Modeling and Analysis of EEG using ICA and AMICA

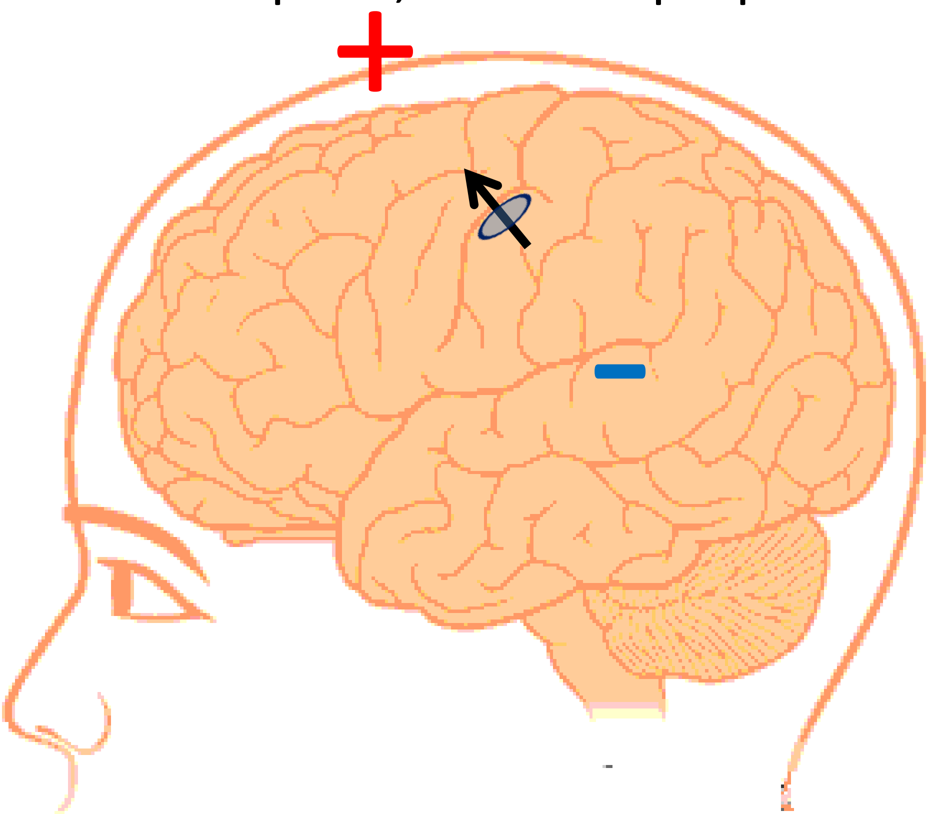
Jason Palmer
Swartz Center for Computational Neuroscience
University of California San Diego, La Jolla, CA

Outline

- Description of EEG data
- The ICA linear superposition model
  - Statistical Linear Model
  - Understanding the nature of independence
- Statistical estimation and likelihood
  - Digression on PCA and dimensionality reduction
- Performance of ICA
  - When does ICA work? Sub- and Super-Gaussianity
  - What factors effect the performance of ICA?
- Fitting source probability densities
  - Generalized Gaussian densities and mixtures
- Comparison of ICA algorithms on EEG data
  - Dipolarity/plausibility vs. Independence
- ICA Mixture model, non-stationarity and data segmentation

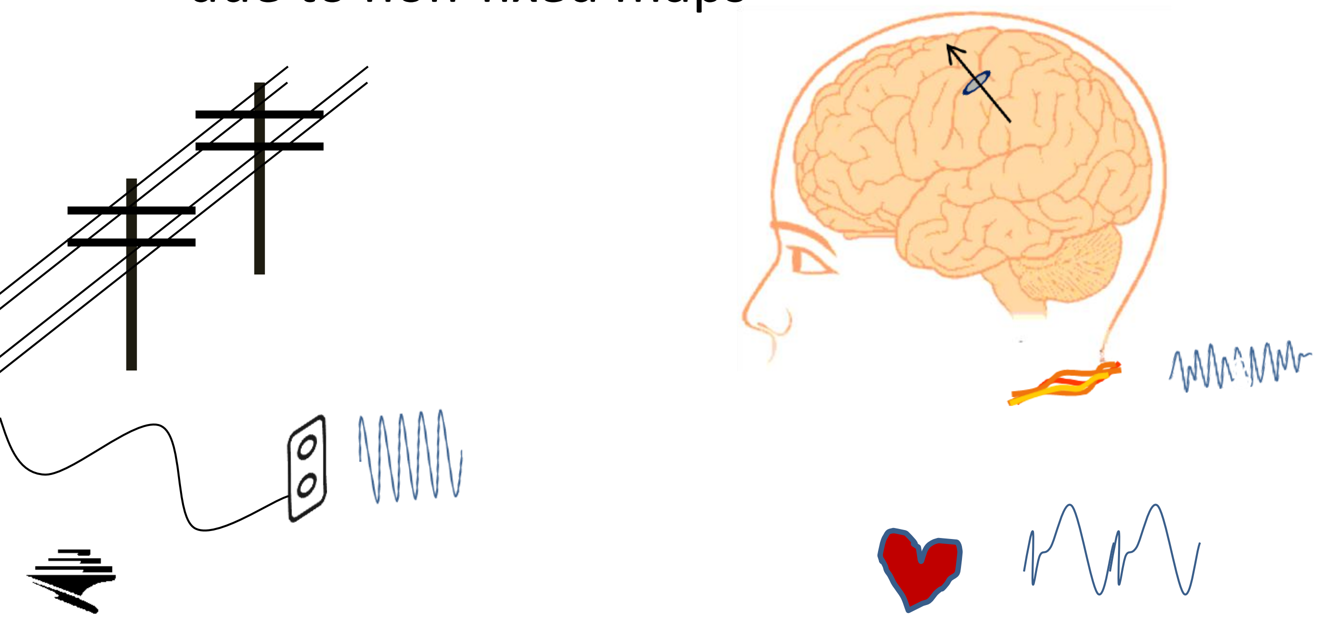
EEG Recordings and Sources

- EEG measures extra-cellular potential fluctuations
- Working theory: local synchronous firing of cortical patches
- In the far field looks dipolar, oriented perpendicular to the cortex



EEG Recordings and Sources

- EEG also picks up firing of muscle cells in scalp
- And heart muscle firing causing heartbeat
- Also non-biological signals, like power line noise
- Sources inside the head are fixed relative to electrodes. Sources outside head (like power line) can be more difficult to separate due to non-fixed maps



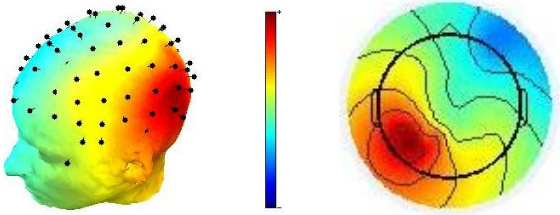
UCSD

Swartz Center for Computational Neuroscience

The ICA (Linear) Model

$\mathbf{a}_i \in \mathbb{R}^n$

- Each source has a an associated "scalp map", or pattern of electric potential measured at the electrodes



- as well as a "source activation" signal driving the map:

$s_i(t)$

- Notation: *i*th map is denoted, $\mathbf{a}_i$, and *i*th source activation, $s_i(t)$

$$\mathbf{A} \triangleq [\mathbf{a}_1\, \mathbf{a}_2\, \cdots\, \mathbf{a}_n] \quad \mathbf{s}(t) \triangleq \begin{bmatrix} s_1(t) \\ \vdots \\ s_n(t) \end{bmatrix} \quad \mathbf{x}(t) \triangleq \begin{bmatrix} x_1(t) \\ \vdots \\ x_n(t) \end{bmatrix}$$

- EEG is linear superposition of *independent sources*:

$$\mathbf{x}(t) \;=\; \mathbf{a}_1 s_1(t) + \cdots + \mathbf{a}_n s_n(t) \;=\; \mathbf{A}\mathbf{s}(t)$$

Independence

- Independence of two events *A* and *B* defined:

  Events are independent if the probability of both occurring is just the product of the probabilities of each happening individually

$$P(A \text{ and } B) = P(A)P(B)$$

- Random variables *X* and *Y* are independent if their events are
- Independence is a stronger property than uncorrelatedness

  Random variables are uncorrelated if the mean (average value) of the product is just the product of the means. Zero mean uncorrelated = *orthogonal*

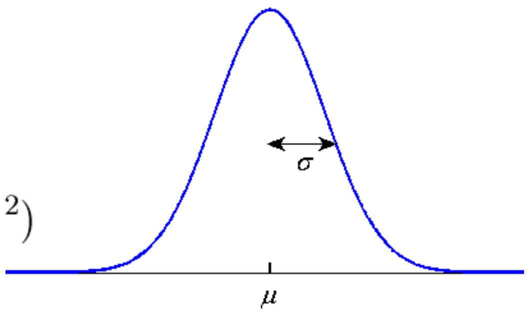
$$E\{XY\} = E\{X\}E\{Y\} \qquad E\{(X-\mu_X)(Y-\mu_Y)\} = 0$$

- Independence implies uncorrelatedness
- But two events can be uncorrelated without being independent. Only in the Gaussian case does uncorrelatedness imply independence.

Independence and PDFs

- A probability density function describes the distribution of a random variable, or random vector, over a space
- E.g. if *X* is a Gaussian, or Normal, random variable, $X \sim \mathcal{N}(x; \mu, \sigma^2)$

$$p_X(x) \ = \ \mathcal{N}(x; \mu, \sigma^2) \ \triangleq \ (2\pi)^{-1/2}\sigma^{-1}\exp(-\tfrac{1}{2}\sigma^{-2}(x-\mu)^2)$$



- In terms of pdfs, r.v.'s are independent if and only if the joint probability density factorizes: $p_{X,Y}(x,y) = p_X(x)p_Y(y)$
- Or in vector notation:

$$p_{\mathbf{s}}(\mathbf{s}) = p_{s_1}(s_1)p_{s_2}(s_2)\cdots p_{s_n}(s_n)$$

- Mutual Information can be used to measure dependence

Only zero when variables are independent (unlike correlation)

$$\mathrm{MI}(X,Y) = I(X;Y) = D\big(p_{X,Y}(x,y)\,\big\|\,p_X(x)p_Y(y)\big) \geq 0$$

UCSD

Swartz Center for Computational Neuroscience

Statistical Model

- The Linear Model (no channel noise): $\mathbf{x} = \mathbf{A}\mathbf{s}$

  Sources are independent: $p_{\mathbf{s}}(\mathbf{s}) = p_{s_1}(s_1)p_{s_2}(s_2)\cdots p_{s_n}(s_n)$

  PDF of the EEG recording can be calculated in closed form:

$$p_{\mathbf{x}}(\mathbf{x}) = |\det \mathbf{A}^{-1}|\,p_{\mathbf{s}}(\mathbf{A}^{-1}\mathbf{x})$$

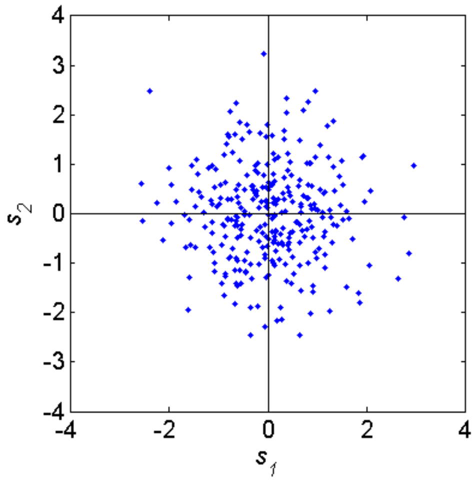
  – Define the "unmixing matrix": $\mathbf{W} \triangleq \mathbf{A}^{-1}$

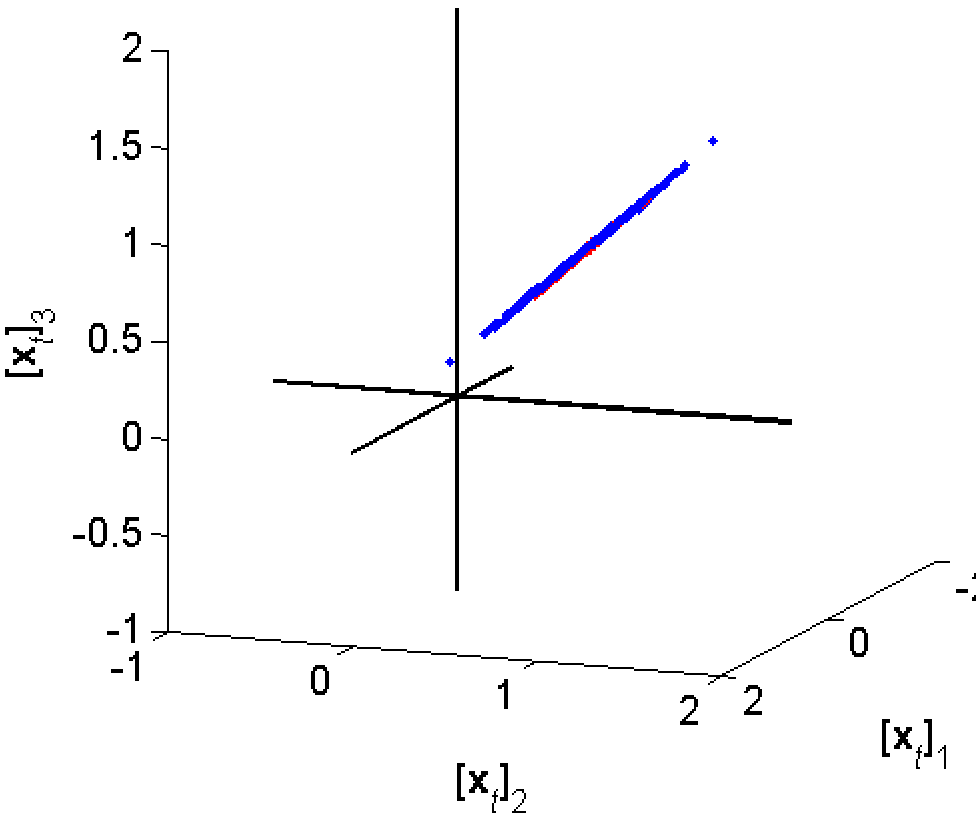
$$p_{\mathbf{x}}(\mathbf{x}) = |\det \mathbf{W}|\,p_{\mathbf{s}}(\mathbf{W}\mathbf{x}) = |\det \mathbf{W}| \prod_{i=1}^{n} p_{s_i}(\mathbf{w}_i^T\mathbf{x})$$

- Key assumption: fewer sources than sensors
  – Necessary to compute density in closed form
  – Can be relaxed to an extent using ICA mixture model, described later

PCA and Dimensionality Reduction

- If there are more channels than sources, then the data will lie in "subspace" of the full space (number of channels)
- PCA can determine a "basis" for this subspace, i.e. a set of vector that "span" the space
- "Sources" are uncorrelated





PCA and Dimensionality Reduction

- De-correlating basis is not unique. Covariance matrix is defined:

$$[\boldsymbol{\Sigma}]_{ij} \triangleq E\{(x_i - \mu_{x_i})(x_j - \mu_{x_j})\}$$

$$\boldsymbol{\Sigma} = E\{(\mathbf{x} - \mu_{\mathbf{x}})(\mathbf{x} - \mu_{\mathbf{x}})^T\}$$

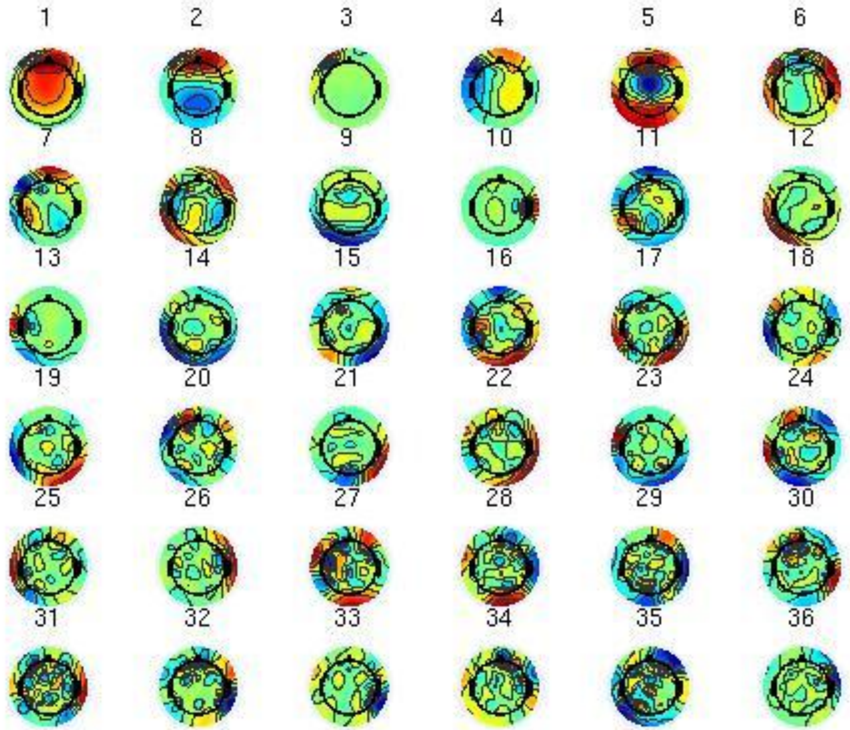
- Factorize the covariance matrix into "square roots": $\boldsymbol{\Sigma} = \mathbf{A}\mathbf{A}^T$
- But there is an infinite number of such roots, e.g.,

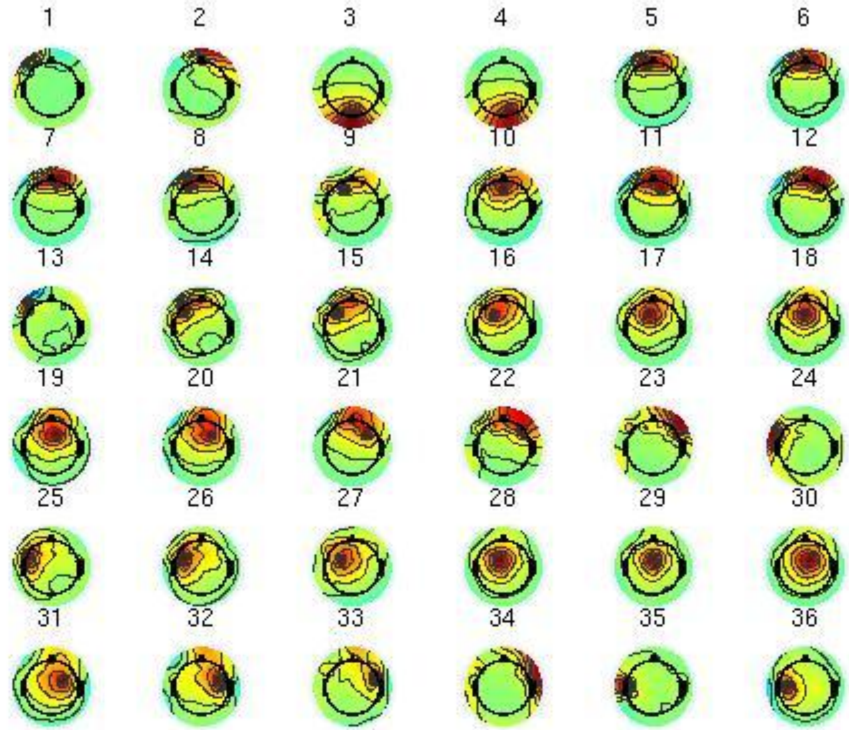
  Cholesky: $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}^T$ Eigen-decomposition: $\boldsymbol{\Sigma} = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^T$

- There is a unique symmetric square root: $\mathbf{Q}\boldsymbol{\Lambda}^{1/2}\mathbf{Q}^T$
- Decorrelating at two different time lags is unique
  - SOBI approach simultaneously approximately diagonalizes covariance matrices at multiple lags—more in comparison section

Eigenvectors and Sphering maps

- We use PCA as a preprocessing step for ICA, potentially reducing the dimension of the data if it is not full rank (e.g. when avg reference is used)
- Not unique! E.g. two decorrelating basis sets: eigenvectors (left), symmetric sphering basis (right). Sphering basis (right) more biologically realistic





Statistical Model (again)

- The instantaneous Linear Model (no channel noise):

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

- Sources are independent:

$$p_{\mathbf{s}}(\mathbf{s}) = p_{s_1}(s_1)p_{s_2}(s_2)\cdots p_{s_n}(s_n)$$

- PDF of the EEG recording can be calculated in closed form:

$$p_{\mathbf{x}}(\mathbf{x}) = |\det \mathbf{A}^{-1}|\, p_{\mathbf{s}}(\mathbf{A}^{-1}\mathbf{x})$$

- Define the "unmixing matrix": $\mathbf{W} \triangleq \mathbf{A}^{-1}$

$$p_{\mathbf{x}}(\mathbf{x}_t) = |\det \mathbf{W}|\, p_{\mathbf{s}}(\mathbf{W}\mathbf{x}_t) = |\det \mathbf{W}| \prod_{i=1}^{n} p_{s_i}(\mathbf{w}_i^T \mathbf{x}_t)$$

UCSD

Swartz Center for Computational Neuroscience

ICA – Estimation and Optimization

- In the statistical model, the sources (and thus the data) are modeled as temporally independent. For *N* samples (time points):

$$p_{\mathbf{x}_1,\ldots,\mathbf{x}_N}(\mathbf{x}_1,\ldots,\mathbf{x}_N) \;=\; \prod_{t=1}^{N} p_{\mathbf{x}}(\mathbf{x}_t) \;=\; |\det \mathbf{W}|^N \prod_{t=1}^{N} p_{\mathbf{s}}(\mathbf{W}\mathbf{x}_t)$$

- Define the “log likelihood” of the data:

$$L(\mathbf{x}_1,\ldots,\mathbf{x}_N) \;=\; \log|\det \mathbf{W}| + \frac{1}{N}\sum_{t=1}^{N} \log p_{\mathbf{s}}(\mathbf{W}\mathbf{x}_t)$$

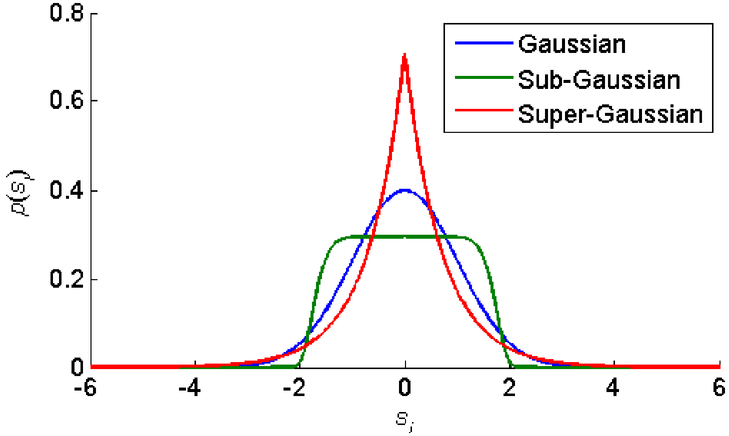
- We wish to maximize the function over the parameters
- The following interpretations of ICA are equivalent:
  - Maximum Likelihood
  - Minimize KL divergence (find model with the best fit to the data)
  - Minimize mutual information

Unknown Source Densities

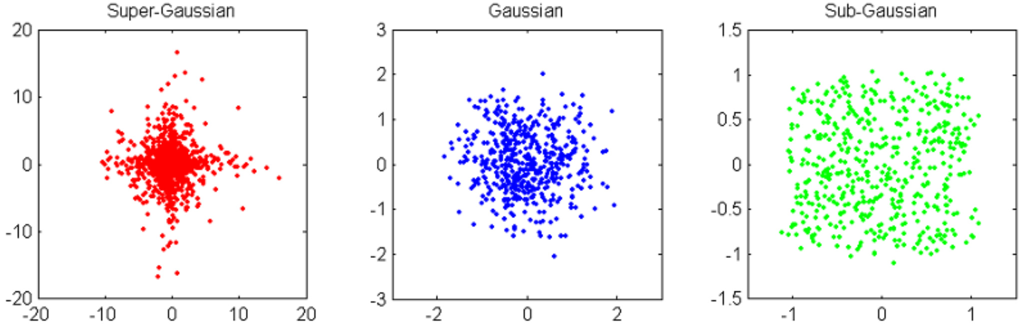
- Likelihood involves unknown source densities:

$$p_{s_i}(s_i),\ \ i = 1, \ldots, n$$

- It turns out, however, that we only need to the know basic form of density — sub-Gaussian or super-Gaussian
  - Gaussian: limiting distribution of sums of random variables
  - Super-Gaussian: heavier tails, sharper peak, positive kurtosis
  - Sub-Gaussian: light tails, like uniform density, negative kurtosis

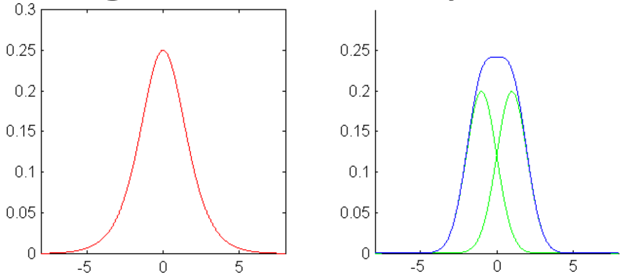


- Scatter plots of two independent random variables:



Unknown Source Densities

- Fixed forms for sub- and super-Gaussian densities are sufficient to separate sources as number of samples goes to infinity
  - E.g. The Fastica (Hyvarinen) and Extended Infomax (Bell, Sejnowski, and Lee) algorithms use this approach. Ext. Infomax logistic and GMM shown:



- Sign of *kurtosis*, or normalized kurtosis, can be used to determine online what an estimated source is:

$$\kappa_4 \triangleq E\{X^4\} - 3E\{X^2\}^2 \qquad \kappa \triangleq \frac{E\{X^4\}}{\sigma^4} - 3$$

  Positive = super-Gaussian, negative = sub-Gaussian

- However, asymptotic stability does not guarantee good performance for a finite number of samples (fixed *N*)

Measuring Performance of ICA

- Given a statistical model, we can calculate asymptotic lower bound on variance (error) in our parameter estimates
- Formulate as problem of estimating: $\hat{\mathbf{C}} \triangleq \hat{\mathbf{W}}\mathbf{A} \approx \mathbf{PD}$
- Generally, if sub- and super-Gaussian chosen correctly, the expected value of $\hat{\mathbf{C}}$ is a permuted diagonal matrix—ICA works
- But the variance in the estimates generally differs. We have:

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

$$\hat{s}_i(t) = \hat{c}_{i1}s_1(t) + \cdots + \hat{c}_{in}s_n(t) \approx \hat{c}_{ij}s_j(t)$$

- Then in terms of normalized (unit variance) sources:

$$\frac{\hat{s}_i(t)}{\sigma_i} = \hat{c}_{i1}\frac{\sigma_1}{\sigma_i}\frac{s_1(t)}{\sigma_1} + \cdots + \hat{c}_{in}\frac{\sigma_n}{\sigma_i}\frac{s_n(t)}{\sigma_n} \approx \hat{c}_{ik}\frac{\sigma_k}{\sigma_i}\frac{s_k(t)}{\sigma_k}$$

Measuring Performance of ICA

- The *j*th interfering (normalized) source is multiplied by: $\hat{c}_{ij}\dfrac{\sigma_j}{\sigma_i}$
- Define:

$$\phi_i \triangleq -E\left\{\frac{d^2}{ds_i^2}\log p_{s_i}(s_i)\right\}$$

- We can bound the variance of the *j*th "contaminating coefficient":

$$E\{\hat{c}_{ij}^2\}\frac{\sigma_j^2}{\sigma_i^2} \geq \frac{1}{N}\frac{\phi_j\sigma_j^2}{(\phi_i\sigma_i^2)(\phi_j\sigma_j^2)-1}$$

- So optimal performance in ICA is characterized by: $\mathcal{L}_i \triangleq \phi_i\sigma_i^2$

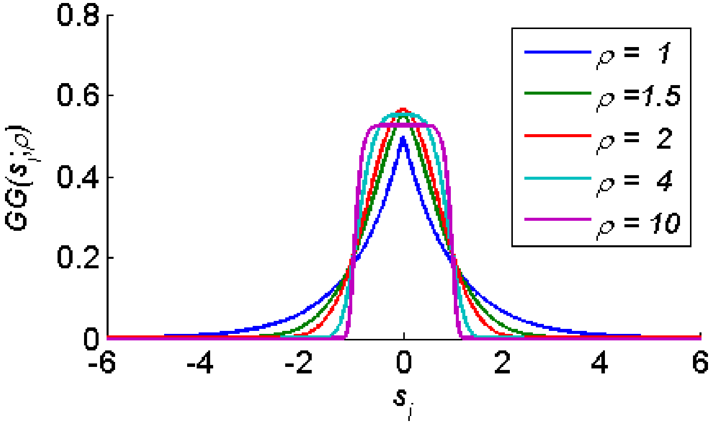
$$\text{Optimal Rejection Rate} = \frac{1}{N}\frac{\mathcal{L}_j}{\mathcal{L}_i\mathcal{L}_j - 1}$$

- This bound gives the optimal performance achievable assuming that each source density is known. Similar but more complicated expression can be derived in terms of approximating densities.

Generalized Gaussian Densities

- The Generalized Gaussian density has the following basic form, where rho is the "shape parameter":

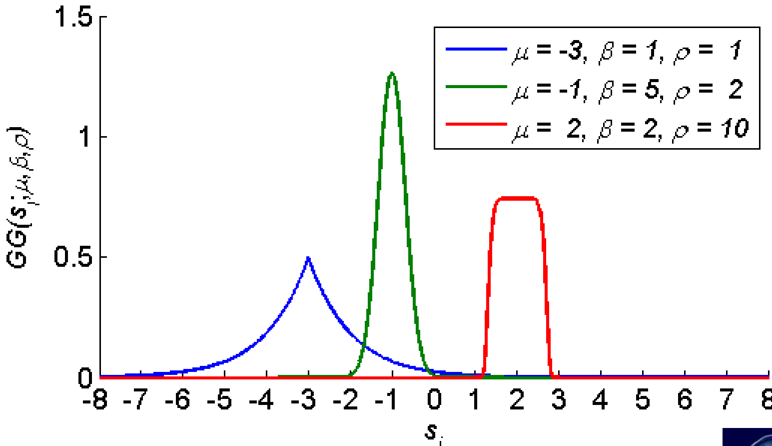
$$\mathcal{GG}(s;\rho) \triangleq \frac{1}{2\,\Gamma(1+1/\rho)}\exp(-|s|^{\rho})$$



- Adding location and scale parameters:

$$\mathcal{GG}(s;\mu,\beta,\rho) \triangleq$$

$$\frac{\beta^{1/2}}{2\,\Gamma(1+1/\rho)}\exp(-\beta^{\rho/2}|s-\mu|^{\rho})$$



UCSD

Swartz Center for Computational Neuroscience

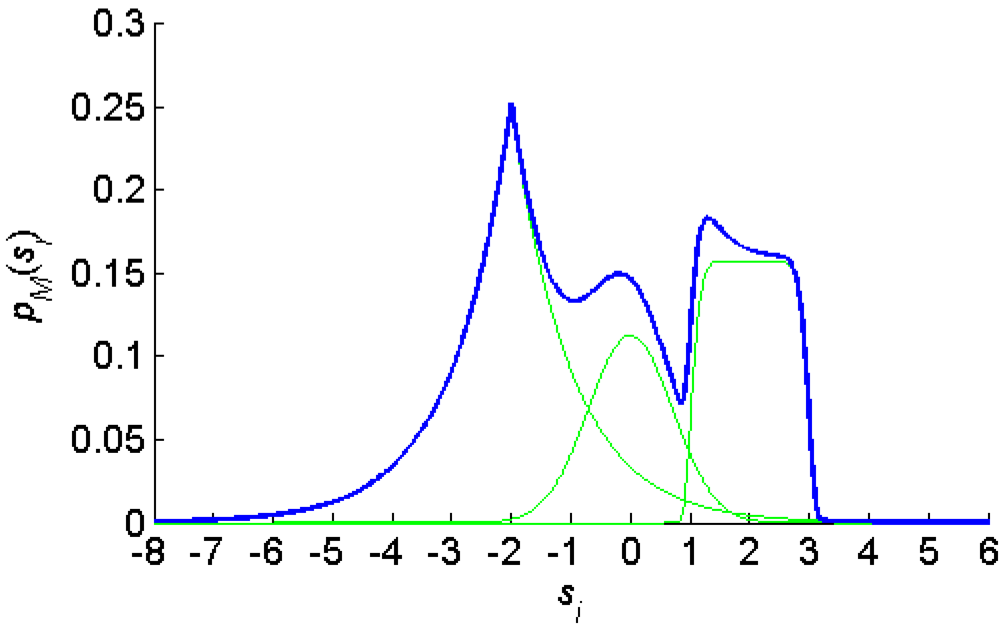
Generalized Gaussian Mixtures

- More complex densities can be constructed using a mixture model:

$$p_M(s_i) = \sum_{j=1}^{m} \alpha_j p_j(s_i)$$

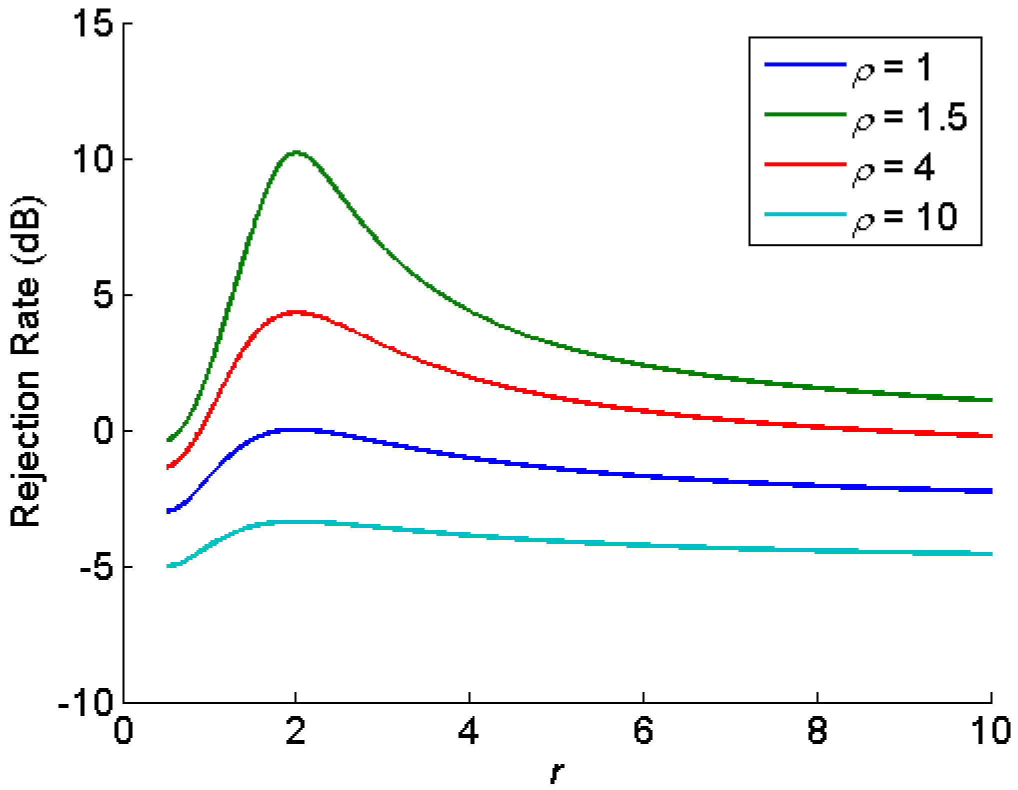
- For example, consider, the following model:

$$\frac{1}{2}\mathcal{GG}(s_i; -2, 1, 1) + \frac{2}{10}\mathcal{GG}(s_i; 0, 1, 2) + \frac{3}{10}\mathcal{GG}(s_i; 2, 1, 10)$$



Generalized Gaussian Interference

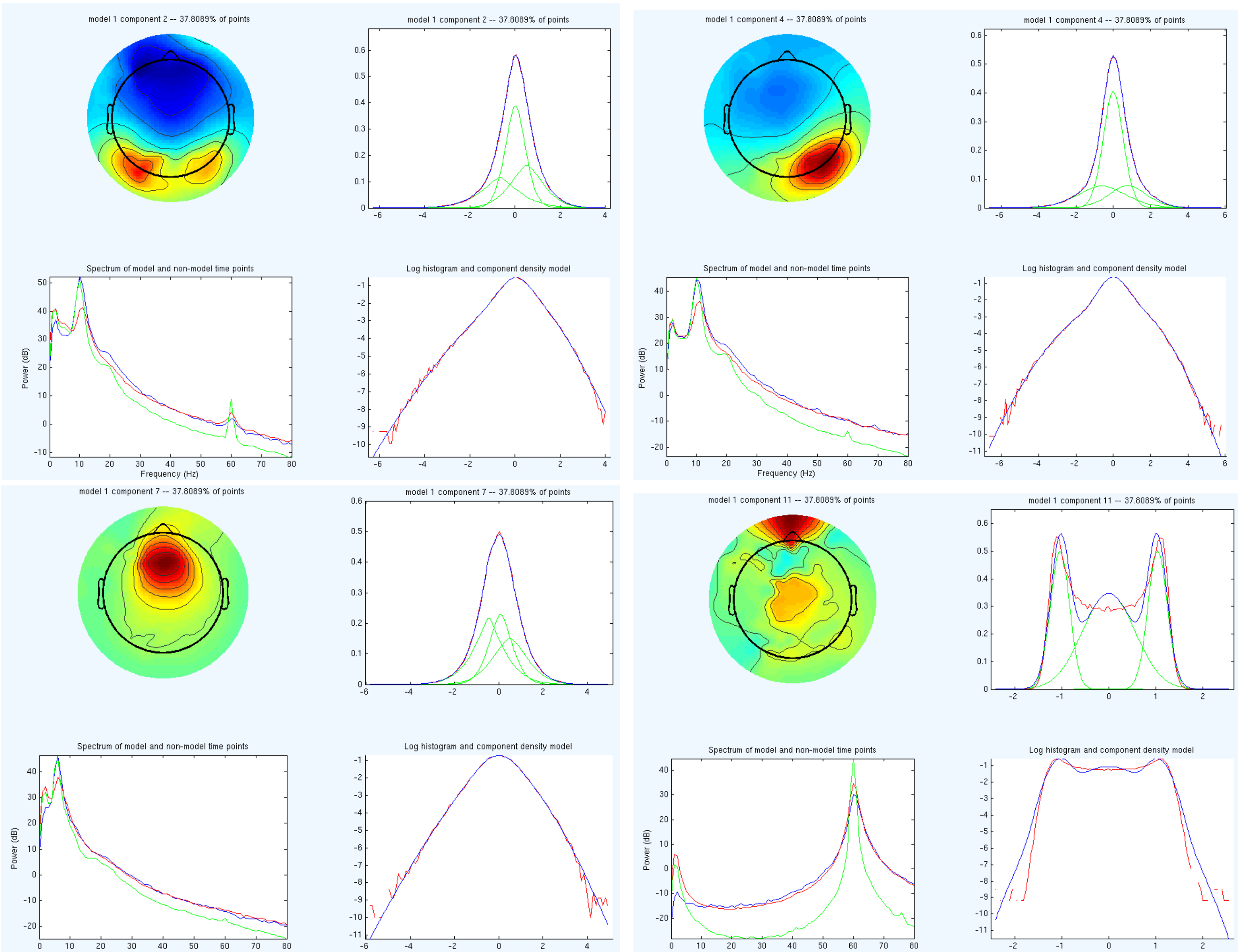
- Now suppose we are trying to estimate a source with a Generalized Gaussian density with shape parameter ρ, which is mixed with a Generalized Gaussian with shape r



EEG Source Model Examples

α

α



θ

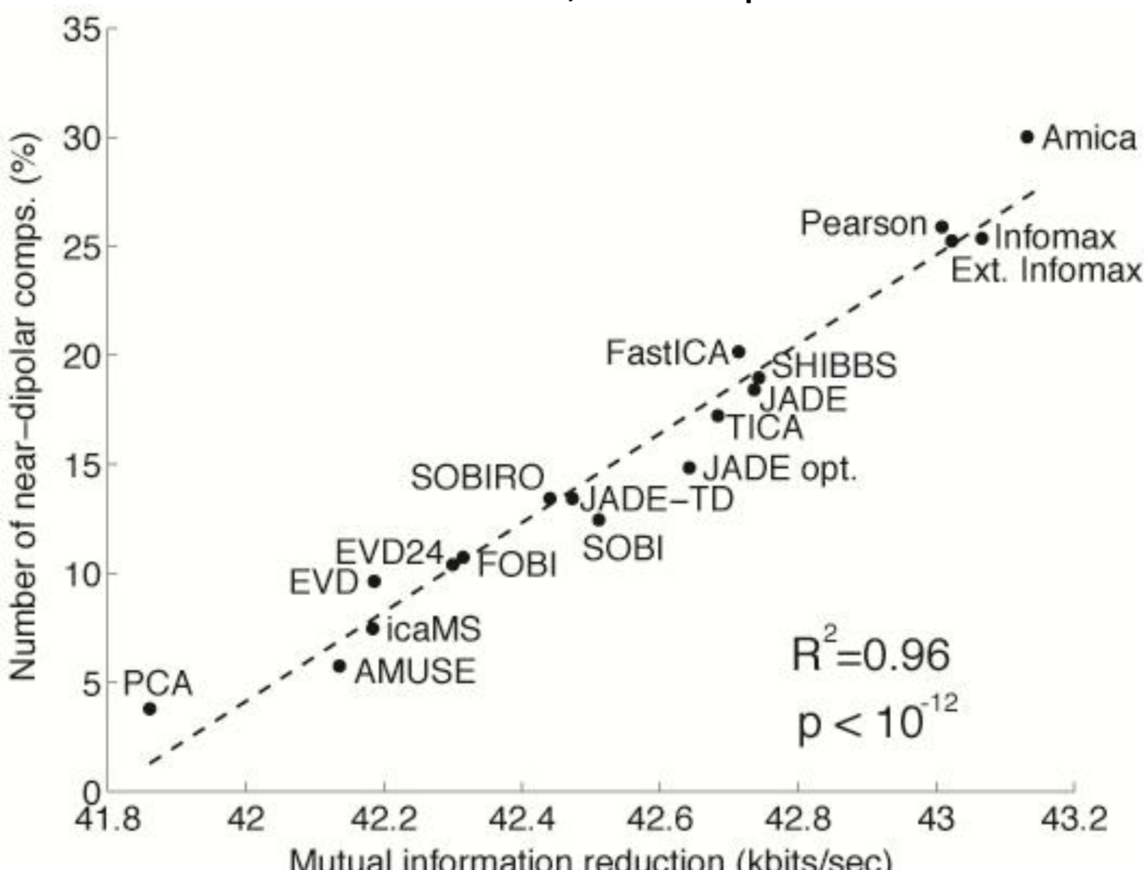
Power line

UCSD

Swartz Center for Computational Neuroscience

An EEG Study

- Experiment with 14 datasets of 71 channel Sternberg data
- 22 different ICA algorithms downloaded and tested
- Amount of MI removed from data, vs. # dipoles with < 5% residual variance

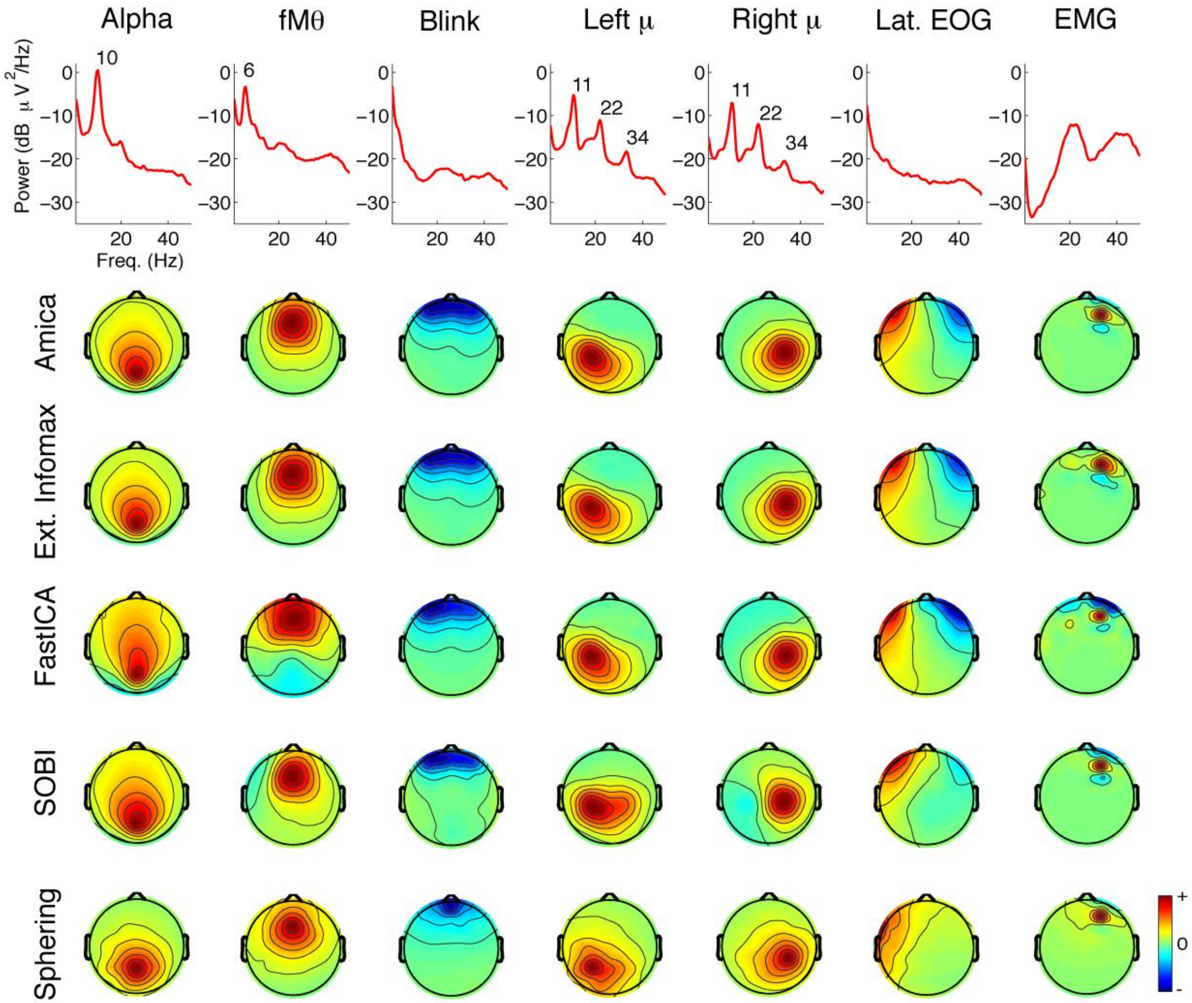


UCSD

Swartz Center for Computational Neuroscience

An EEG Study

- The more dependence we remove, the better our components look!



The ICA Mixture Model

- The statistical framework can be extended to estimate multiple ICA models simultaneously – ICA mixture model (Lee et al)

$$p_{\mathbf{x}}(\mathbf{x}_t) = \sum_{h=1}^{M} \gamma_h \, p_h(\mathbf{x}_t \,|\, h_t = h)$$

- M models indexed by $h \in \{1,\ldots,M\}$. Each model has its own maps, $\mathbf{A}_h$, sources, $\mathbf{s}_h$ (with source models), and "centers" $\mathbf{c}_h$
- Given which model is active, i.e. given h, the model is linear:

$$p_{\mathbf{x}}(\mathbf{x}_t \,|\, h_t = h) = |\det \mathbf{W}_h| \, p_{\mathbf{s}_h}\big(\mathbf{W}_h(\mathbf{x}_t - \mathbf{c}_h)\big)$$

- Different models are active at different times—one model active at a time. At each time t, choose an "active" model index with probabilities $\gamma_1, \ldots, \gamma_M$.

Posterior Likelihood

- Once the model parameters have been learned, we can use Bayes' Rule to compute the posterior likelihood that $h_t = h$:

$$P(h_t = h \,|\, \mathbf{x}_t) \;=\; \frac{p(\mathbf{x}_t \,|\, h_t = h) P(h_t = h)}{\sum_{h'=1}^{M} p(\mathbf{x}_t \,|\, h_t = h') P(h_t = h')}$$

- This gives us the likelihood of each model at each time point

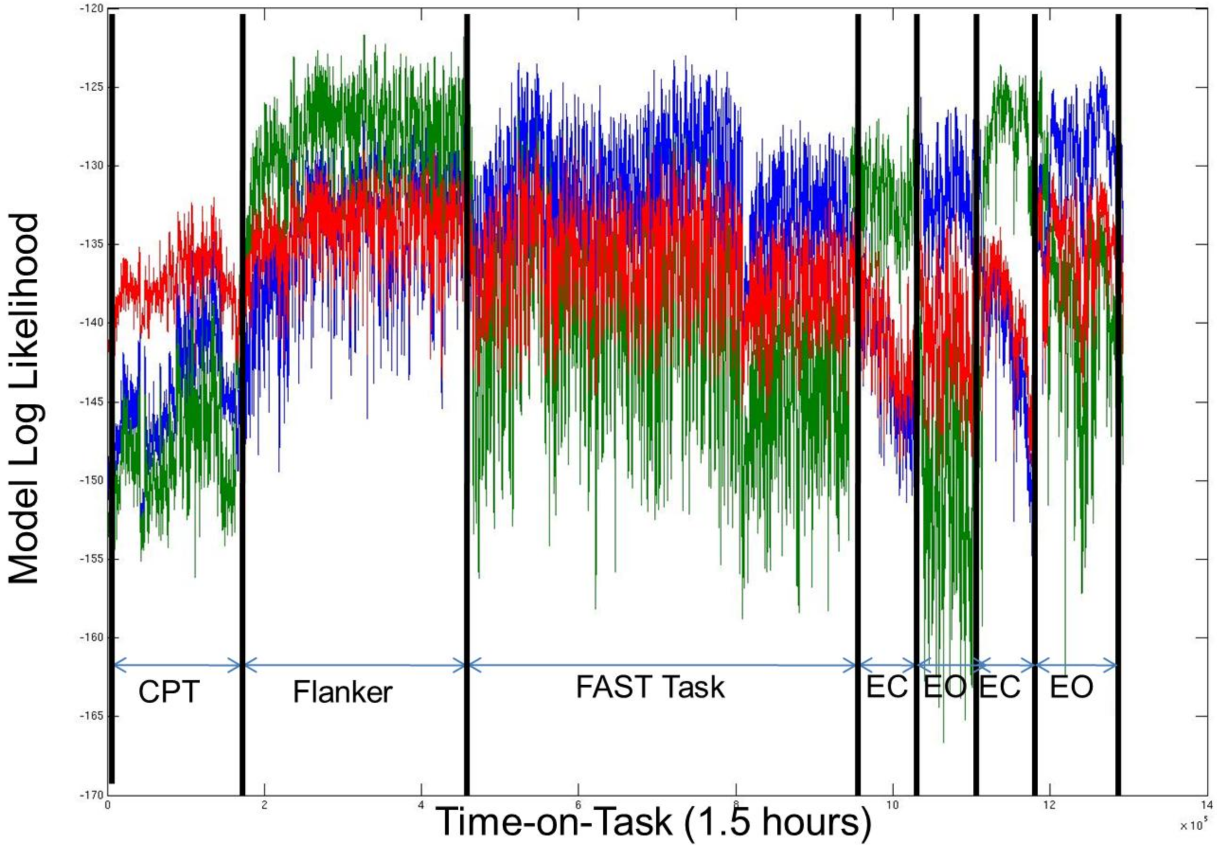
- Segmentation is performed by declaring the active model at time t to be the one with highest posterior likelihood

$$\hat{h}_t = \arg\max_h P(h_t = h \,|\, \mathbf{x}_t)$$

Segmentation Example: Multi-task

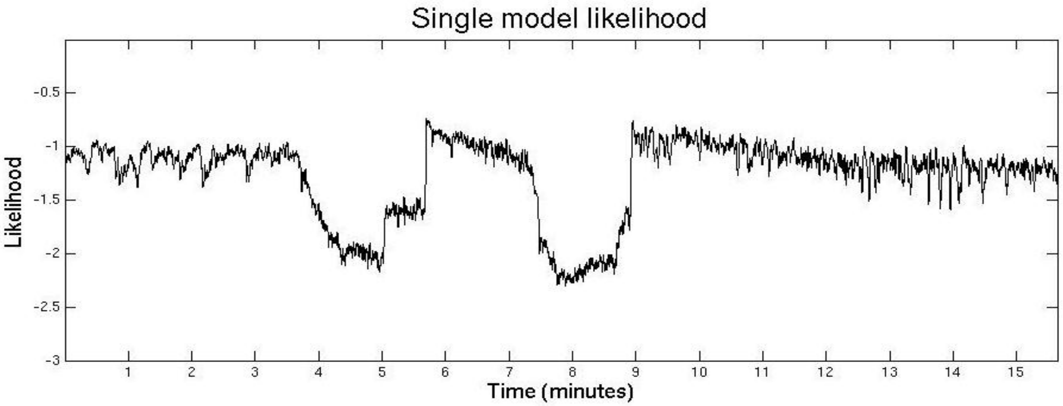
- 3 models learned on 1.5 hour recording with multiple tasks

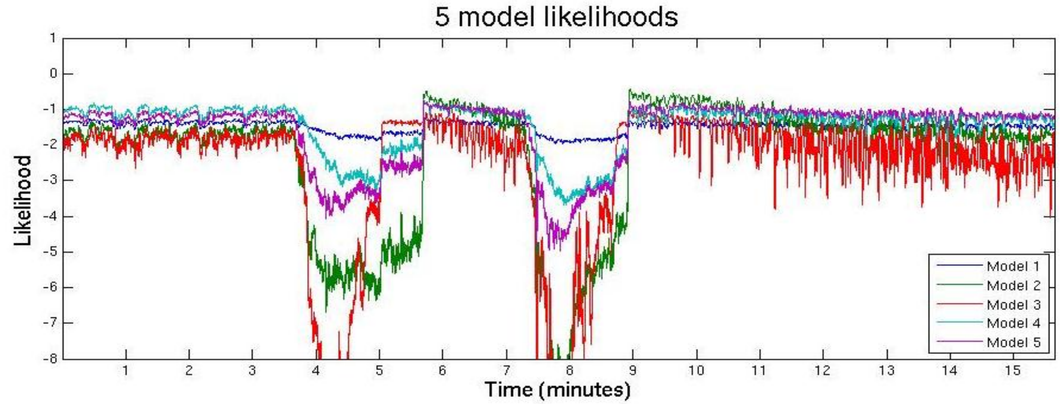
  Continuous Performance Task (CPT), Eriksen Flanker, Fast Response, Eyes Closed (EC), Eyes Open (EO)



Segmentation Example: Epileptic Seizure

- 15 minutes of ECoG array recording with 2 seizures
- Single model trained on seizure data shows drop in likelihood at seizures
- 5 models were estimated
- The models segmented the seizure and non-seizure data, as well as different periods within seizures
- The segmentation is the same in two seizures
- Segmentation is consistent using different number of models in the training





Conclusions

- EEG data is well modeled using the instantaneous ICA linear model
  - PCA can be used for preprocessing, decorrelation, dimensionality reduction, not unique!
- ICA basis is unique if sources are non-Gaussian. In principle we only need to match general from (sub- or super-Gaussian) of density
- However, variance in the estimate (separation quality, map quality) depends on how precisely source density is modeled.
- We presented a general framework for modeling arbitrary independent source distributions using adaptive mixtures of Generalized Gaussian mixture model – AMICA
- We extended single-model AMICA to multiple models to account for data non-stationarity using the ICA mixture model, and showed how to estimate posterior model likelihood for unsupervised data segmentation
- More detail and background, as well as code for all functions and figure generation, is available on the EEGLAB Wiki:
  - Linear_Representation_and_Basis_Vectors
  - Random_Variables_and_Probability_Density_Functions
  - Amica
  - Amica_download

Acknowledgements

- The ICA comparison study was done in collaboration with Arnaud Delorme
- Much of the data used during algorithm development and testing was collected by Julie Onton at SCCN
- The Multi-task study was done in collaboration with Gráinne McLoughlin (KCL)
- Understanding and interpretation of EEG is directed by Scott Makeig (UCSD)
- Computing resources tirelessly maintained by Robert Buffington at SCCN
- To the conference organizers: Moltes gràcies!
- I gràcies als vostès per escoltant!