Independent Component Analysis (ICA): A latent variable model

- Find something "interesting" in signals:

  'cocktail party problem'

  EEG, ECG signals, FMRI data

- Feature extraction.

Generative Model

$$\mathbf{x}(t) = \mathbf{A}\mathbf{S}(t) + \textbf{ noise}$$

- $\mathbf{x} = (x_1, \ldots, x_d)$ vector of observed data (signals, images), $t =$ index
- $\mathbf{S} = (s_1, \ldots, s_m)$ vector of statistically independent latent source variables (unknown!)
- $\mathbf{A}$: $(d \times m)$ Mixing Matrix (unknown parameter !)

Goal:

Demix the signals and recover sources

$$\widehat{\mathbf{S}}(t) = \mathbf{W}\mathbf{x}(t)$$

with $\mathbf{W} = \mathbf{A}^{-1}$ for square matrices and no noise.

Ambiguities: Permutation of Sources, Scaling $s_i \to \lambda s_i$.

Some Interpretations of ICA

- $x_i(t) = \sum_j A_{ij} s_j(t)$

  $x_i(t)$ is signal at sensor i & $s_j(t)$ speaker j at time t.

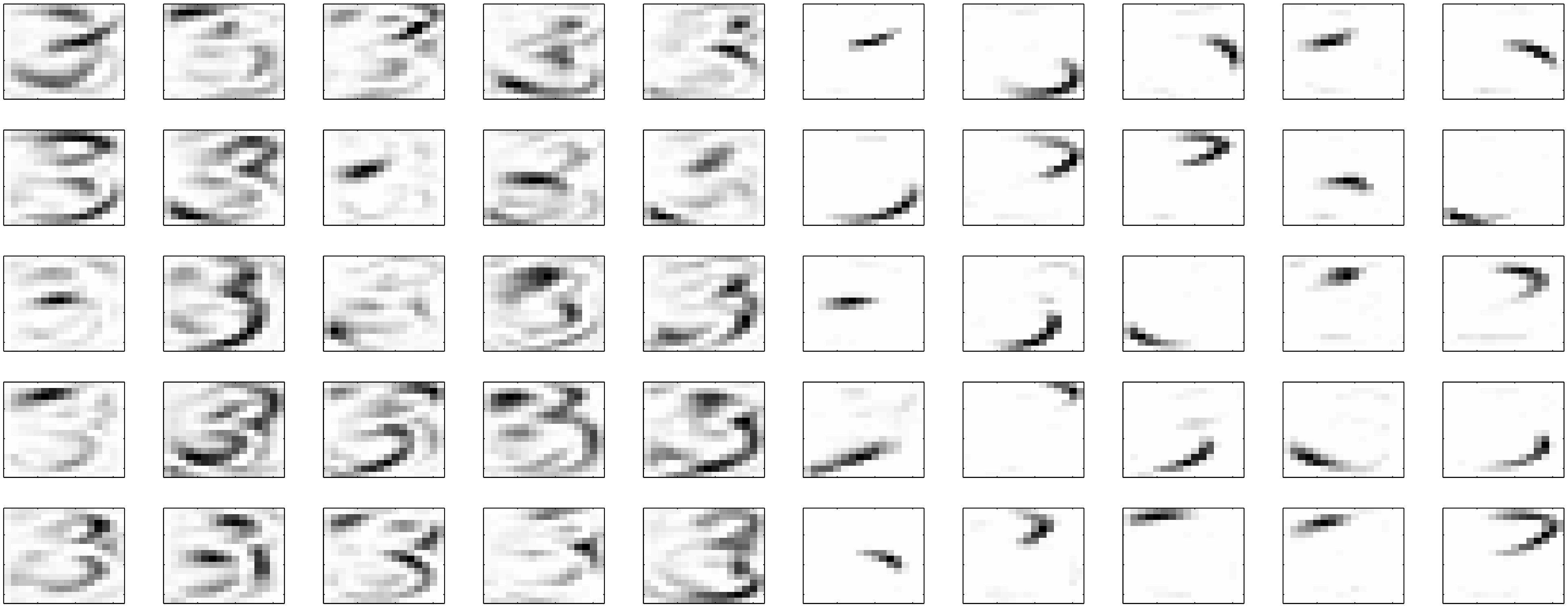
- $x_i(t) = \sum_j A_{ij} s_j(t)$

  Vector $x_i(t)$ of pixel intensities of image t is expanded into features $\mathbf{A}_{\bullet j}$ and the $s_j(t)$ are the statistically independent coefficients.

- $x_t(i) = \sum_j A_{tj} s_j(i)$

  $x_t(i)$ intensity of each pixel i at time t is a time dependent mixture of time independent activity pattern $s_j(i)$.

Feature Extraction

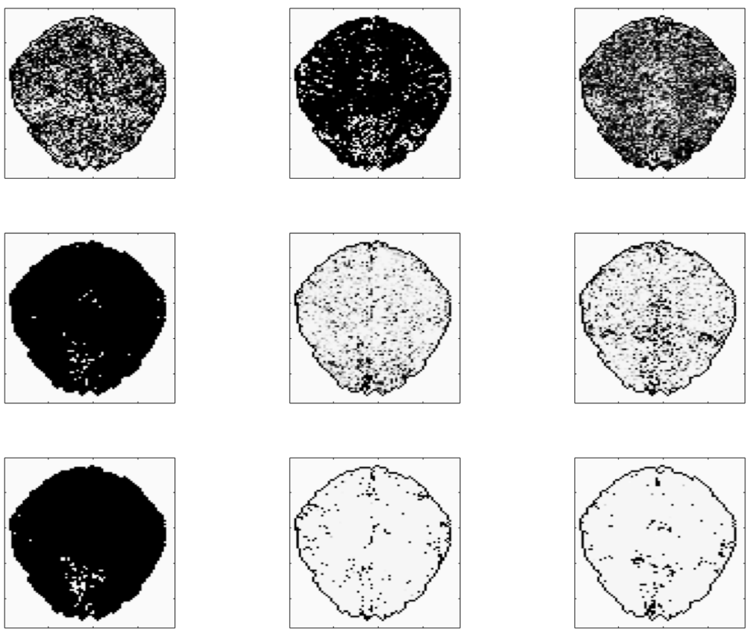


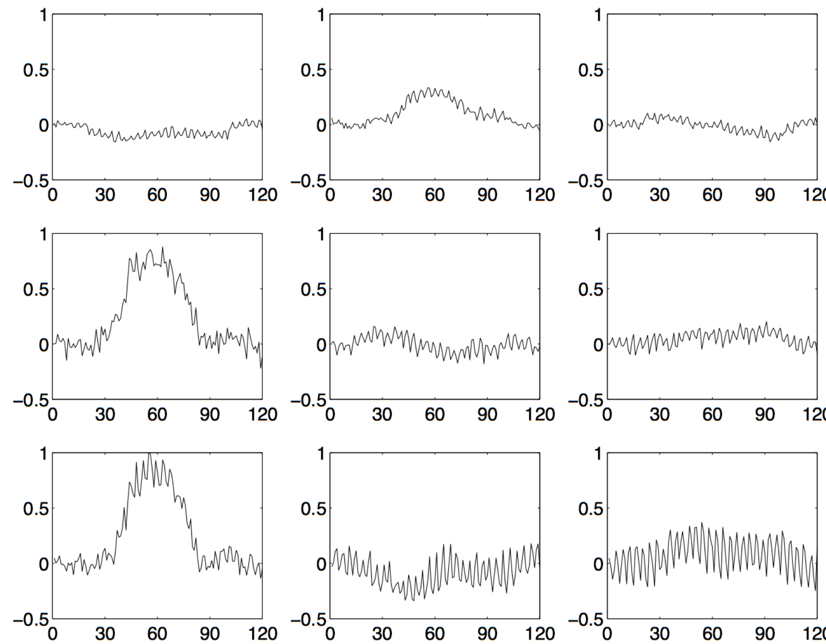
left: unconstrained **right:** constrained (positive) mixing matrix $\mathbf{A}$.

$x_i(t) =$ sequence of 500 images (handwritten '3's).$p(s) = e^{-s}$, $s \geq 0$.
Shown are the $m = 25$ columns $A_{\bullet j}$ of the matrix $\mathbf{A}$.

Functional Magnetic Resonance Imaging (fMRI) from: Højen–Sørensen, Hansen & Winther.

left: Posterior mean sources **right:** responses $A_{\bullet i}$ for $i = 1, \ldots, 9$.





Computing the Likelihood

Assume no noise and $d = m$

- Assume all n data are independent (no temporal structure):

$$p(D|\mathbf{A}) = \prod_{t=1}^{n} p(\mathbf{x}(t)|\mathbf{A})$$

- Look at a single data point: $p(\mathbf{x}|\mathbf{A}) = \int d\mathbf{S}\, p(\mathbf{x}|\mathbf{A}, \mathbf{S})\, p(\mathbf{S})$

  with $p(\mathbf{S}) = \prod_{i=1}^{d} p_i(s_i)$ (ICA assumption) and

  $p(\mathbf{x}|\mathbf{A}, \mathbf{S}) = \prod_{k=1}^{d} \delta\left(x_k - (\mathbf{AS})_k\right)$ Dirac - δ distributions (i.e. no noise).

The Likelihood cont'd

$p(\mathbf{x}|\mathbf{A}) = \int d\mathbf{S}\, p(\mathbf{x}|\mathbf{A},\mathbf{S})\, p(\mathbf{S}) = \frac{1}{|\det \mathbf{A}|} \prod_{i=1}^{d} p_i((\mathbf{A}^{-1}\mathbf{x})_i)$

With $\mathbf{W} = \mathbf{A}^{-1}$, we get for the negative log-likelihood

$-\ln p(D|\mathbf{W}) = -n \ln |\det \mathbf{W}| - \sum_t \sum_i \ln p_i((\mathbf{W}\mathbf{x}(t))_i)$

which must be minimized with respect to the matrix $\mathbf{W}$.

Modeling the sources

- Relation to PCA

  Let $\mathbf{U}$ matrix of eigenvectors of covariance matrix, i.e. $\mathbf{\Sigma U} = \mathbf{U}\Lambda$. If we set $\mathbf{W} = \mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{U}^T$, then the vector

  $\mathbf{Wx} \doteq \mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{U}^T\mathbf{x}$ has decorrelated components with unit variance.

  For Gaussian signals: decorrelated = independent!

  BUT any $\mathbf{QW}$ with orthogonal $\mathbf{Q}$ (i.e. $\mathbf{QQ}^T = \mathbf{I}$) will also decorrelate the signal: Estimation of "true" mixing matrix impossible for Gaussian signals/sources. Rotating a spherical Gaussian doesn't change its shape!

- Hence, *assume* non-Gaussian sources like e.g.

  the **super–Gaussian** $p_i(s) \propto \frac{1}{e^s + e^{-s}}$.

Disadvantages of Simple Model

- Noise ?
- Constraints on Mixing Matrix (positivity) ?
- Number of sources $\neq$ number of sensors ?
- How many sources are enough ?

Other approaches I: Minimize Mutual Information

Goal: Find $\mathbf{W}$ such that $\mathbf{S} \doteq \mathbf{W}\mathbf{x}$ has independent components.

Minimize Mutual information

$$I = \int d\mathbf{S}\, p(\mathbf{S}) \ln \frac{p(\mathbf{S})}{\prod_{i=1}^{m} p_i(s_i)}$$

with respect to $\mathbf{W}$. Problem: Find good estimate for I from data sample $\mathbf{x}(1), \mathbf{x}(2), \ldots, \mathbf{x}(T)$.

Practical Solutions:

- Approximate I using low order cumulants.
- Assume source model, eg $p(s) = \frac{1}{\pi \cosh(s)}$ - equivalent to Maximum Likelihood (Bell & Sejnowski, Cardoso & Laheld, MacKay)

Other approaches II: Non $-$ Gaussianity

Mixing sources $\sim \sum$ of independent random variables $\sim$ Gaussian distribution.

Demixing Make distribution $p(S)$ of $S \doteq \mathbf{W}\mathbf{x} = \mathbf{W}\mathbf{A}\mathbf{x}$ as non Gaussian as possible!

Possible 'contrast functions' for Minimization

- Higher Cumulants such as kurt(s) $\doteq E[s^4] - 3(E[s^2])^2$ (Hyvärinen's *FastICA*)

- 'Negentropy': $H_{Gauss} - H[\mathbf{S}]$.

More approaches

- Use temporal structure
- Kernel ICA
- ...

The ICA model was an example of

Latent Variable Models

- Simple models (like exponential families) allow for simple analytic parameter estimation by Maximum Likelihood.

- More complex models explain data by hidden (unobserved) variables, the so called latent variables. Such models are very useful in practice.

- However, even Maximum Likelihood (ML) estimation can become a hard computational task.

Overview

- Latent variable models: Definition
- Examples
- ML with the EM Algorithm

Latent variable Models: Definition

$\mathbf{y}$ = observed variables.

$\boldsymbol{\theta} = (\boldsymbol{\theta}_{\mathbf{y}}, \boldsymbol{\theta}_{\mathbf{x}})$ sets of parameters.

$\mathbf{x}$ = latent, unobserved variables.

Total likelihood

$$p(\mathbf{y}|\boldsymbol{\theta}) = \sum_{\mathbf{x}} p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}_{\mathbf{y}}) p(\mathbf{x}|\boldsymbol{\theta}_{\mathbf{x}})$$

If the $\mathbf{x}$'s would be known, ML would often be easy!

Example I: Mixtures of Gaussians

Model for multimodal densities

$$
\begin{aligned}
p(y|\{\mu_c, \sigma_c, p(c)\}_{c=1}^{K}) &= \sum_c p(c) \frac{1}{\sqrt{2\pi\sigma_c^2}} \exp\left[-\frac{(y-\mu_c)^2}{2\sigma_k^2}\right] \\
&\equiv \sum_c p(c) p(y|c, \boldsymbol{\theta})
\end{aligned}
$$

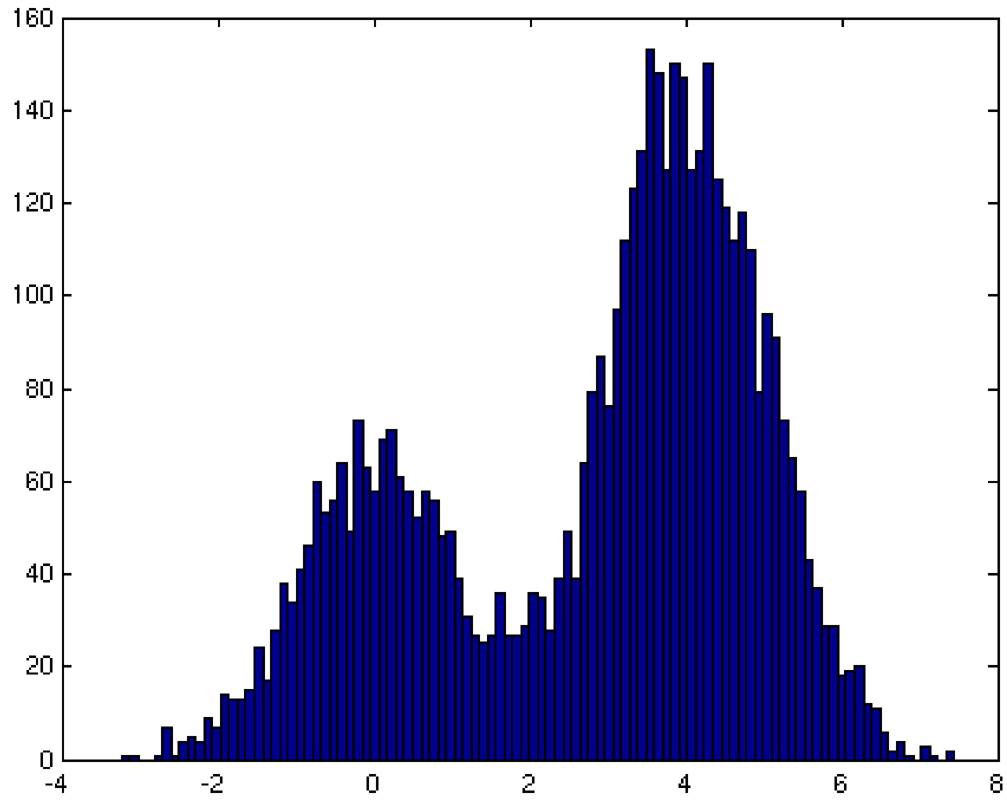
Total likelihood $p(D|\boldsymbol{\theta}) = \prod_i p(y_i|\boldsymbol{\theta})$

y_i observed, component c_i hidden,

$\boldsymbol{\theta} = \{\mu_c, \sigma_c, p(c)\}_{c=1}^{K}$ parameters to be estimated by ML.

Take $\nabla_{\boldsymbol{\theta}} \ln p(D|\boldsymbol{\theta}) = 0$ results in complicated set of nonlinear equations.

Data from a mixture of 2 Gaussians



Example II: Hidden Markov Models

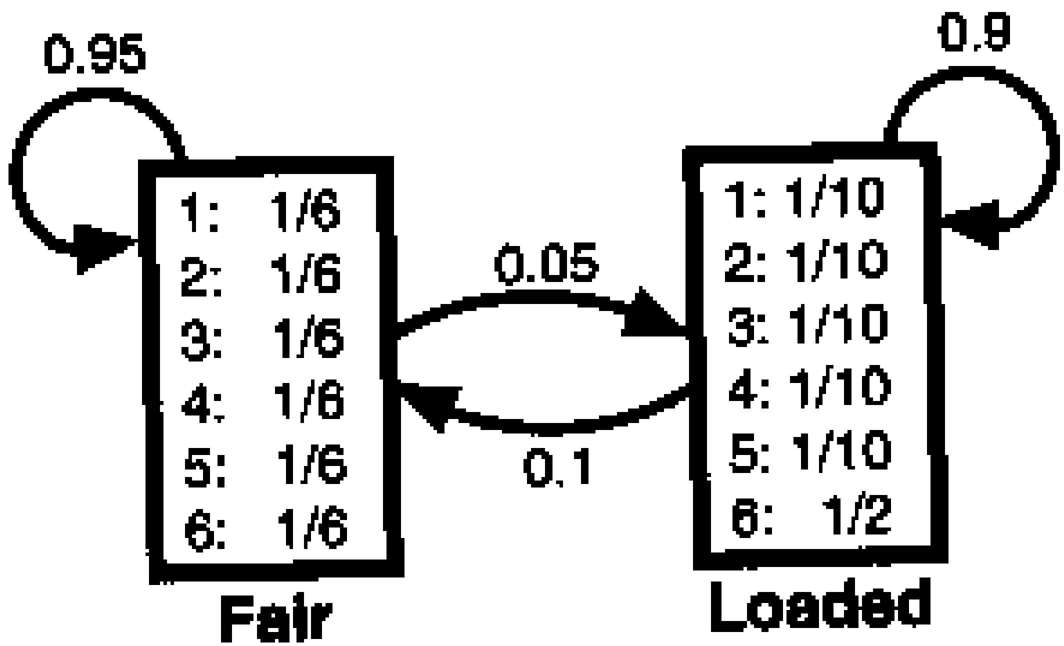
Modelling dependencies in one dimensional data structures, eg

- Speech recognition (Word models etc)

- Biosequences (DNA, proteins)

Example: The occasional dishonest casino (Durbin et al)

The HMM



Hidden Markov Models: Definitions

- Observations $\mathbf{y} = (y_1, y_2, \ldots, y_T)$ are *independent given* the sequence of states $\mathbf{S} = (s_1, s_2, \ldots, s_T)$. ie

$$P(\mathbf{y}|\mathbf{S}) = \prod_{i=1}^{T} P(y_i|s_i) = \prod_{i=1}^{T} b_{s_i}(y_i)$$

  with the matrix of *emission probabilities* $b_k(l) = P(y = l|s = k)$.

- States are not observed (hidden) and generated from a *Markov chain*

$$P(\mathbf{S}) = \pi_{s_1} P(s_2|s_1) P(s_3|s_2) \ldots P(s_T|s_{T-1}) \ .$$

- The total probability of the observed sequences is obtained by marginalization of the joint probability $P(\mathbf{y}, \mathbf{S}) = P(\mathbf{y}|\mathbf{S})P(\mathbf{S})$ over the states

$$P(\mathbf{y}) = \sum_{\mathbf{S}} P(\mathbf{y}|\mathbf{S}) P(\mathbf{S})$$

  For N states, there are N^T different paths in the sum!!