Independent Component Analysis (ICA) of EEG, Concepts and Methods Part 2 – Methods

Jason Palmer

Swartz Center for Computational Neuroscience
Institute for Neural Computation
University of California San Diego, La Jolla CA

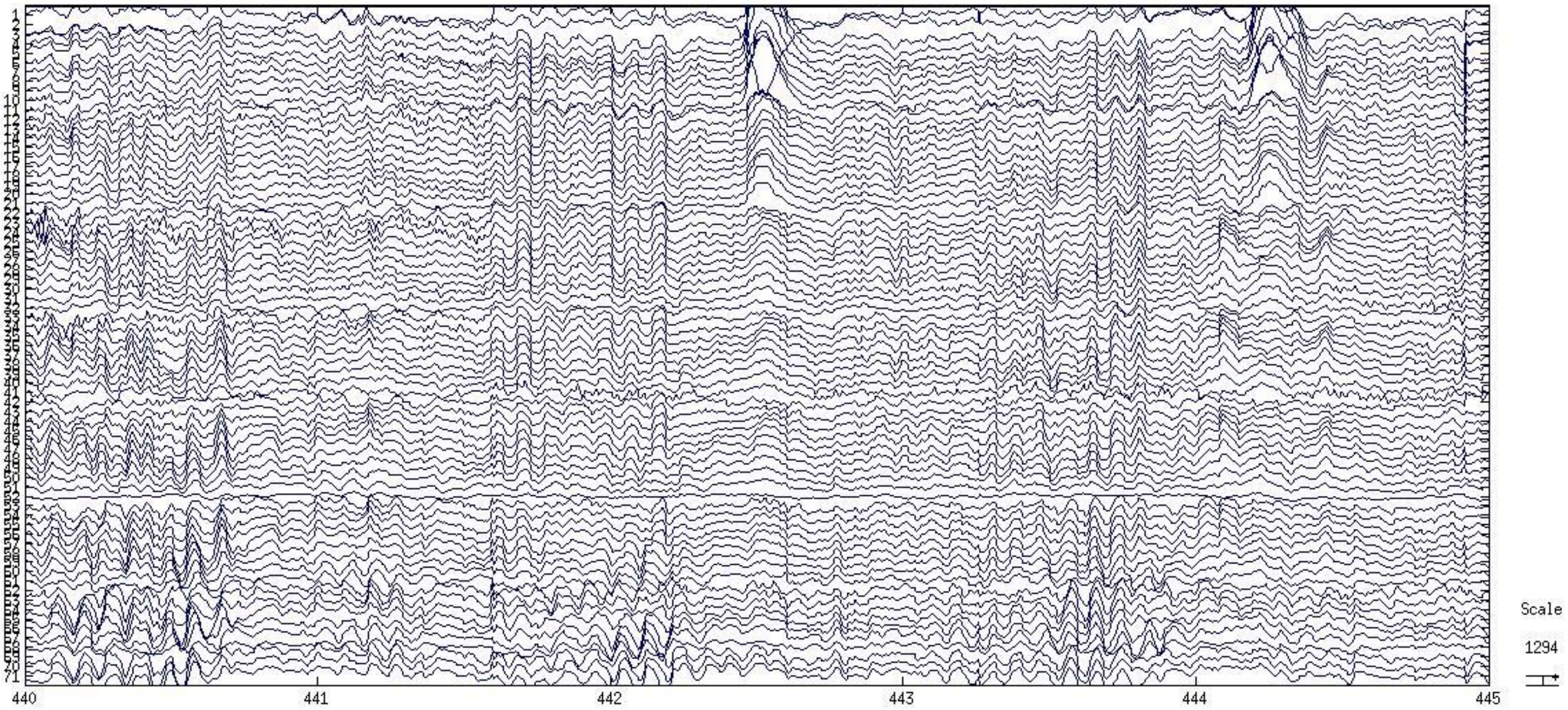
November 18, 2010

Outline

- Decorrelation – PCA and Sphering
- Statistical dependence
- ICA strategies
- Maximum likelihood and minimum mutual information
  - Modeling source densities
- Multiple models and non-stationarity
  - ICA mixture model

EEG Data

- Raw EEG records large number of simultaneously active sources
- From physics, we know that EEG at one instant is simply the sum of all source activity at that instant



Decorrelation

- Our first thought is decorrelation, i.e. find **A** and **S** such that the rows of **S** are orthogonal
- Unfortunately decorrelation is not unique, there are an infinite number of such **A**, **S** pairs
- One example is PCA, which projects onto eigenvectors of covariance matrix:

$$\mathbf{X}\mathbf{X}^{\mathrm{T}} / N = \mathbf{U}\mathbf{D}\mathbf{U}^{\mathrm{T}}$$

where the columns of **U** are the eigenvectors

Linear superposition model

- Basic linear model:

$$n \times N \longrightarrow X = \overset{n \times n}{\overset{\swarrow}{A}} Y \longleftarrow n \times N$$

- Eigen-decomposition of covariance:

$$XX^T / N = UDU^T$$

- PCA decomposition:

$$X = (UD^{1/2})(D^{-1/2} U^T X) = A_{PCA} Y_{PCA}$$

- Sphering decomposition:

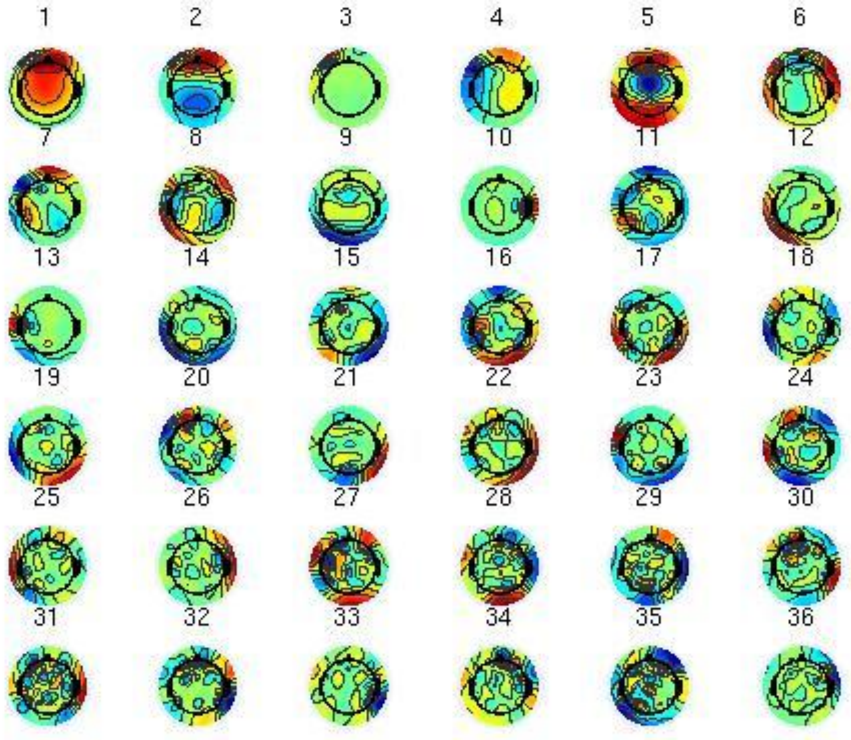
$$X = (UD^{1/2} U^T)(UD^{-1/2} U^T X) = A_{SPH} Y_{SPH}$$

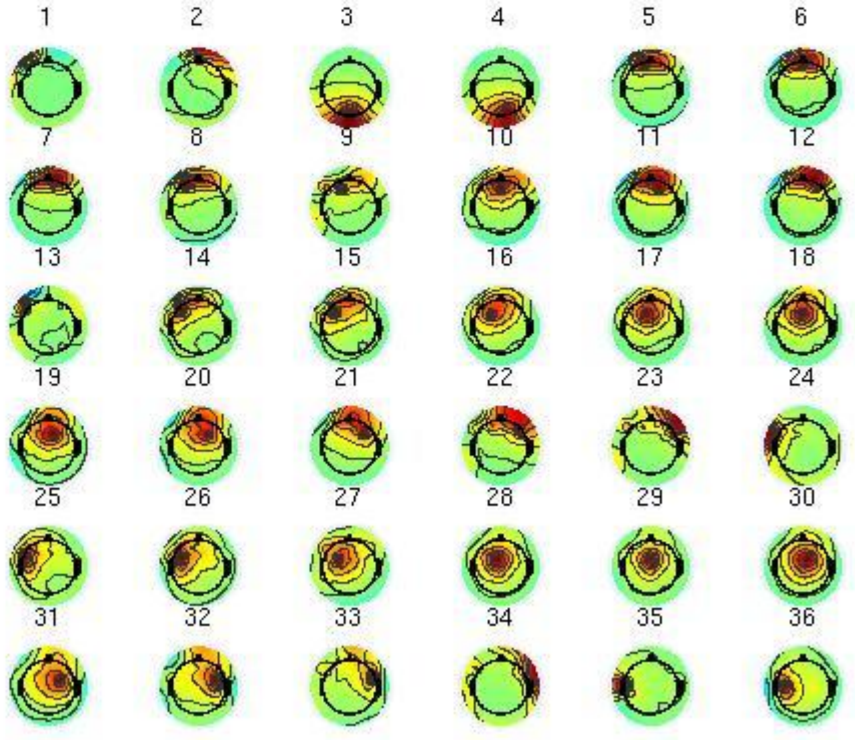
- ICA then decomposes $Y_{SPH} = A_{ICA} Y_{ICA}$ so that:

$$X = \underset{\text{icawinv}}{(A_{SPH} A_{ICA})}\, Y_{ICA}\,, \quad Y_{ICA} = (\underset{\text{icaweights}}{W_{ICA}} \underset{\text{icasphere}}{W_{SPH}})\, X$$

PCA and Sphering component maps

- PCA maps (left) are eigenvectors–orthogonal,unrealistic
- Sphering components (right) – all radial, localized





Independent Component Analysis

- Rather than try to reduce (or eliminate) correlation between sources, try to reduce statistical dependence

- Independence is defined mathematically by factorizability of the joint probability density:

$$\mathrm{p}_{\mathbf{s}}\big(s_1(t), s_1(t), \ldots, s_n(t)\big) = \mathrm{p}_1\big(s_1(t)\big) \cdot \mathrm{p}_2\big(\mathrm{s}_2(t)\big) \cdots \mathrm{p}_n\big(\mathrm{s}_n(t)\big)$$

- Mutual information is a measure of how much the joint density differs from the product of the marginal densities, specifically it is the Kullback-Leibler divergence of joint from product of marginals

Mutual Information and Maximum Likelihood Estimation

- To estimate the sources in the model **X**=**AS**, we look for an unmixing matrix **W** = **A**$^{-1}$ such that **Y**=**WX** where **Y** is a scaled rearrangement of the sources **S**

- Since the model is fairly simple (linear with independent sources, temporally i.i.d.) we can calculate the likelihood

$$p(\mathbf{X}) = \prod_t p_{\mathbf{x}}(\mathbf{x}(t)), \qquad \log p(\mathbf{X}) = \sum_t \log|\det \mathbf{W}| + \log p_{\mathbf{y}}(\mathbf{Wx}(t))$$

- It turns out that the likelihood is related to the mutual information

$$I(y_1; y_2; \ldots ; y_n) = -N^{-1} \log p(\mathbf{X}) + \sum_i KL\big(q(y_i)\,||\,p(y_i)\big)$$

- So we can minimize the mutual information by maximizing the likelihood over **W** and minimizing the divergence of the model source model densities $q_i(y_i)$ from the actual densities $p_i(y_i)$

What is a Mixture Model?

- A mixture model is a probabilistic combination of several models:

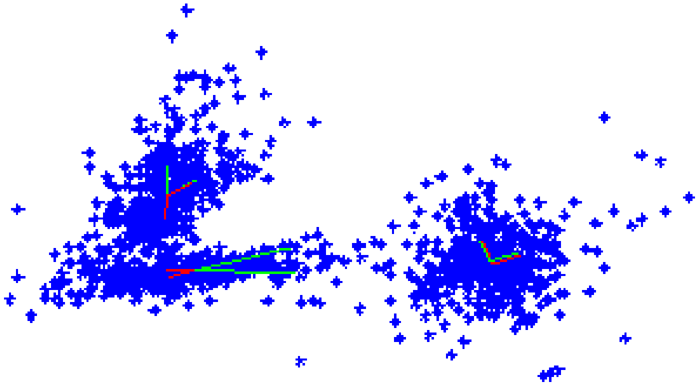
mixture proportions

means

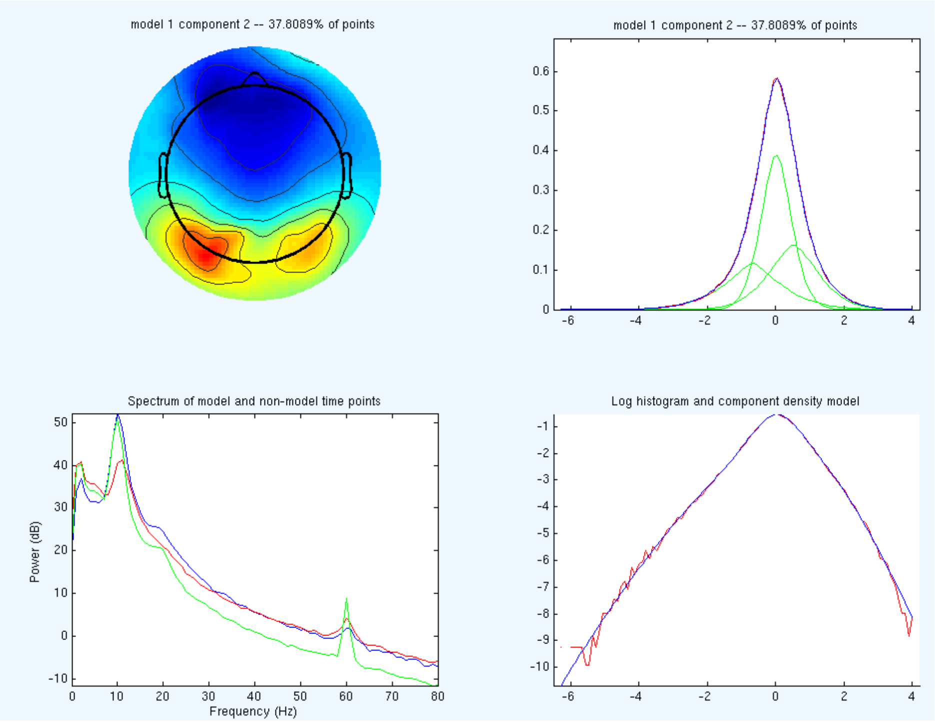
$$p(x) = \sum_{j=1}^{M} \gamma_j \, p_j\!\left(\frac{x-\mu_j}{\sigma_j}\right)$$

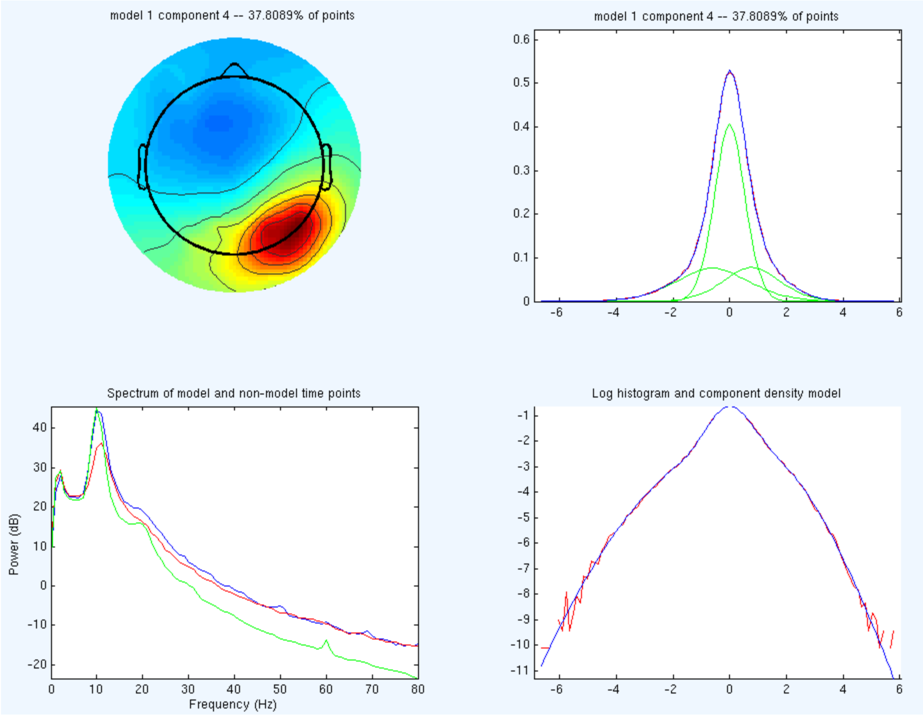
scales

- Each data point modeled as being generated by one of the models in the mixture

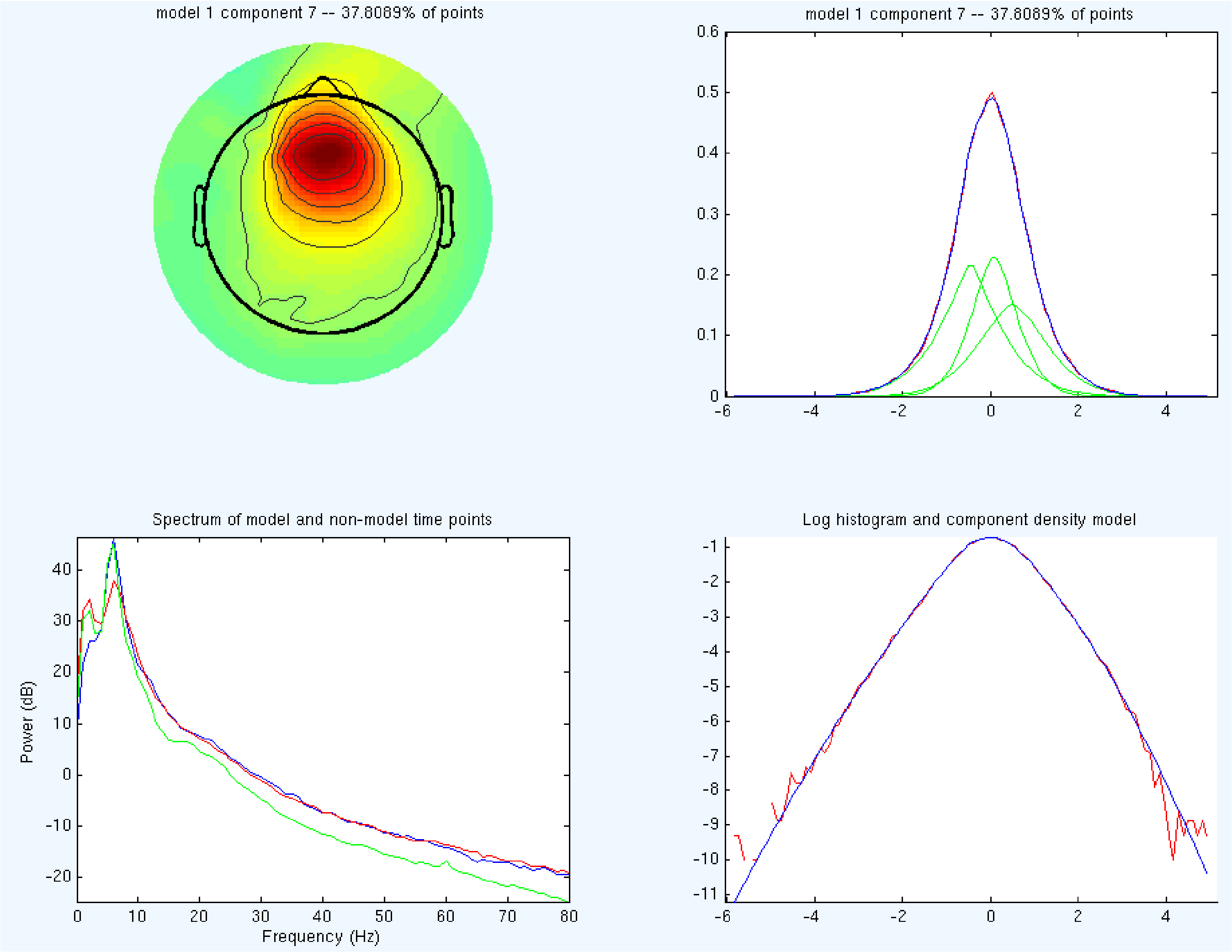


Alpha components



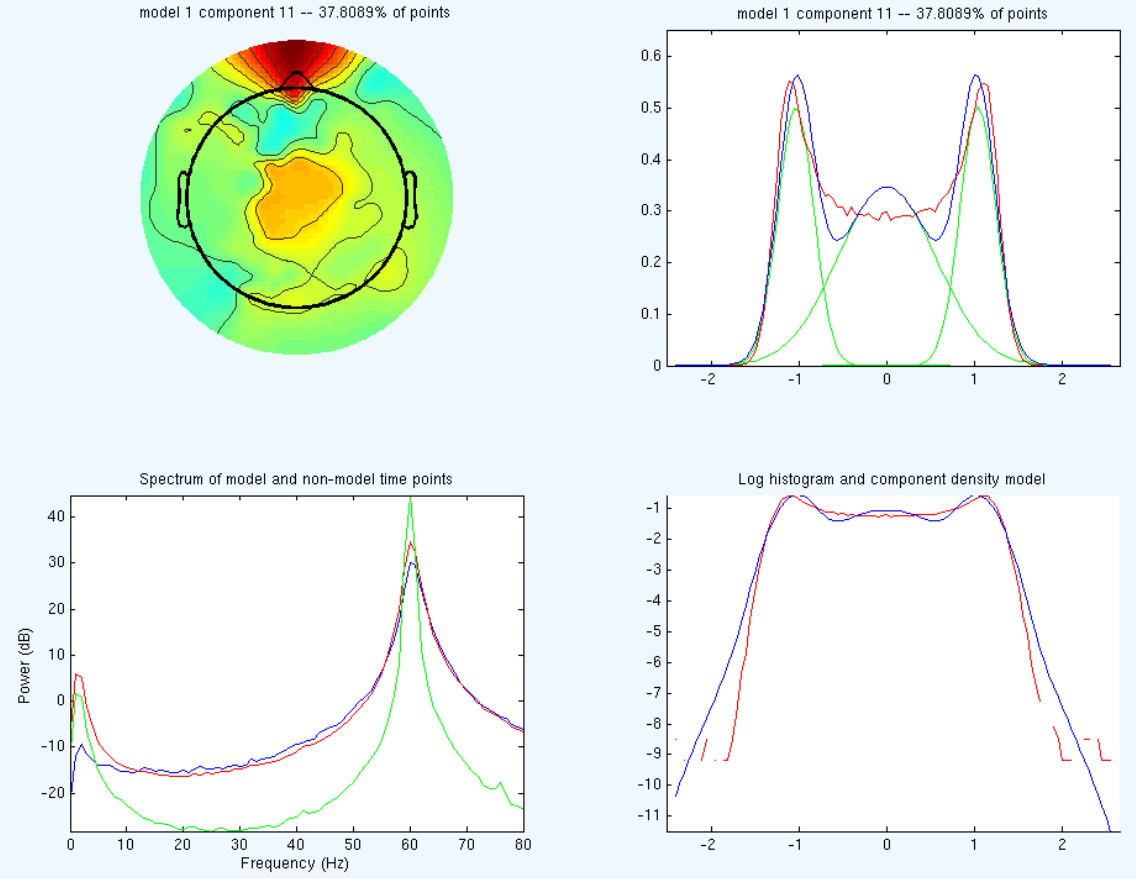


Frontal midline θ

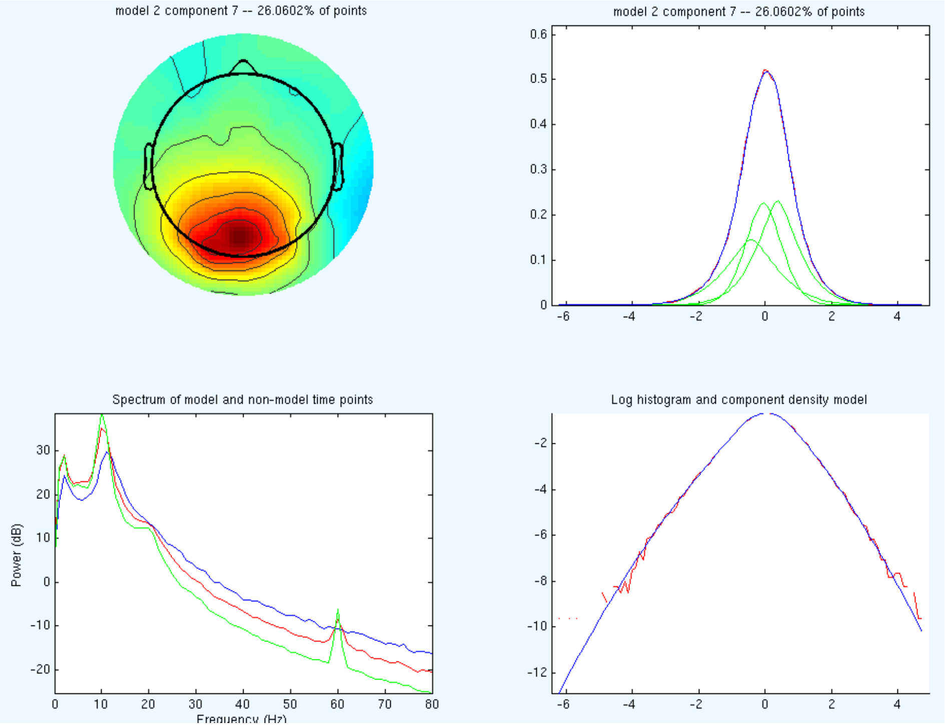


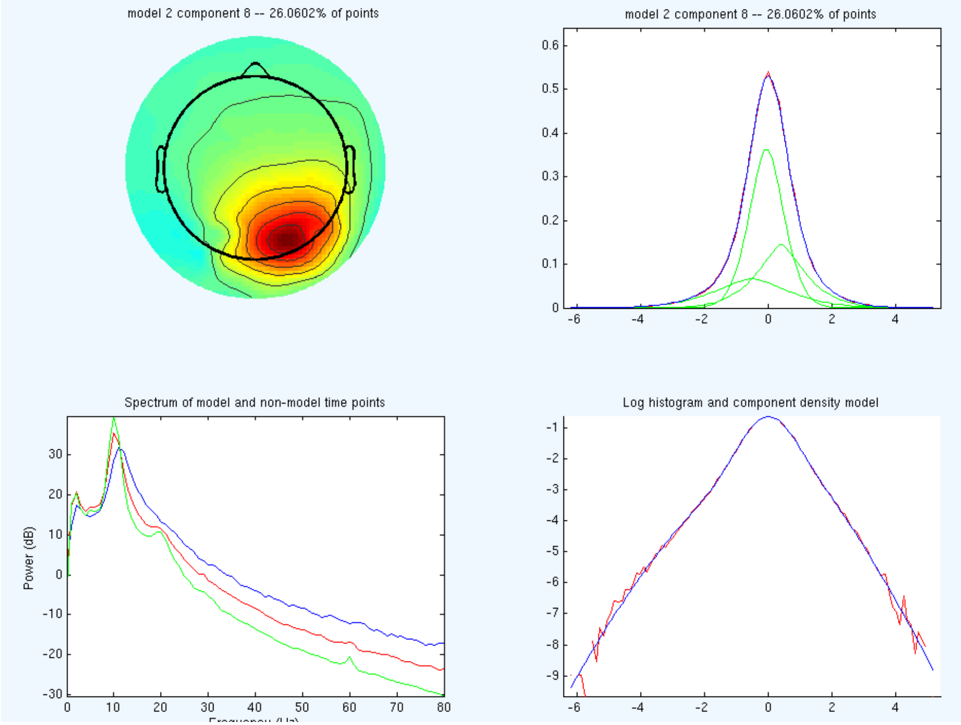
Power line component

- Sub-Gaussian component represented by mixture model of Generalized Gaussian densities



Alpha components



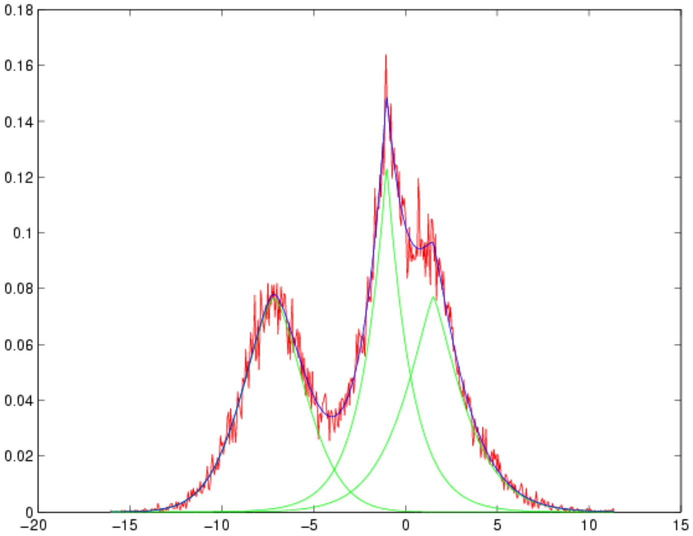


Source Density Mixture Model

- Each source density mixture component has unknown location, scale, and shape:

$$q_{hi}\big(s_i(t)\big) = \sum_{j=1}^{m} \alpha_{hij}\sqrt{\beta_{hij}}\, q_{hij}\Big(\sqrt{\beta_{hij}}\big(s_i(t) - \mu_{hij}\big)\,;\rho_{hij}\Big)$$

- Generalized Gaussian mixture model is convenient and flexible



ICA Algorithms – strategies

- Look for sources with *independent* activity
- Mutual information and likelihood
  - Approx. MI via cumulant expansion of source density
  - Maximum likelihood
    - Fixed source densities – Infomax, FastICA
    - Adaptive / parametric source densities – Pearson, Amica, Extended Infomax
- Multiple lag decorrelation – SOBI, AMUSE, etc.
- Tensor diagonalization – JADE, SHIBBS, FOBI
- Multiple lag tensor diagonalization – JADE-TD

Maximum Likelihood Framework

- Probabilistic model of EEG data is a classical linear model:

  $$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t)$$

  where the sources $\mathbf{s}(t)$ are *independent* (density is product of marginal densities):

  $$p_{\mathbf{s}}(\mathbf{s}(t)) = p_1(s_1(t)) \cdot p_2(s_2(t)) \cdots p_n(s_n(t))$$

- We estimate the unmixing matrix $\mathbf{W}=\mathbf{A}^{-1}$ and estimate sources $\mathbf{y}$:

  $$\mathbf{y}(t) = \mathbf{W}\mathbf{x}(t)$$

- Then the likelihood (prob. dens.) of one time point is:

  $$p_{\mathbf{x}}(\mathbf{x}(t)) = |\det \mathbf{W}|\ p_{\mathbf{s}}(\mathbf{y}(t))$$

- The log likelihood of the data $\mathbf{X}$ assuming temporal independence is:

  $$p(\mathbf{X}) = \prod_t p_{\mathbf{x}}(\mathbf{x}(t)), \qquad \log p(\mathbf{X}) = \sum_t \log|\det \mathbf{W}| + \log p_{\mathbf{s}}(\mathbf{W}\mathbf{x}(t))$$

- We maximize this function (optimize) with respect to $\mathbf{W}$

Mutual Information Reduction (MIR)

- Entropy of linear transformation, $\mathbf{y} = W\mathbf{x}$

$$h(\mathbf{y}) \quad = \quad \log|\det W| + h(\mathbf{x})$$

- Mutual information (instantaneous) for linear transformation:

$$I(\mathbf{y}) \quad = \quad h(y_1) + \ldots + h(y_n) \;-\; \log|\det W| \;-\; h(\mathbf{x})$$

- Total mutual information reduction (MIR) due to linear transformation

$$\begin{aligned} \text{MIR} \quad = \quad I(\mathbf{x}) - I(\mathbf{y}) \quad &= \quad [h(x_1) + \ldots + h(x_n)] \;-\; h(\mathbf{x}) \\ &\quad - [h(y_1) + \ldots + h(y_n)] \;+ \log|\det W| + h(\mathbf{x}) \end{aligned}$$

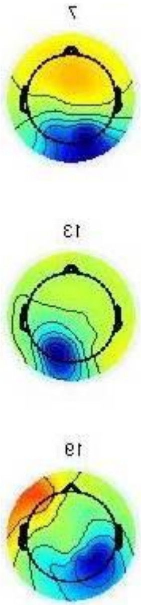
$$= \quad \log|\det W| \;+\; [h(x_1) + \ldots + h(x_n)] \;-\; [h(y_1) + \ldots + h(y_n)]$$

- Similar to ML since entropy $h(y) = E\{-\log p(y)\}$

UCSD

Swartz Center for Computational Neuroscience

Dipolarity and biological plausibility

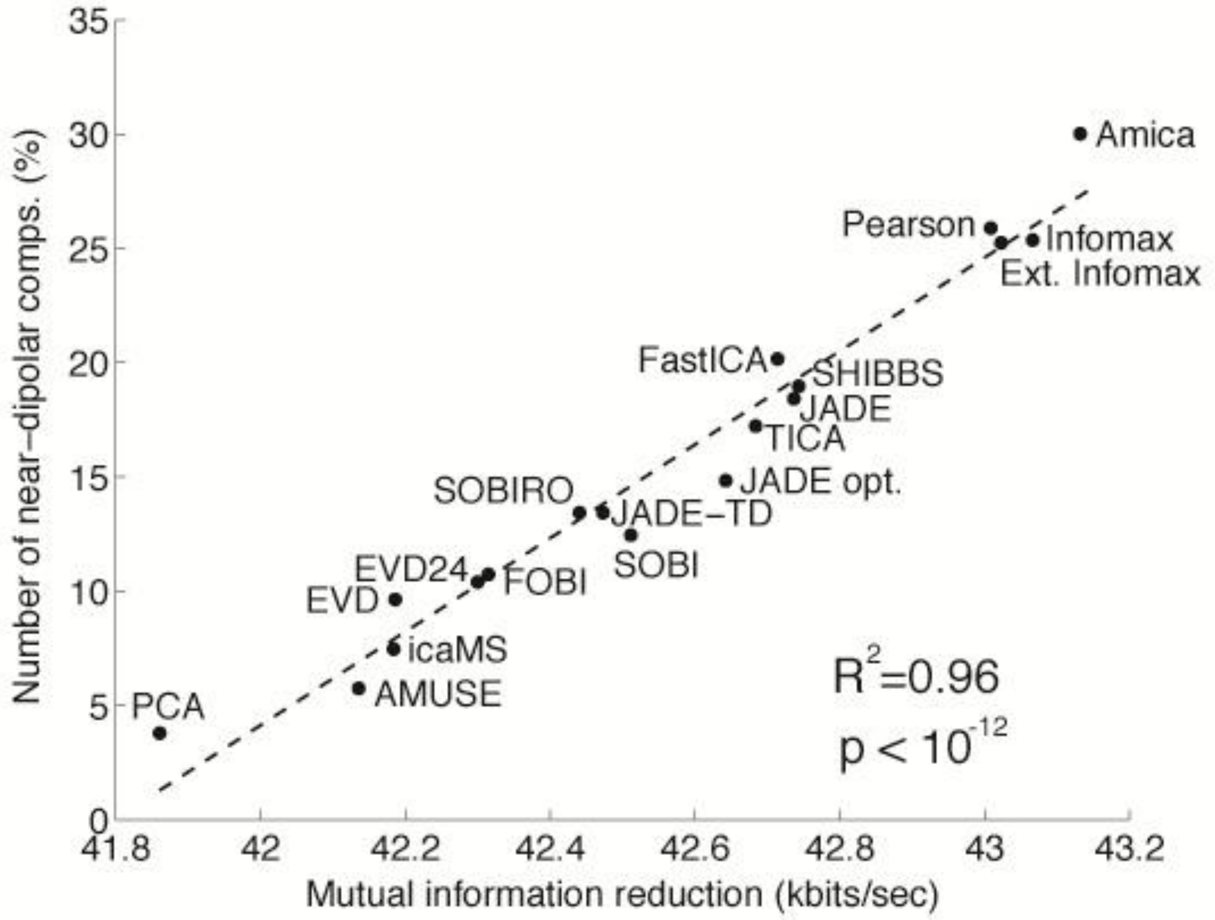


- Dipolarity is measured by fitting a single dipole (projection) to the measured component map and computing *residual variance*

- The dipolarity of a decomposition is the percentage of the estimated components with a residual variance (squared error in dipole fit) less than some threshold (typically 5%)

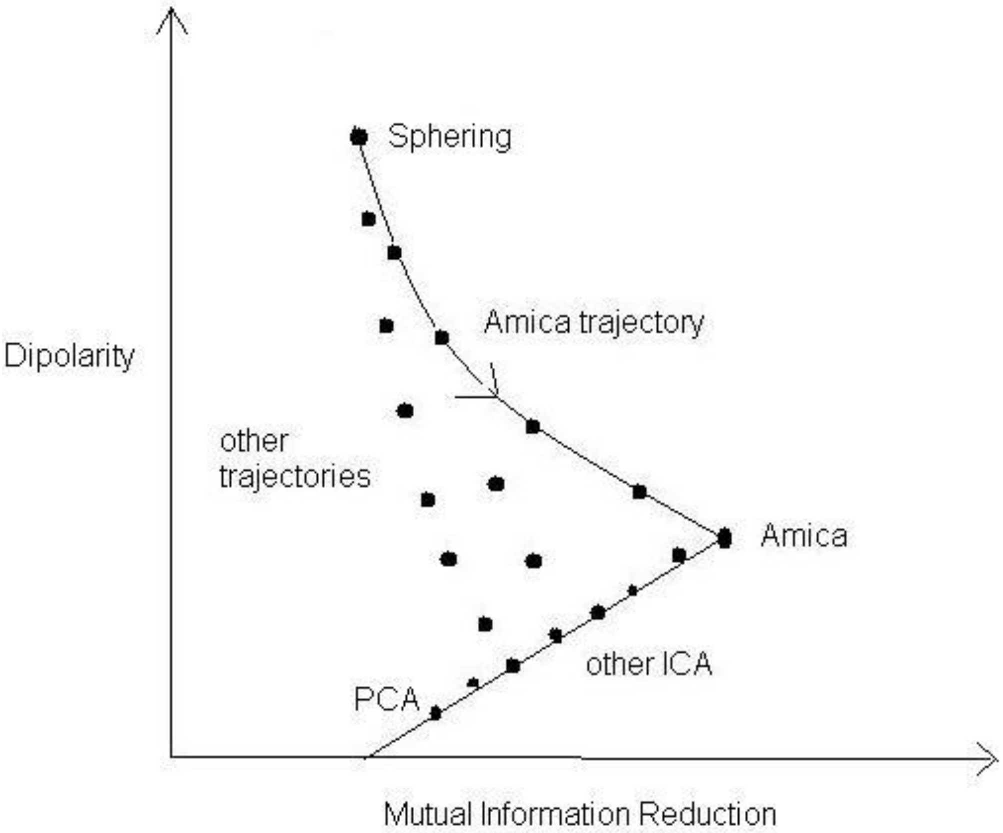
Comparison Dipolarity vs. MIR

Experiment with 14 datasets of 71 channel data, 22 ICA algorithms tested



Artificial dipolarity of sphering

- The Sphering decorrelating basis (not plotted in previous plot) scores high dipolarity because it consists mainly of radial dipoles (with high MI)



What does this tell us?

- The EEG sources really do have some delayed dependence. By trying to eliminate dependence at all lags, the time domain algorithms yield unrealistic (non-dipolar) components. Sophisticated algorithms that are instantaneous only, like JADE, do better.

- Algorithms that enforce decorrelation, like FastICA and JADE, seem to yeild less biologically plausible components. Sources actually have some dependence.

- Algorithms that don’t enforce decorrelation, and that have adaptive source densities (like Ext. Infomax, Pearson, Amica) or have good density models to start with (Infomax) seem to do the best. There is a known higher variance in the component estimate when decorrelation is enforced, so this makes sense.

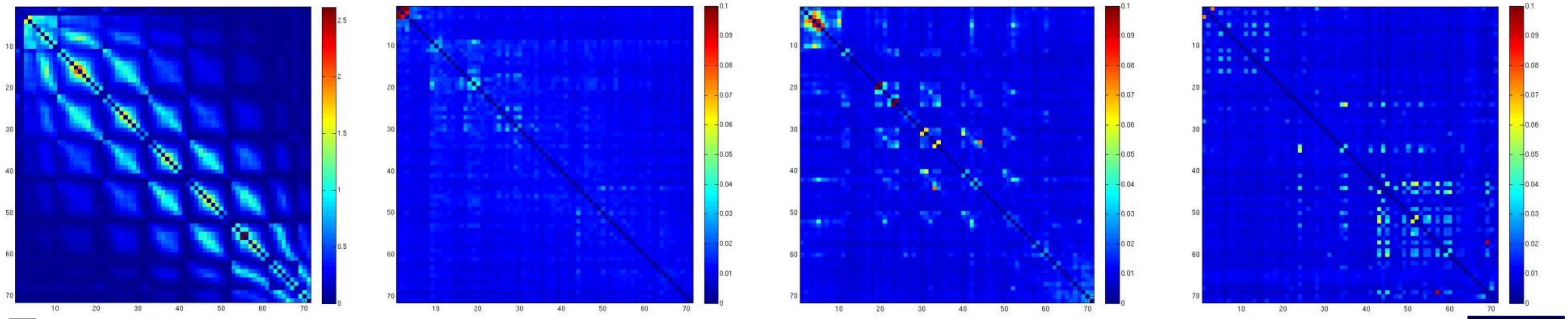
- Among the ML / min mutual info type algorithms, the better the source density is modeled, the better the algorithm does in both MIR and dipolarity. There is a known penalty in asymptotic minimum variance (CRLB) when source density model is misspecified.

Pairwise mutual information

- Pairwise mutual information (PMI):

$$[\mathrm{M}]_{ij} = \mathrm{I}(x_i; x_j) = h(x_i) + h(x_j) - h(x_i, x_j)$$

- Comparison of PMI for original data, PCA (data projected onto eigenvectors), Sphered data, ICA



ICA Mixture Model

- Want to model observations ***x***(*t*), *t* = 1,…,*N*, different models "active" at different times
- Bayesian linear mixture model, *h* = 1, . . . , *M* :

$$\boxed{\mathbf{x}(t) = \mathbf{A}_h\mathbf{s}(t) + \mathbf{c}_h}$$

- Conditionally linear given the model, $\mathbf{W}_h \triangleq \mathbf{A}_h^{-1}$:

$$\boxed{p(\mathbf{x}(t)\,|\,h) = |\det \mathbf{W}_h|\, q_h\big(\mathbf{W}_h(\mathbf{x}(t) - \mathbf{c}_h)\big)}$$

- Samples are modeled as independent in time:

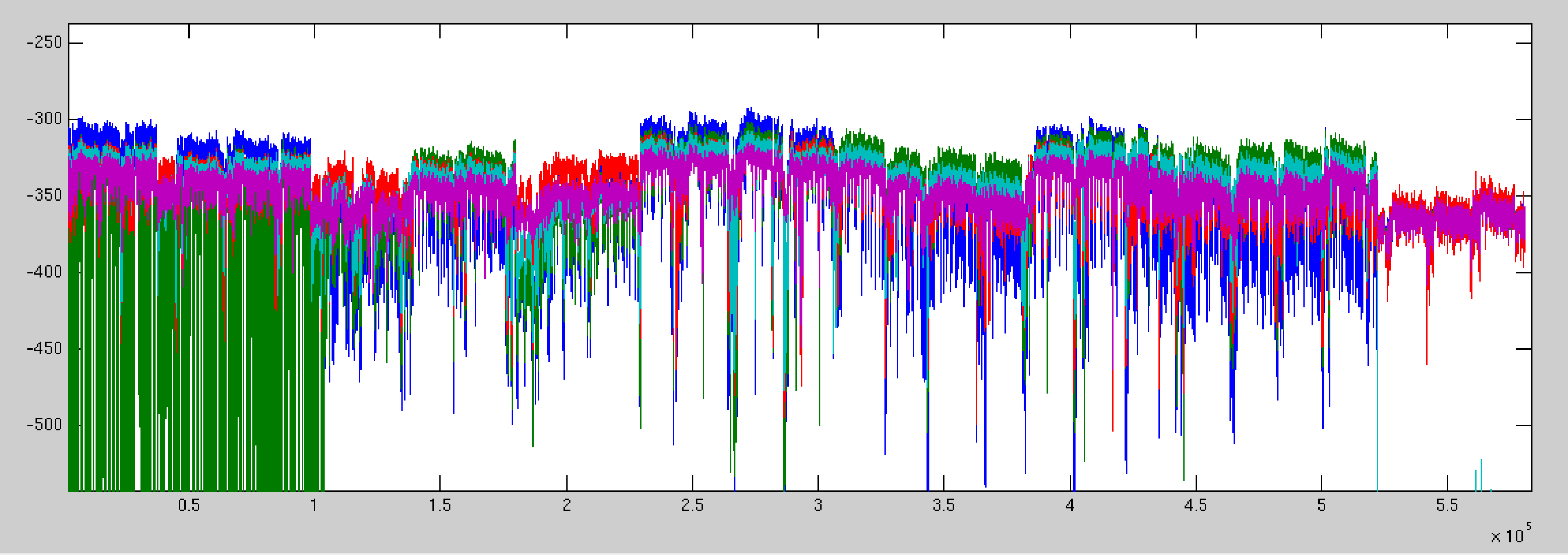
$$\boxed{p(\mathbf{X};\Theta) = \prod_{t=1}^{N}\sum_{h=1}^{M}\gamma_h\, p(\mathbf{x}(t)\,|\,h)}$$

UCSD

Swartz Center for Computational Neuroscience

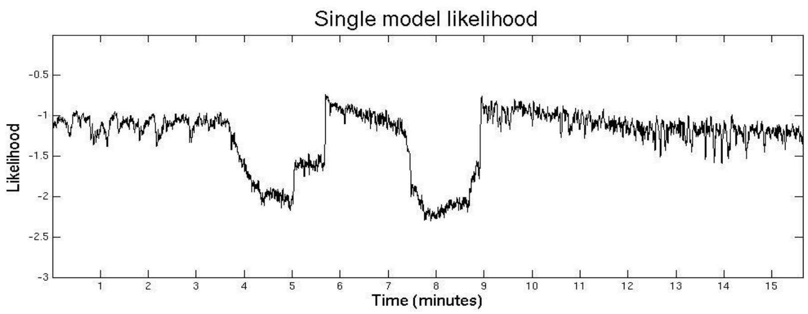
Example segmentation

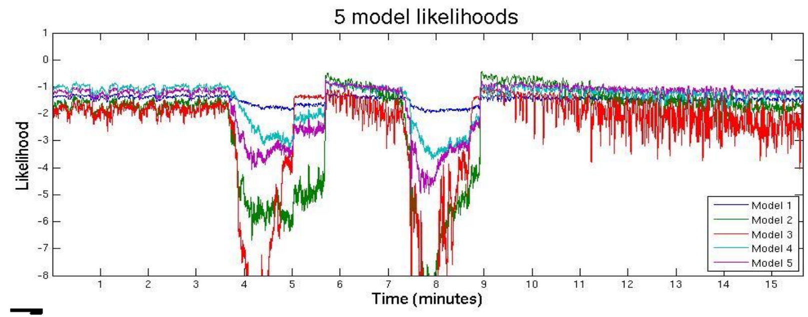
- Task trials are represented by blue, green, and red models
- Red model contains muscle activity not present in blue and green
- Non-task periods represented by cyan and magenta models

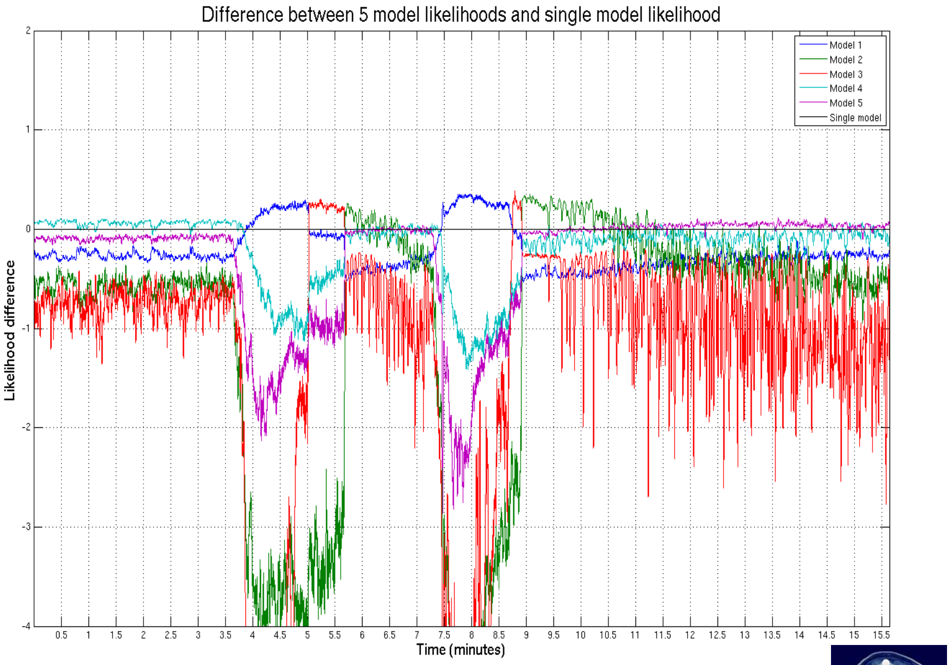


Epilepsy

- Data: 15 minutes from 1 subject containing 2 seizures
- Single model does not represent seizure well
- We learned 5 models – new models consistently adapt to portions of seizure

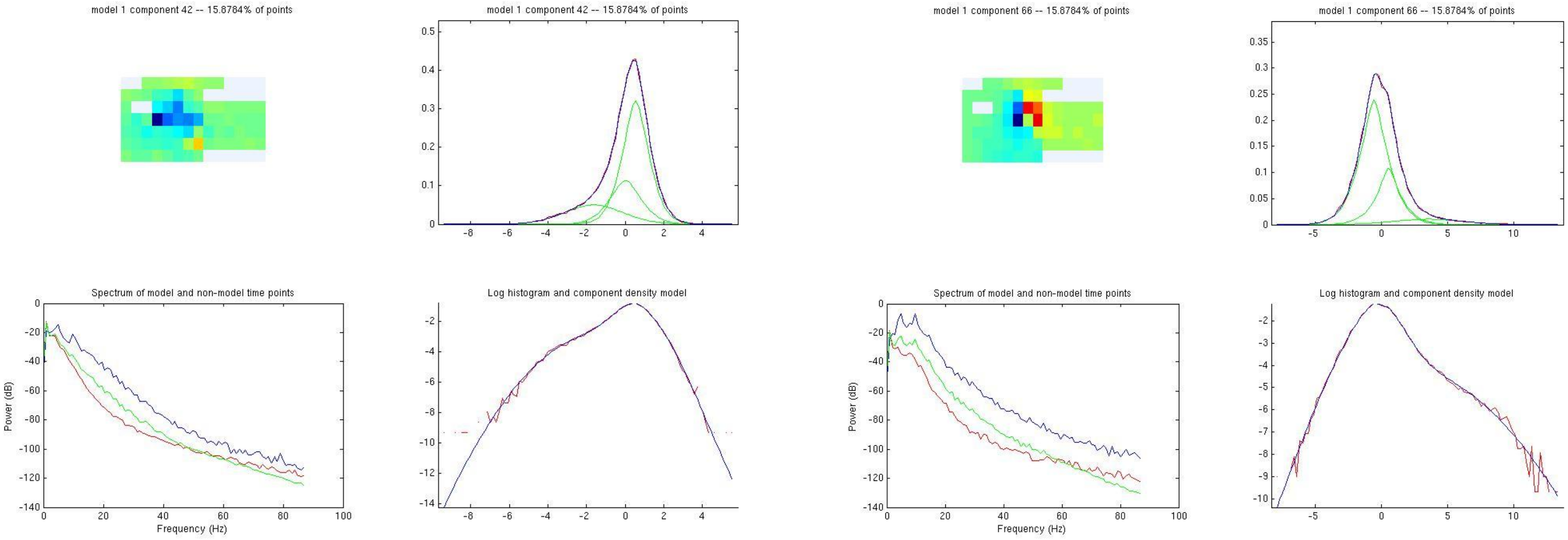

Single model likelihood


5 model likelihoods


Difference between 5 model likelihoods and single model likelihood

Epilepsy Grid Maps

- Maps from grid of electrodes placed intercranially over seizure area
- Source probability densities are fit by mixture model



Conclusion

- ICA is essentially an optimization problem
- Instantaneous ICA algorithms with adaptive source densities yield best EEG components
- Lagged decorrelation algorithms (SOBI, etc.) enforce decorrelation at all time shifts at the cost of biological plausibility
- Some EEG sources may be instantaneously dependent, e.g. alpha, and scalp muscle
- Strategy of minimizing mutual information nevertheless sound because dependent subspaces are separated from rest of sources
- Multiple ICA models can be learned to deal with non-stationarity