# Independent Component Analysis

## P9120 Group 3

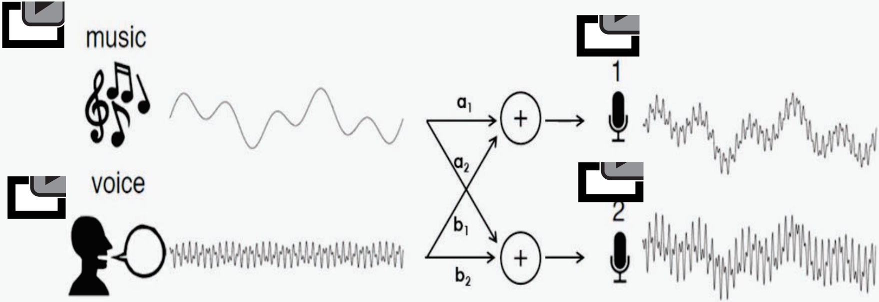
Ruoyu Ji Haokun Yuan Apoorva Srinivasan Tianchen Xu

Biostatistics Department
Columbia University

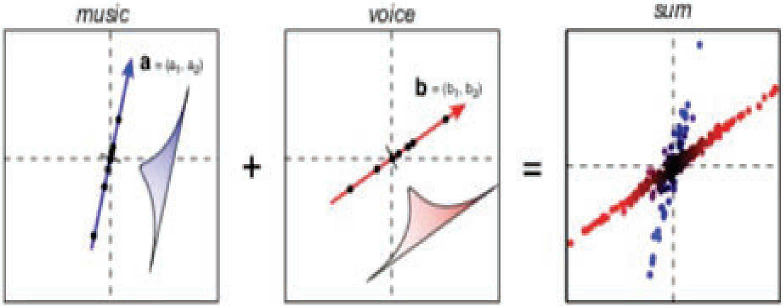
October 24, 2019

# Table of Contents

# Motivating Example: Cocktail Party Problem



# Example data



# Mathematical Formulation of Independent Component Analysis

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

$$\mathbf{A} = \begin{bmatrix} a_1 & a_2 \\ b_1 & b_2 \end{bmatrix}$$

**Assumptions:**

1) The mixing matrix is invertible
2) The sources are statistically independent
3) The independents components have non-Gaussian distributions
4) Data has been centered. **x** and **s** are zero-mean vectors

**Goal:**

In this setting, the goal is of ICA is to find an unmixing matrix W that is appropriate $A^{-1}$ so that, $\hat{\mathbf{s}} \approx s$

$$\hat{\mathbf{s}} = \boldsymbol{W}\boldsymbol{x}$$

Seems impossible: Find two unknowns **A** and **s** by only observing their matrix product **x ????**

# Singular Value Decomposition(SVD)

- Divide and conquer!
- Lets focus on **A** first.

$$\mathbf{A} = \mathbf{U\Sigma V^T}$$

Where,

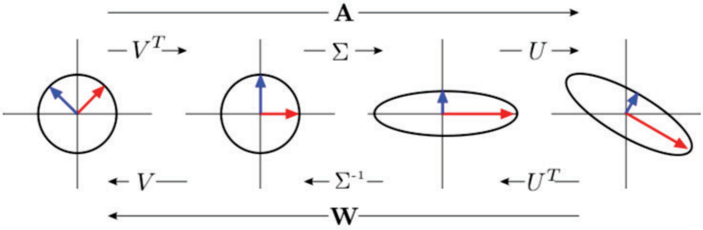
U = orthogonal matrix(eigenvector of $A^TA$)

$\boldsymbol{\Sigma}$ = Diagonal matrix non-negative diagonal entries(eigenvalue)

V = orthogonal matrix(eigenvector of $AA^{T)}$

$$\mathbf{W} = \mathbf{A}^{-1} = \mathbf{V\Sigma^{-1}U^T}$$

# Graphical Depiction of SVD



# Finding un-mixing matrix W

1. Examine the covariance of the data $\mathbf{x}$ in order to calculate $\mathbf{U}$ and $\mathbf{\Sigma}$
2. Return to the assumption of independence of $\mathbf{s}$ to solve for $\mathbf{V}$.

# Examining the covariance of the data

- As a reminder, the covariance is the expected value of the outer product of individual data points $\langle \mathbf{x}\mathbf{x}^T \rangle$

$$
\begin{aligned}
\langle \mathbf{x}\mathbf{x}^T \rangle &= \langle (\mathbf{A}\mathbf{s})(\mathbf{A}\mathbf{s})^T \rangle \\
&= \langle (\mathbf{U}\Sigma\mathbf{V}^T\mathbf{s})\,(\mathbf{U}\Sigma\mathbf{V}^T\mathbf{s})^T \rangle \\
&= \mathbf{U}\Sigma\mathbf{V}^T \langle \mathbf{s}\,\mathbf{s}^T \rangle \mathbf{V}\Sigma\mathbf{U}^T
\end{aligned}
$$

Assumption: covariance of the source **s** is *whitened* $\langle \mathbf{s}\mathbf{s}^T \rangle = \mathbf{I}$,

$$\langle \mathbf{x}\mathbf{x}^T \rangle = \mathbf{U}\Sigma^2\mathbf{U}^T.$$

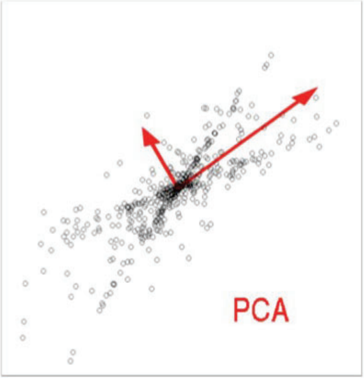
# PCA vs ICA

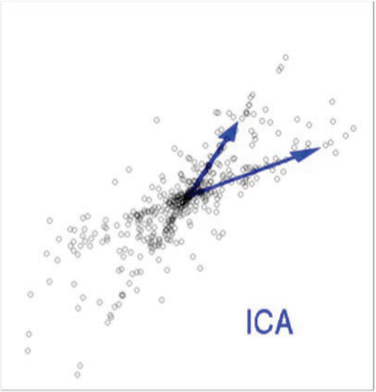
- Similarity:

1. Feature extraction
2. Dimension reduction

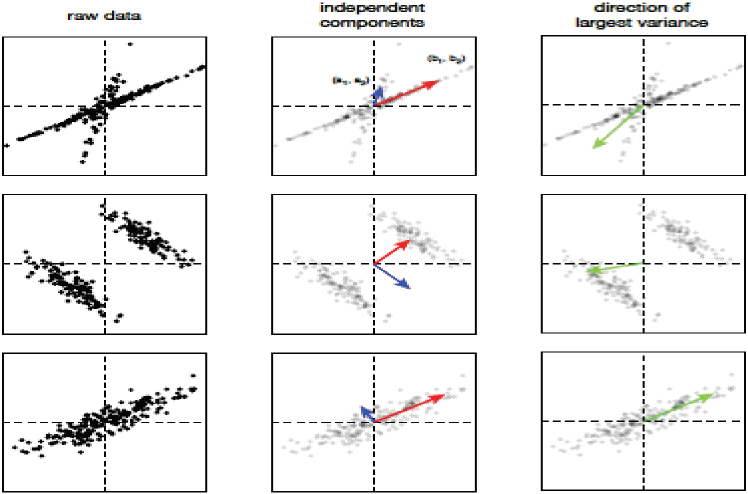
- Differences:
- PCA just removes correlations, **not** higher order dependence , ICA removes correlations, **and** higher order dependence
- Two different sound signals need not be orthogonal as in PCA, even if they are independent

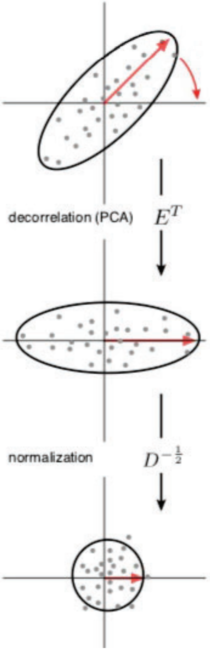
# PCA vs ICA



# PCA vs ICA


raw data
independent components
direction of largest variance




# Whitening

- Whitening is an operation that removes all linear dependencies in a data set (i.e. second-order correlations) and normalizes the variance along all dimensions.

- In our problem whitening simplifies the ICA problem down to finding a single rotation matrix V.

$$\hat{\mathbf{s}} = \mathrm{V}\boldsymbol{x}_{\mathrm{w}}$$

- Where,

$$\mathrm{x_w} = (\mathrm{D}^{-1/2}\,\mathrm{E}^{\mathrm{T}})\mathrm{x}$$

# Estimating V

- The goal of ICA can now be stated succinctly. Find a rotation matrix V such that $\hat{\mathbf{s}}$ is statistically independent.

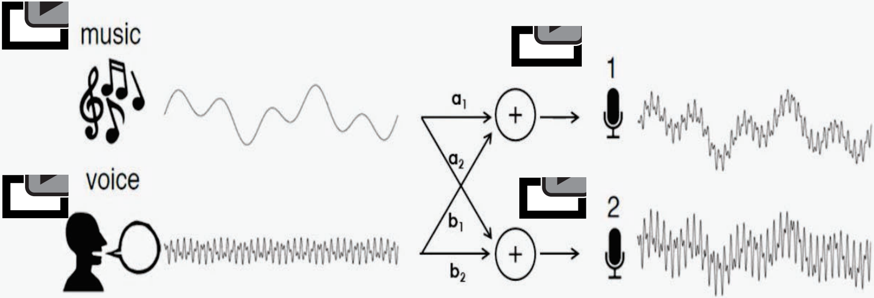
- Using an optimization technique from information theory, we estimate

$$\mathbf{V} = \arg\min_{\mathbf{V}} \sum_i H\left[(\mathbf{V}\mathbf{x}_w)_i\right]$$

# Putting it all together

- We have just found **U**, **V** and $\boldsymbol{\Sigma}$ !!
- Original formulation: **x= As**
- We needed to estimate sampled source signals using samples of observed data $\hat{\mathbf{s}} = \mathbf{W}\mathbf{x}$
- where, $\mathbf{W} = \mathbf{A}^{-1} = \mathbf{V}\boldsymbol{\Sigma}^{-1}\mathbf{U}^{\mathbf{T}}$

# Results



# Notation - Factor Analysis Model

Let's introduce the notation first. Suppose observed $\boldsymbol{X} \in \mathbb{R}^{N \times p} (N > p)$ can be decomposed to two matrices $\boldsymbol{S} \in \mathbb{R}^{N \times p}$, $A \in \mathbb{R}^{p \times p}$:

$$\boldsymbol{X} = \boldsymbol{S}\boldsymbol{A}^T$$

$$\Leftrightarrow \begin{pmatrix} \boxed{\begin{matrix} x_{11} \\ x_{21} \\ \vdots \\ x_{Np} \end{matrix}} & \boxed{\begin{matrix} x_{12} \\ x_{22} \\ \\ x_{N2} \end{matrix}} & \begin{matrix} \cdots \\ \cdots \\ \\ \cdots \end{matrix} & \boxed{\begin{matrix} x_{1p} \\ x_{2p} \\ \vdots \\ x_{Np} \end{matrix}} \end{pmatrix} = \begin{pmatrix} \boxed{\begin{matrix} s_{11} \\ s_{21} \\ \vdots \\ s_{Np} \end{matrix}} & \boxed{\begin{matrix} s_{12} \\ s_{22} \\ \\ s_{N2} \end{matrix}} & \begin{matrix} \cdots \\ \cdots \\ \\ \cdots \end{matrix} & \boxed{\begin{matrix} s_{1p} \\ s_{2p} \\ \vdots \\ s_{Np} \end{matrix}} \end{pmatrix} \begin{pmatrix} \boxed{\begin{matrix} a_{11} \\ a_{12} \\ \vdots \\ a_{1p} \end{matrix}} & \begin{matrix} \cdots \\ \cdots \\ \\ \cdots \end{matrix} & \boxed{\begin{matrix} a_{p1} \\ a_{p2} \\ \vdots \\ a_{pp} \end{matrix}} \end{pmatrix}$$

| | original $p$ variables | transformed $p$ variables | transformation |
|---|---|---|---|
| **FA:** | | *Score matrix* | *Loading matrix* |
| **ICA:** | | *Source matrix* | *Mixing matrix* |

# Notation - ICA Model

$$\boldsymbol{X} = \boldsymbol{S}\boldsymbol{A}^T$$

$$\Leftrightarrow \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & & & \vdots \\ x_{Np} & x_{N2} & \cdots & x_{Np} \end{pmatrix} = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & & & \vdots \\ s_{Np} & s_{N2} & \cdots & s_{Np} \end{pmatrix} \begin{pmatrix} a_{11} & \cdots & a_{p1} \\ a_{12} & \cdots & a_{p2} \\ \vdots & & \vdots \\ a_{1p} & \cdots & a_{pp} \end{pmatrix}$$

---

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

$$\Leftrightarrow \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix} = \begin{pmatrix} a_{11} & \cdots & a_{1p} \\ a_{21} & \cdots & a_{2p} \\ \vdots & & \vdots \\ a_{p1} & \cdots & a_{pp} \end{pmatrix} \begin{pmatrix} S_1 \\ S_2 \\ \vdots \\ S_p \end{pmatrix}$$

where $X$ and $S$ are $p$ dimensional r.vs.

# Rotation Problem

The model is unidentifiable even we constrain $S$ to be orthonormal:

$$
\begin{aligned}
X &= \boldsymbol{A}S \\
&= \boldsymbol{A}\boldsymbol{R}^T\boldsymbol{R}S \\
&= \boldsymbol{A}^*S^*
\end{aligned}
$$

where $\boldsymbol{R}$ is any orthonormal matrices.

- *Quartimax:* To maximize the sum of all loadings raised to power 4 in $S$. It thus minimizes the number of factors needed to explain a variable.
- *Varimax:* To maximize variance of the squared loadings in each factor in S. As the result, each factor has only few variables with large loadings by the factor.
- ......

# Rotation - ICA Assumption

- The starting point for ICA is the very simple assumption that the components $S_i$ are statistically independent.
- It will be shown that we must also assume that the independent component must have non-Gaussian distributions. However, in the basic model we do not assume these distributions known (if they are known, the problem is considerably simplified.)
- Then, after estimating the matrix $\boldsymbol{A}$, we can compute its inverse, say $\boldsymbol{W}$, and obtain the independent component simply by:

$$S = \boldsymbol{A}^{-1}X = \boldsymbol{W}X.$$

# Ambiguities of ICA

1. The variances of the independent components $S_i$ cannot be determined.
   - Since both $S$ and $\boldsymbol{A}$ are unknown, any scalar multiplier of source $S_i$ can be cancelled by dividing the corresponding column of $\boldsymbol{A}$ with the same scalar value.
   - The most natural way to assume that each 'source' has unit variance $\text{Var}(S_i^2) = 1$.
2. The order of the independent components cannot be determined.
   - Again, since $S$ and $\boldsymbol{A}$ are unknown, order of the terms in the model can be changed freely, and we can call any of the independent components the first one.

# Model Estimation

- Distribution of a sum of independent random variables tends toward a Gaussian distribution.
- Thus, a sum of two independent random variables usually has a distribution that is closer to gaussian than any of the two original random variables.

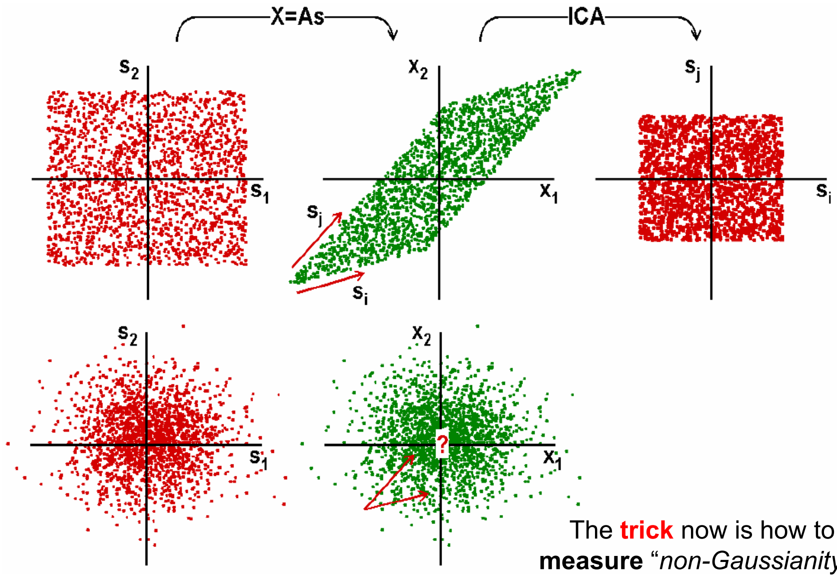
**Problem Formulation**

To find a linear mapping $\boldsymbol{W}$ such that the unmixed sequences $u$,

$$u = W^T X$$

where $W$ is a row of $\boldsymbol{W}$, are maximally statistically independent.

# Necessary Non-Gaussianity




The **trick** now is how to **measure** "*non-Gaussianity*"?

# Measurement of Non-Gaussianity

- *Kurtosis*
  - Defined by: $\mathrm{kurt}(y) = \mathrm{E}(y^4) - 3(\mathrm{E}(y^2))^2$.
  - The kurtosis for a Gaussian is zero.
  - Kurtosis is very sensitive to outliers when its value has to be estimated from a measured sample.
- *Differential entropy*
  - Defined by: $H(y) = -\int g(y) \ln g(y)\, dy$ where $g(y)$ is the density of $y$.
  - The Gaussian random variable has the largest entropy among all random variables of equal variance, which means that entropy can be used to measure non-gaussianity.
- *Negentropy*
  - Defined by: $J(y) = H(z) - H(y)$ where $z$ is a normally distributed variable.

# Approximation of Negentropy

- *Classical approximation:*

$$J(y) = \frac{1}{12}\,\mathrm{E}(y^3)^2 + \frac{1}{48}\,\mathrm{kurt}(y)^2$$

- *New approximation based on maximum-entropy principle:*

$$J(y) = \sum_{i=1}^{p} k_i\,[\mathrm{E}\{G_i(y)\} - \mathrm{E}\{\mathrm{G_i(z)}\}]^2$$

where $k_i$ are some positive constants, and $z$ is a Gaussian variable of zero mean and unit variance.The functions $G_i$ are some nonquadratic functions such as:

$$G(u) = \frac{1}{a}\ln\cosh(au), \quad G(u) = -e^{-\frac{u^2}{2}}, \; \dots \qquad (1 \le a \le 2)$$

# Mutual Information

- Defined by:

$$I(y_1, y_2, \cdots y_p) = \sum_{i=1}^{p} H(y_i) - H(y)$$

where $y_i$ are the components of the random variable $y$.

- It is the natural measure of the dependence between random variables. Its value is always nonnegative, and zero if and only if the variables are statistically dependent.
- Let $u = \boldsymbol{W}X$, then

$$I(u_1, u_2, \cdots, u_p) = \sum_{i=1}^{p} H(u_i) - H(X) - \ln|\det(\boldsymbol{W})|.$$

This is equivalent to minimizing the sum of the entropies of the separate components of $u$.

# Use of Negentropy

- Recall *Negentropy*:
  Defined by: $J(y) = H(z) - H(y)$ where $z$ is a normally distributed. It has many approximation forms.

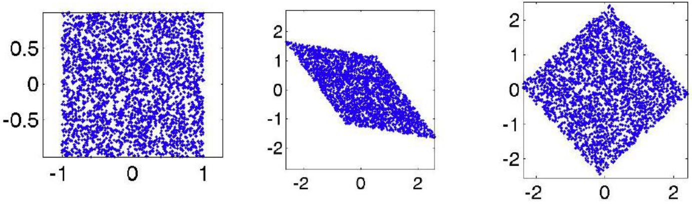
- Entropy and negentropy differ only by a constant and sign. Therefore, finding an invertible transformation $\boldsymbol{W}$ that minimizes the mutual information is roughly equivalent to finding directions in which negentropy is maximized.

# Preprocessing for ICA - Centering

- The most basic and necessary preprocessing is to center the data matrix $\boldsymbol{X}$ that is, subtract the mean vector, to make the data a zero mean variable. With this, $S$ can be considered to be zero mean, as well.
- After estimating the mixing matrix $A$ the mean vector of $S$ can be added back to the centered estimates of $S$ to complete the estimation.
- The mean vector of $S$ is given by $\boldsymbol{A}^{-1}\mu$ where $\mu$ is the mean vector of the data matrix $\boldsymbol{X}$.

# Preprocessing for ICA - Whitening

- By eigen-value decomposition (EVD): $\mathrm{E}(\boldsymbol{X}\boldsymbol{X}^T) = \boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^T$, where $\mathbf{V}$ is the orthogonal matrix of eigenvectors and $\mathbf{D}$ is the diagonal matrix of eigenvalues.
- transformation of $\boldsymbol{X}$: $\tilde{\boldsymbol{X}} = \boldsymbol{V}\boldsymbol{D}^{-1/2}\boldsymbol{V}^T\boldsymbol{X}$, then $\mathrm{Cov}(\tilde{\boldsymbol{X}}) = \boldsymbol{I}$.
- Since $\boldsymbol{S}$ has covariance $\mathbf{I}$, then for $\tilde{\boldsymbol{X}} = \tilde{\boldsymbol{A}}\boldsymbol{S}$, we have $\boldsymbol{A}$ is orthogonal which only contains $p(p-1)/2$ degrees of freedom instead of $p^2$.




original mixed whitened

# Preprocessing for ICA - Sphering

- Centering + Whitening = Sphering
- Sphering removes the first and second order statistics of the data. Both the mean and covariance are set to zero and the variance are equalized.
- Because it is a very simple and standard procedure, much simpler than any ICA algorithms, it is a good idea to reduce the complexity of the problem this way.
- We look at the eigenvalues of $\mathrm{E}(\boldsymbol{X}\boldsymbol{X}^T)$ and discard those that are too small. This has often the effect of reducing noise. Moreover, dimension reduction prevents over learning, which can sometimes be observed in ICA.

# A Direct Approach to ICA

Recall that many approaches to ICA , are based on minimizing the approximation on entropy.

# Product Density ICA

Independent Component by definition have a joint product density:

$$f_S(s) = \prod_{i=j}^{p} f_j(s_j)$$

And $f_j$ can be represented as a tilted Gaussian Density:

$$f_j(s_j) = \phi(s_j)e^{g_j(s_j)}$$

The log-likelihood for the observed data $X = \boldsymbol{A}S$:

$$l(\boldsymbol{A}, \{g_j\}_l^p; \boldsymbol{X}) = \sum_{i=1}^{N} \sum_{j=1}^{p} [log \phi_j(a_j^T x_i) + g_j(a_j^T x_i)]$$

# Penalized Density

$$\sum_{j=1}^{p}\left[\frac{1}{N}\sum_{i=1}^{N}[\log\phi_j(a_j^T x_i)+g_j(a_j^T x_i)]-\int\phi(t)e^{g_j(t)}dt-\lambda_j\int\{g_j^{'''}(t)\}^2(t)dt\right]$$

- The first enforces the density constraint $\int\phi(t)e^{g_j(t)}dt=1$ on any solution $\hat{g}_j$.
- The second is a roughness penalty, which guarantees that the solution $\hat{g}_j$ is a quartic-spline with knots at the oserved values of $s_{ij}=a_j^T x_i$.

# Product Density ICA Algorithm

**Algorithm 14.3** *Product Density ICA Algorithm:* `ProDenICA`

1. Initialize $\mathbf{A}$ (random Gaussian matrix followed by orthogonalization).
2. Alternate until convergence of $\mathbf{A}$:
   (a) Given $\mathbf{A}$, optimize (14.91) w.r.t. $g_j$ (separately for each $j$).
   (b) Given $g_j$, $j = 1, \dots, p$, perform one step of a fixed point algorithm towards finding the optimal $\mathbf{A}$.

# Product Density ICA Algorithm 2a)

Given $\boldsymbol{A}$ optimize log-likelihood function w.r.t. $g_j$

$$\sum_{j=1}^{p} \left[ \frac{1}{N} \sum_{i=1}^{N} [\log \phi_j(a_j^T x_i) + g_j(a_j^T x_i)] - \int \phi(t) e^{g_j(t)} dt - \lambda_j \int \{g_j^{'''}(t)\}^2 (t) dt \right]$$

For simplicity, we focus on a single coordinate:

$$\frac{1}{N} \sum_{i=1}^{N} [\log \phi(s_i) + g(s_i)] - \int \phi(t) e^{g(t)} dt - \lambda_j \int \{g^{'''}(t)\}^2 (t) dt$$

Since the function involves integral, an approximation is needed.

# Product Density ICA Algorithm 2a)

Construct a fine grid L with values $s_l^*$ in increments of $\triangle$.

$$y_l^* = \frac{\# s_i \in (s_l^* - \triangle/2, s_l^* + \triangle/2)}{N}$$

Likelihood function can be written as

$$\sum_{l=1}^{L} \left\{ y_l^* [\log \phi(s_l^*) + g(s_l^*)] - \triangle \phi(s_l^*) e^{g(s_l^*)} \right\} - \lambda_j \int \{g^{'''}(t)\}^2 (t) dt$$

# Product Density ICA Algorithm 2b)

Given $\hat{g}_j$, perform one step of fixed point algorithm towards finding the optimal A.
Optimize log-likelihood function w.r.t. $\boldsymbol{A}$ is equivalent to maximize:

$$C(A) = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{p} \hat{g}_j(a_j^T x_i)$$

$C(\boldsymbol{A})$ is the log-likelihood ratio between the fitted density and Gaussian and can be seen as a estimate of negentropy.

# 2b) Fixed Point Algorithm

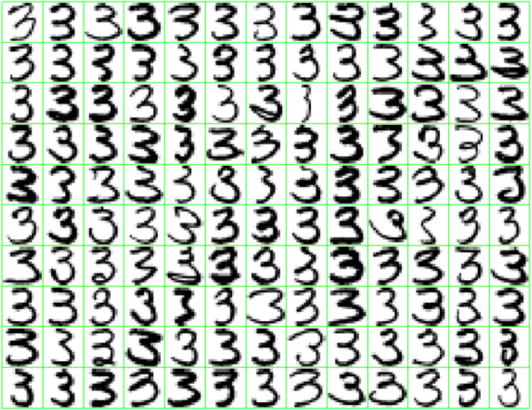
- For each $j$ update

$$a_j \leftarrow E\left\{X\hat{g}_j'(a_j^T x_i) - E[\hat{g}_j''(a_j^T x_i)]\right\}$$

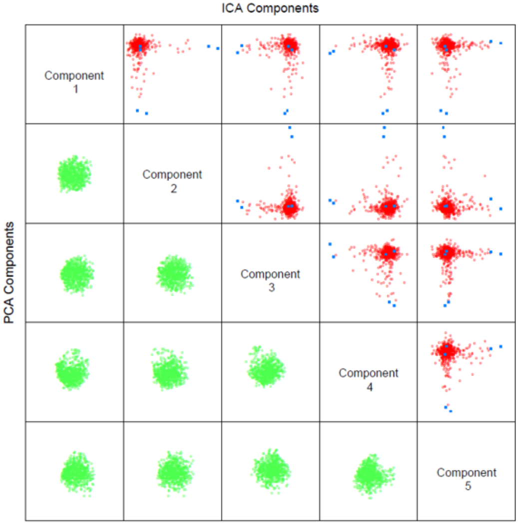
- Orthogonalize $\boldsymbol{A}$ using $(\boldsymbol{A}\boldsymbol{A}^T)^{-\frac{1}{2}}\boldsymbol{A}$
  Compute the SVD of $\boldsymbol{A}$, $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{\boldsymbol{T}}$, and then replace $\boldsymbol{A}$ with $\boldsymbol{A} \leftarrow \boldsymbol{U}\boldsymbol{V}^{\boldsymbol{T}}$

# Example: Handwritten Digits

- Digitized $16 \times 16$ grayscale images
- Points in 256-dimensional space
- High-dimensional data

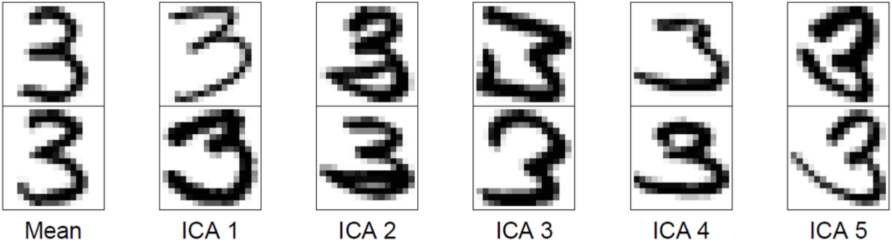


# Comparison: standardized first five ICA components vs standardized first five PCA components



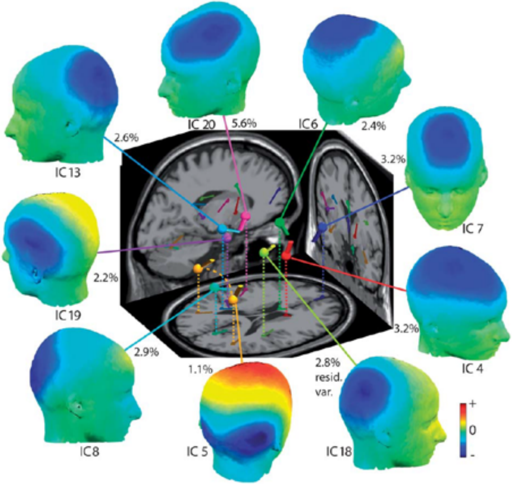
# Nature of each ICA component

- extreme digits vs central digits



# Example: EEG (electroencephalographic) Time Courses

- Domain: Brain Dynamics



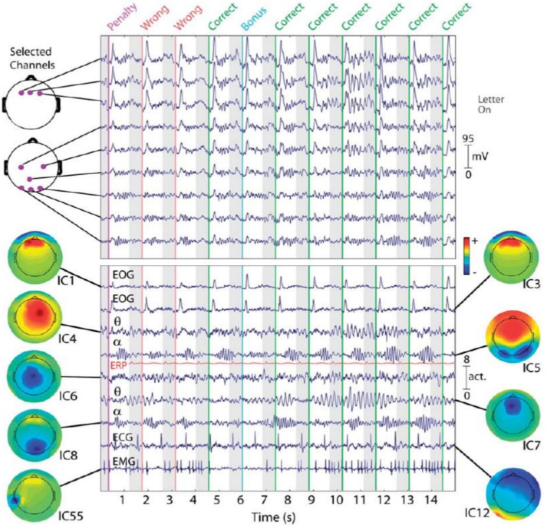
# Experiment

- Subjects wear a cap embedded with a lattice of 100 EEG electrodes, which record brain activity at different locations on the scalp.
- Data are from a single subject performing a standard "two-back" learning task over a 30-minute period, i.e., the subject is presented with a letter (B, H, J, C, F, or K) at roughly 1500-ms intervals, and responds by pressing one of two buttons to indicate whether the letter presented is the same or different from that presented two steps back.
- Depending on the answer, the subject earns or loses points, and occasionally earns bonus or loses penalty points for infrequent correct or incorrect trials, respectively.

- Goal: untangle the components of signals in multi-channel EEG data.
- Key assumption: signals recorded at each scalp electrode are a mixture of independent potentials arising from different cortical activities, as well as non-cortical artifact domains.
- ICA method

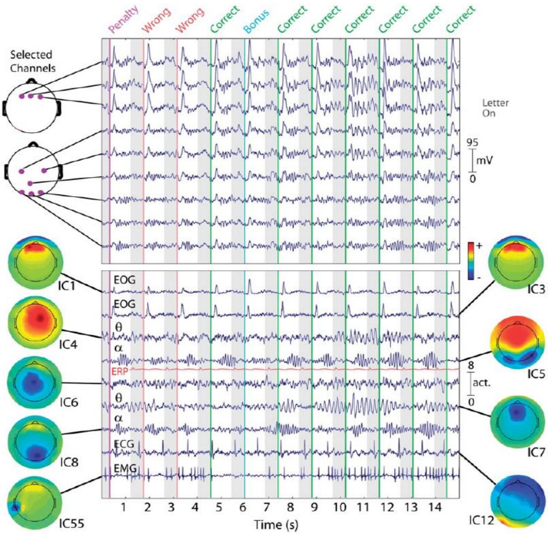
The top half:
EEG data from 9 (of 100) electrodes/channels over the course of 15 (of 1917) seconds.

The bottom half:
the “activities”/time courses of 9 (of 100) independent components during the same period.



The top half:
time-course data show spatial correlation in EEG signals – the signals of nearby sensors look very similar.

The bottom half:
ICA component activities are temporally distinct, i.e., maximally independent over time.



Thank you! :)