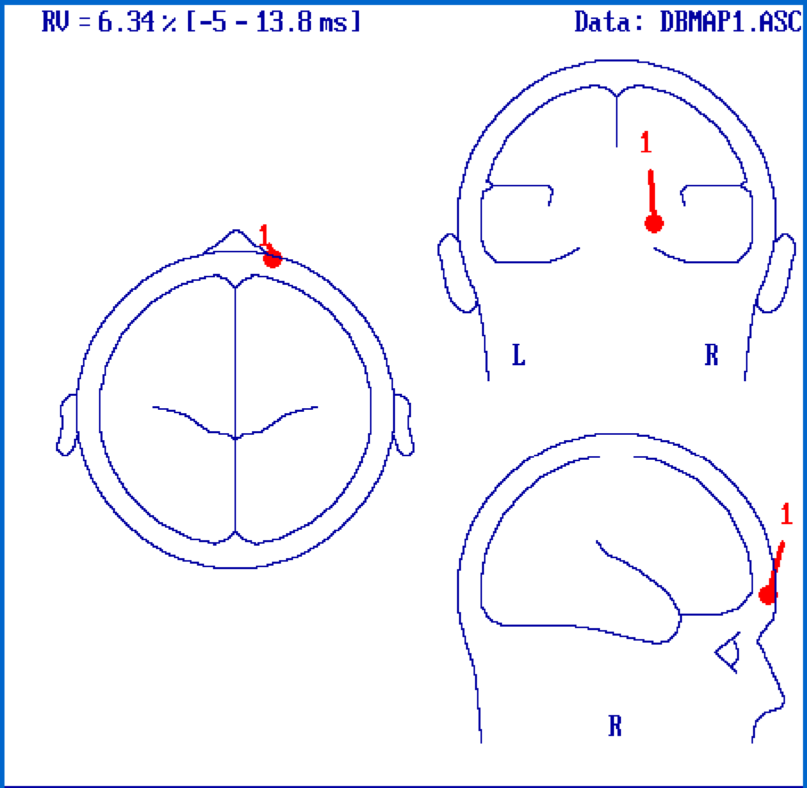


Component 2

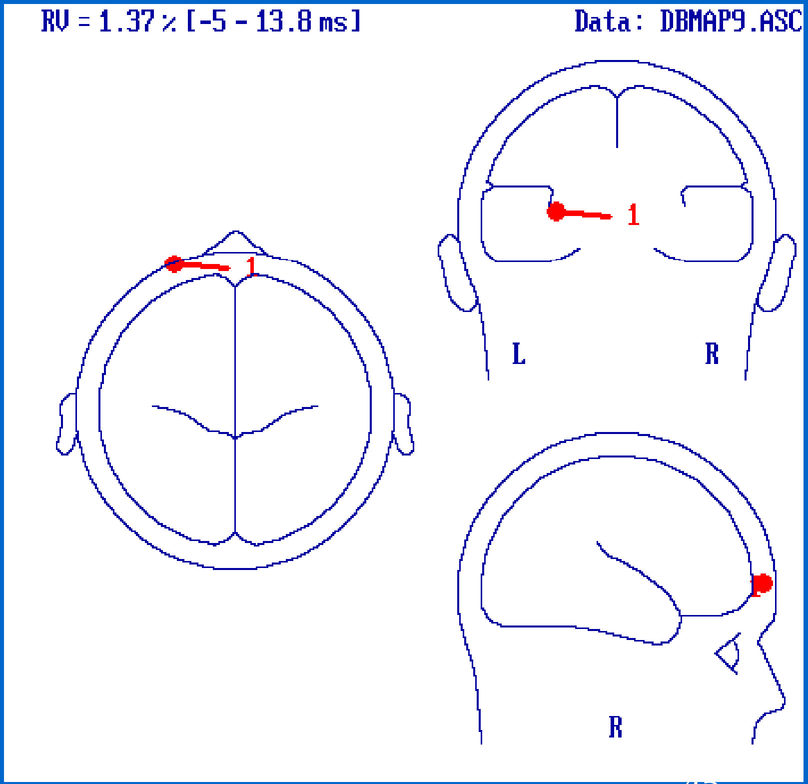


Single-dipole BESA Modeling

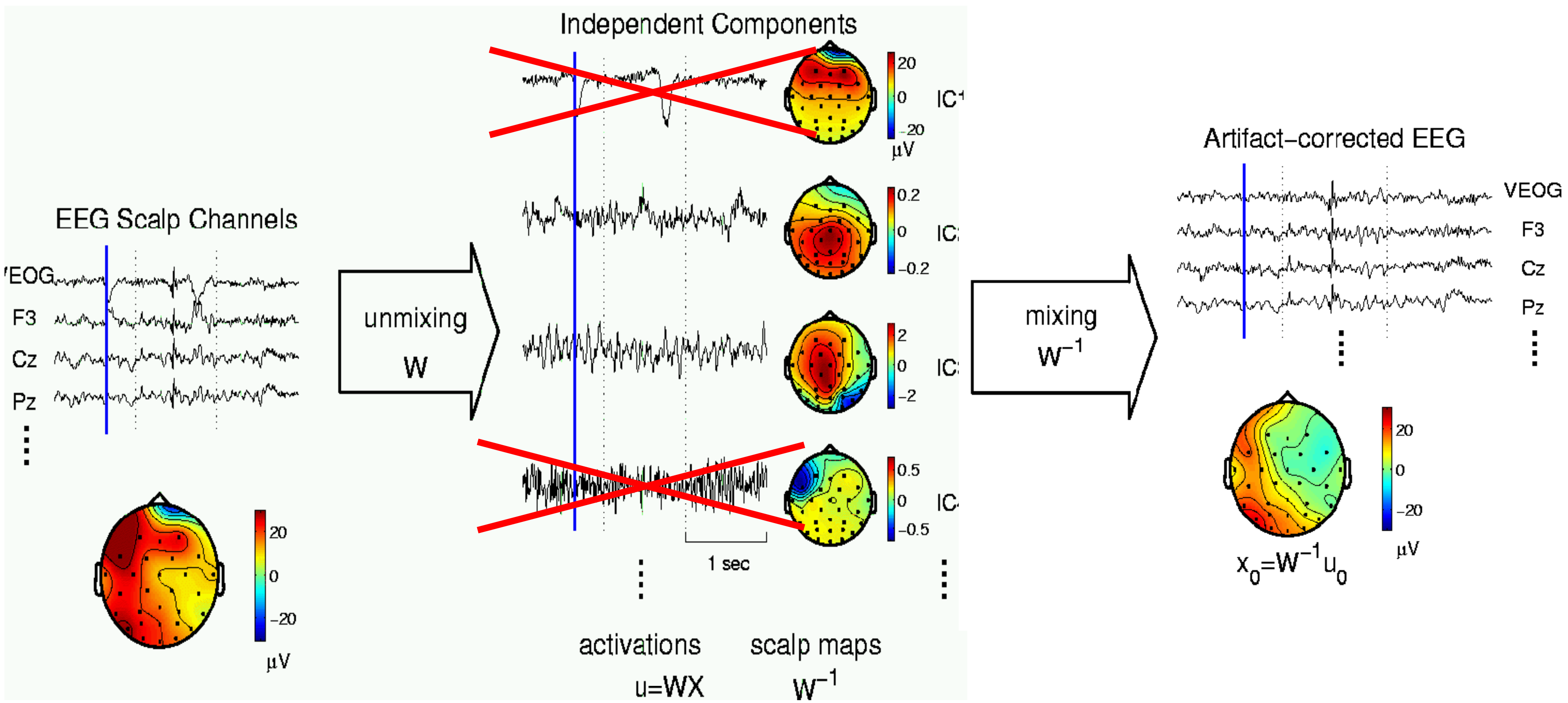
Component 1



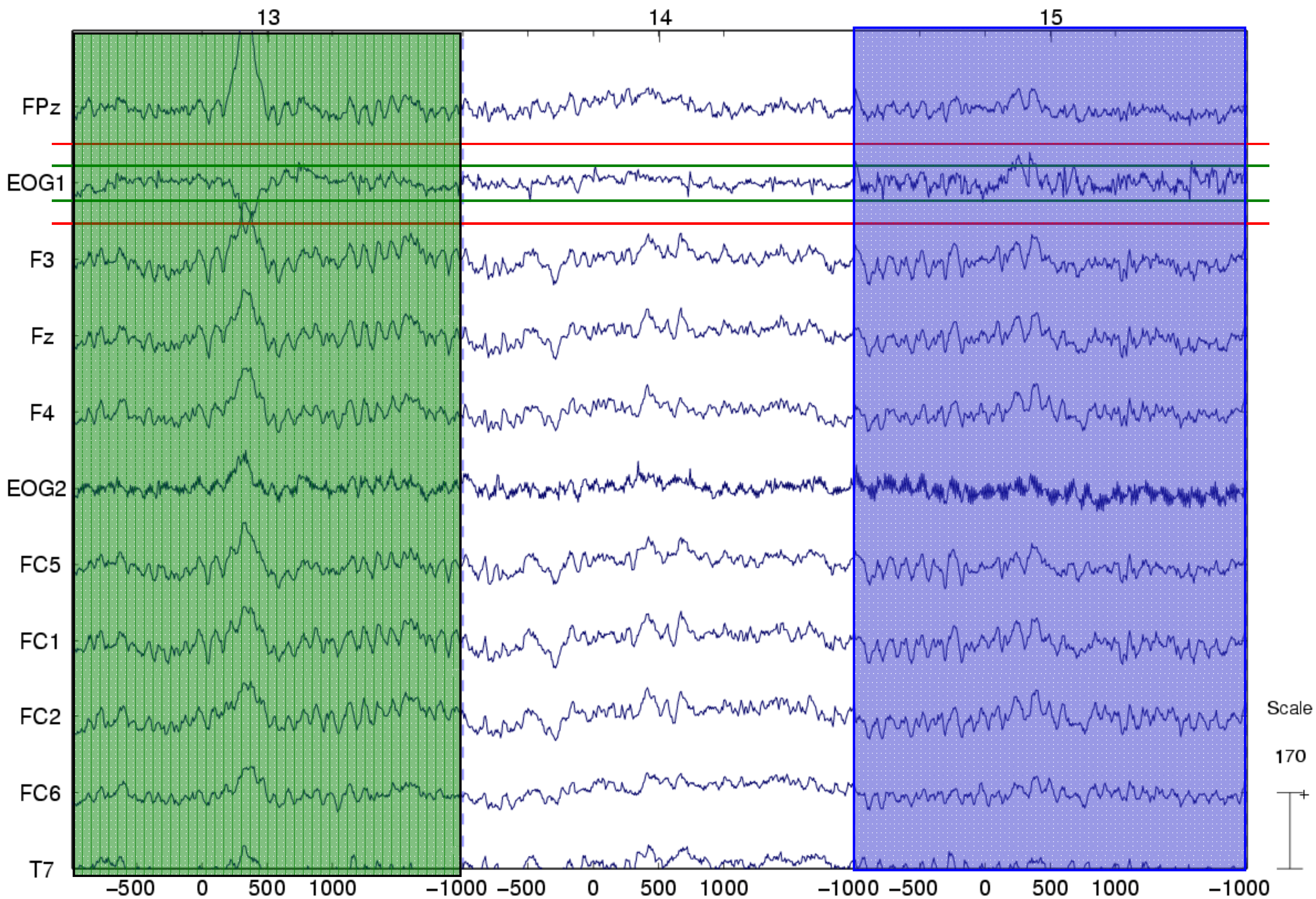
Component 2



ICA-based Artifact Correction



Split Single Trials based on EOG

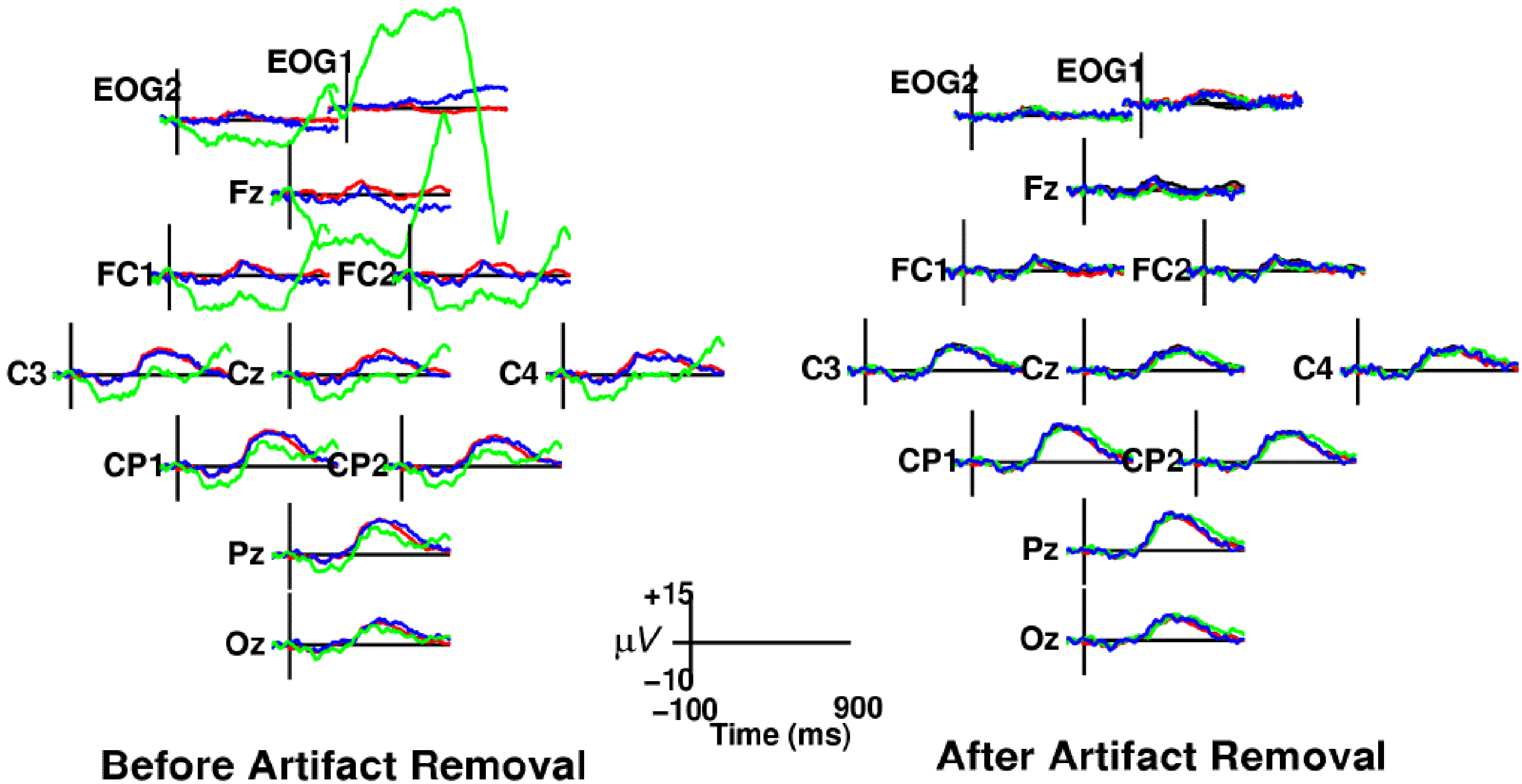


Heavily contaminated

Clean trials

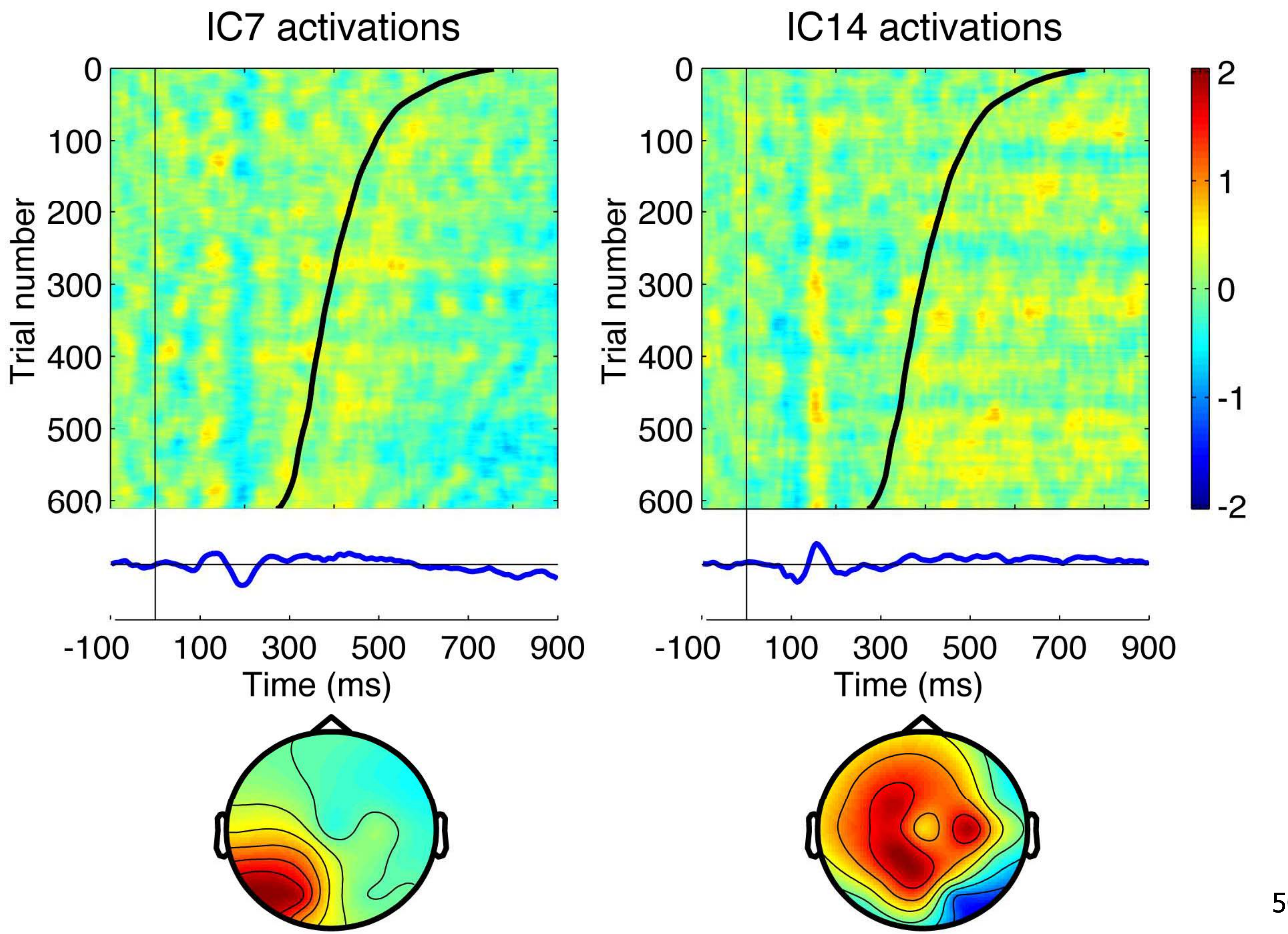
Contaminated

Averages of Least, Moderately and Heavily Contaminated Trials

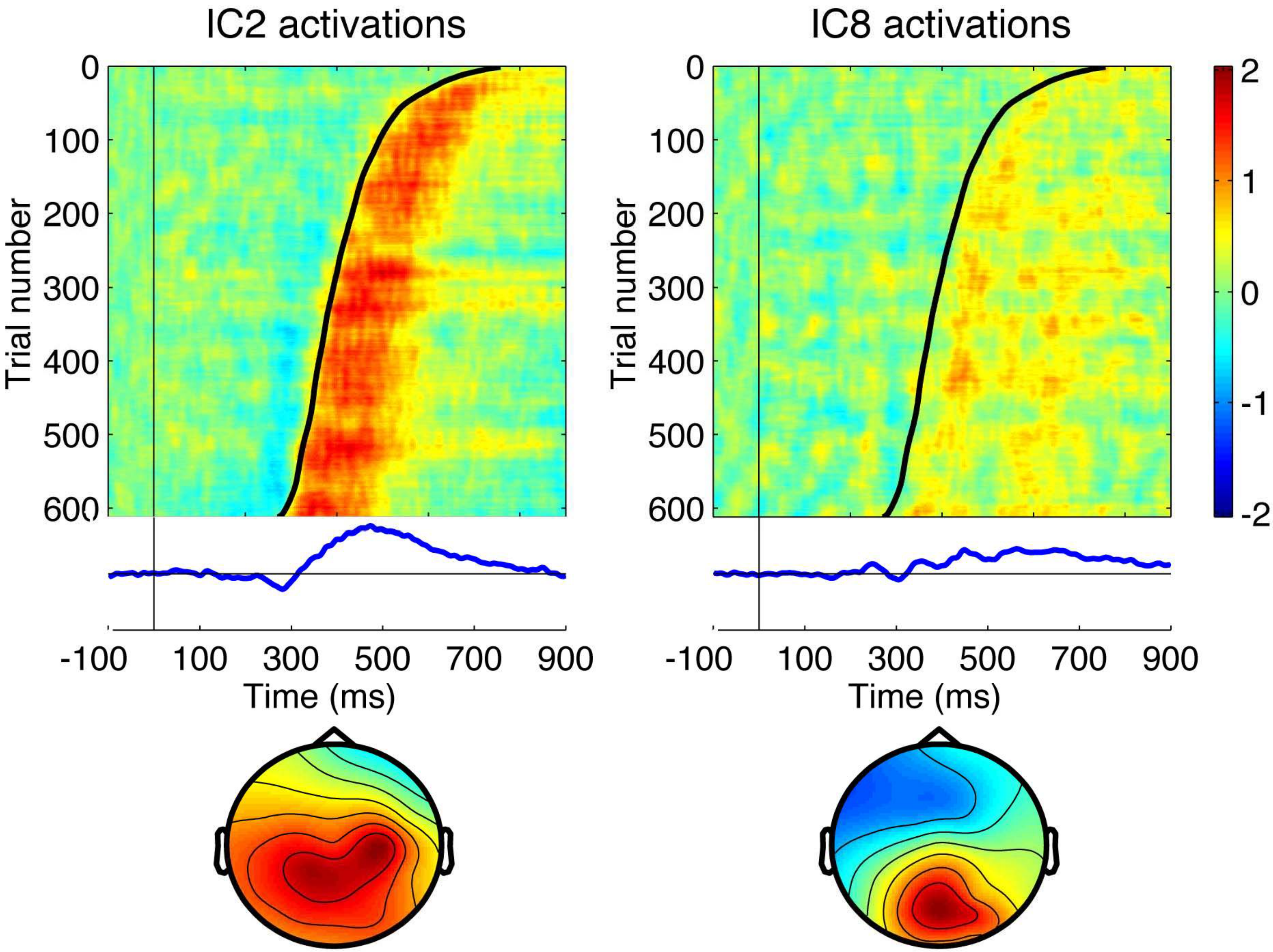


From Jung et al., *Clinical Neurophysiology*, 2000.

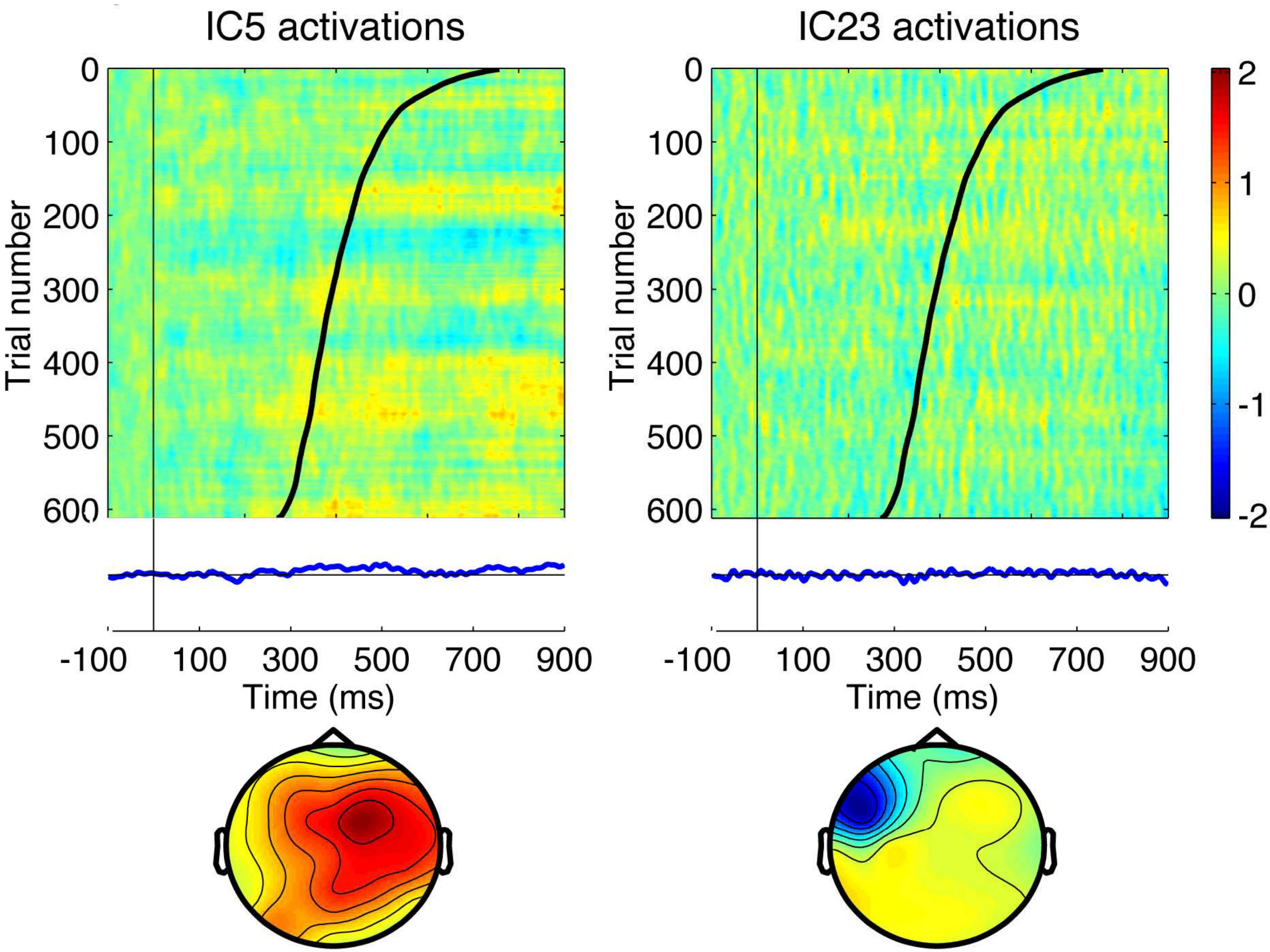
Stimulus-locked

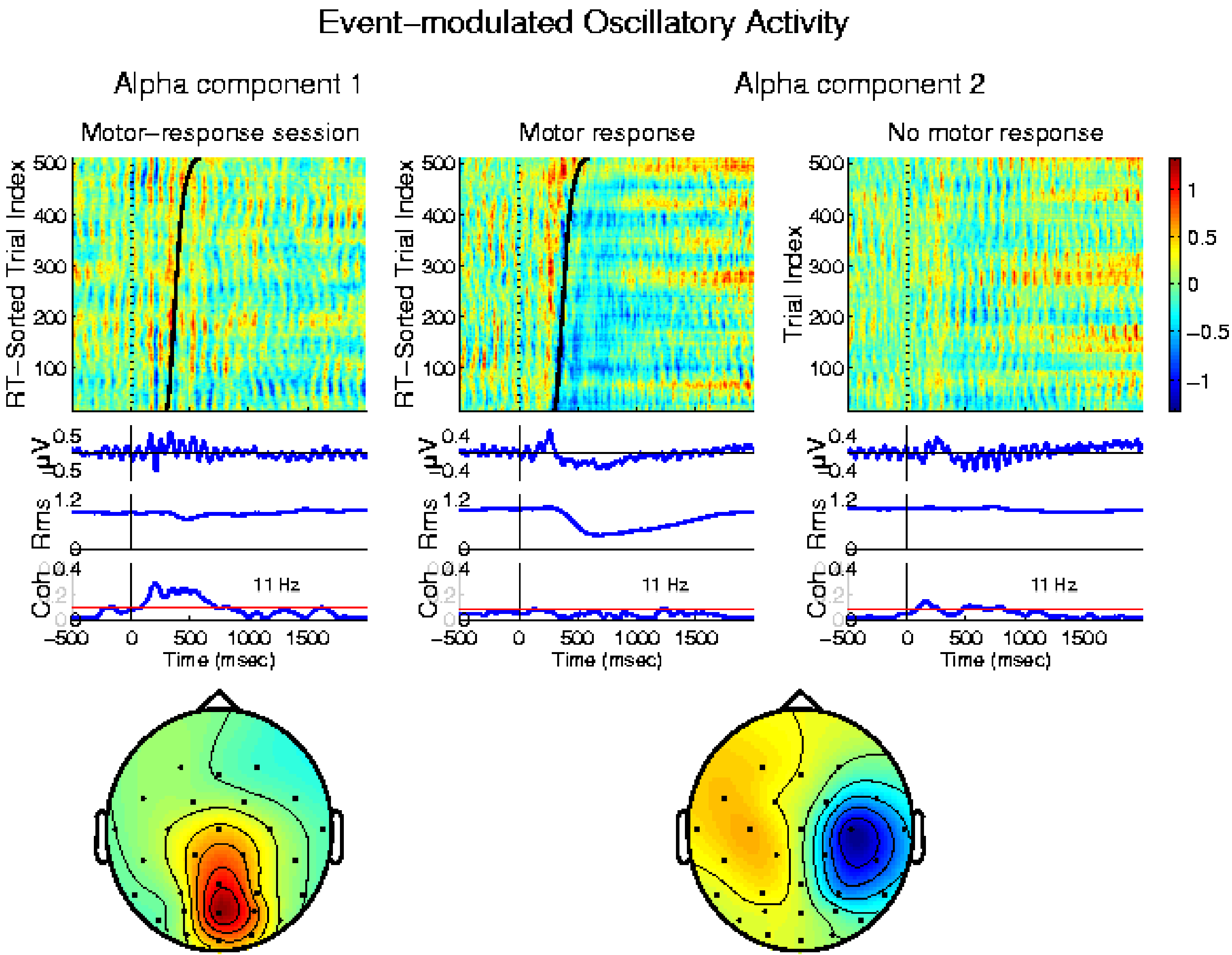


Response-locked



Non-phase locked





Event–modulated Oscillatory Activity
Alpha component 1
Alpha component 2
Motor–response session
Motor response
No motor response
RT–Sorted Trial Index
Trial Index
500
400
300
200
100
1
0.5
0
–0.5
–1
μV
0.5
–0.5
0.4
–0.4
Rms
1.2
0
Coh
0.4
0.2
0
11 Hz
–500
0
500
1000
1500
Time (msec)

Characteristics of Independent Components

- Concurrent Activity
- Maximally Temporally Independent
- Overlapping Maps and Spectra
- Dipolar Scalp Maps
- Functionally Independent
- Between-Subject Regularity