EE 6885 Statistical Pattern Recognition

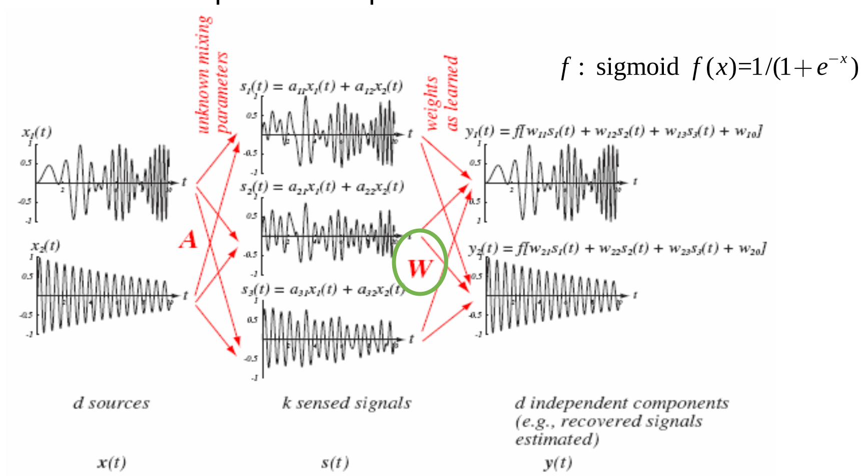
Fall 2005
Prof. Shih-Fu Chang
http://www.ee.columbia.edu/~sfchang

Lecture 19 (11/30/05)

- Topics
  - Feature Dimension Reduction
    - ICA, LDA, MDS 10.13
    - ICA Tutorial:
      - **Aapo Hyvärinen and Erkki Oja, "Independent Component Analysis: Algorithms and Applications,"** ***Neural Networks***, **13(4-5):411-430, 2000**
  - Review of AdaBoost Error Bound (HW#7 P.2)
    - Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," In Computational Learning Theory: Eurocolt '95, pages 23–37. Springer-Verlag, 1995.
- Final Exam
  - Dec. 16th Friday 1:10-3 pm, Mudd Rm 644

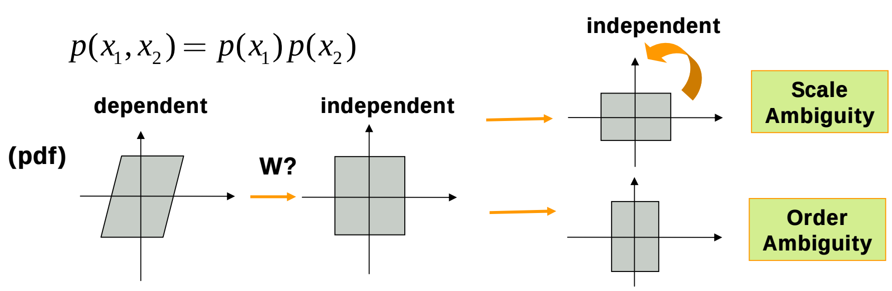
Independent Component Analysis

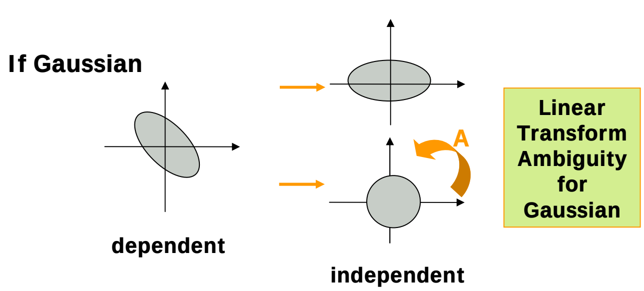
- Seek most independent directions, instead of minimize representation errors (sum-squared-error) as in PCA
- Blind source separation in speech mixture



Independence and Ambiguity in ICA

$$p(x_1, x_2) = p(x_1)\,p(x_2)$$

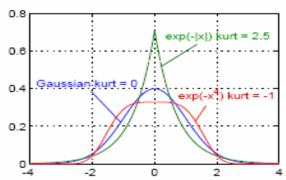




- **Therefore, ICA does not work for Gaussian source components**
- **Focus on non-Gaussianity**
- **Contrary to conventional models**

- Find the best weights to make the output components independent
- How to measure independence?
  - Linear combination of random variables leads to Normal distribution
  - Use the high-order statistics to measure Non-Gaussianity
  - Gradient Decent to weights for discovering each component
    - **Measures of deviations from Gaussianity: 4th moment is Kurtosis ("bulging")**

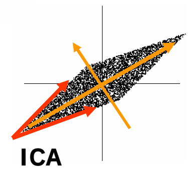
$$kurt(y) = E\left[\left(\frac{y-\mu}{\sigma}\right)^4\right] - 3$$



(from Ellis)

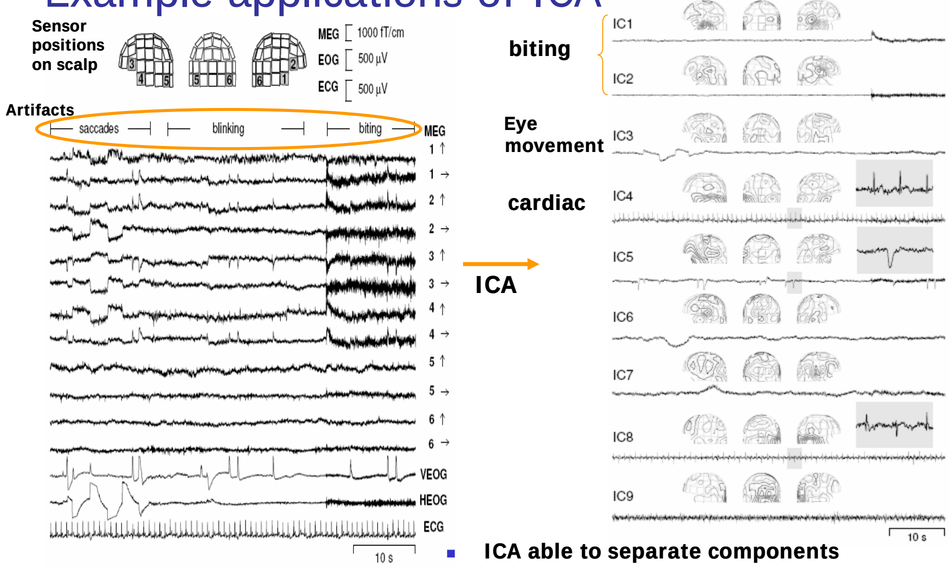
-kurtosis of Gaussian is zero (this def.)
-'heavy tails' $\rightarrow kurt > 0$
-closer to uniform dist. $\rightarrow kurt < 0$

•Directly related to KL divergence from Gaussian PDF



- FastICA Matlab package :
  http://www.cis.hut.fi/projects/ica/fastica/

Example applications of ICA

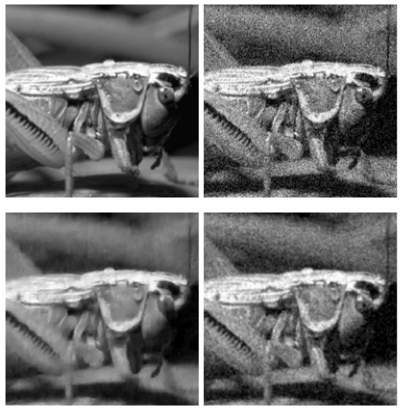


122 MEG signal streams

- **ICA able to separate components corresponding to different artifacts**
- **Note PCA can be applied before ICA**

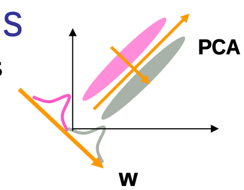
Applications of ICA

- **Noise reduction, image restoration**



- **Assume the statistics of image and noise are independent**

LDA: Linear Discriminant Analysis



Given a set of data $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, and their class labels
Find the best projection dimension, $y_i = \mathbf{w}^t \mathbf{x}_i$
so that y_i are most separable

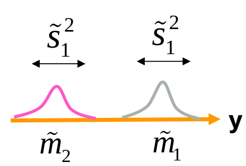
$$\tilde{m}_i = \frac{1}{n_i} \sum_{\mathbf{x} \in D_i} \mathbf{w}^t \mathbf{x} = \mathbf{w}^t \mathbf{m}_i$$

$\mathbf{m}_i$: sample means
$\tilde{m}_i$: sample means of projected points

$$\tilde{s}_i^2 = \frac{1}{n_i} \sum_{\mathbf{y} \in Y_i} (y - \tilde{m}_i)^2$$

$\tilde{s}_1^2 + \tilde{s}_2^2$: within-class scatter

LDA maximizes criterion function: $J(\mathbf{w}) = \dfrac{|\tilde{m}_1 - \tilde{m}_2|^2}{\tilde{s}_1^2 + \tilde{s}_2^2}$



LDA Scatter Matrices

before projection: $\mathbf{S}_i = \sum_{\mathbf{x}\in D_i} (\mathbf{x}-\mathbf{m}_i)(\mathbf{x}-\mathbf{m}_i)^t$

after projecttion: $\tilde{s}_i^{\,2} = \mathbf{w}^t \mathbf{S}_i \mathbf{w}$

$$\tilde{s}_1^{\,2} + \tilde{s}_2^{\,2} = \mathbf{w}^t(\mathbf{S}_1 + \mathbf{S}_2)\mathbf{w} = \mathbf{w}^t \mathbf{S}_{\mathbf{w}} \mathbf{w}$$

$\mathbf{S}_{\mathbf{w}} = \mathbf{S}_1 + \mathbf{S}_2$: within-class scatter matrix

Similarly, between-class scatter matrix $\mathbf{S}_{\mathbf{B}} = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^t$

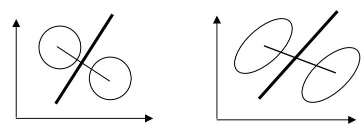
$$\Rightarrow J(\mathbf{w}) = \frac{\mathbf{w}^t \mathbf{S}_{\mathbf{B}} \mathbf{w}}{\mathbf{w}^t \mathbf{S}_{\mathbf{w}} \mathbf{w}}$$

$\mathbf{S}_{\mathbf{w}}$: usually nonsingular $\quad \mathbf{S}_{\mathbf{B}}$: rank 1

$$\Rightarrow \mathbf{w}_{opt} = \arg\max J(\mathbf{w})$$
$$= \mathbf{S}_{\mathbf{w}}^{-1}(\mathbf{m}_1 - \mathbf{m}_2)$$

Recall the Gaussian Cases

$$w = \Sigma^{-1}(\mu_i - \mu_j)$$



Mean difference vector in the PCA space

Multi-Dimensional Scaling (MDS)

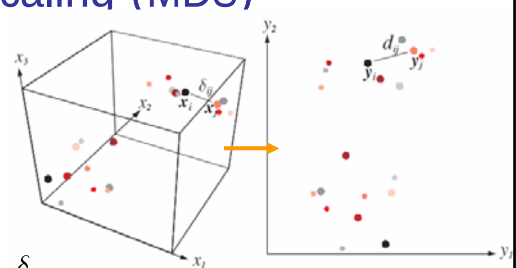
- **Visualize the data points in a lower-dim space**
- **How to preserve the original structure (e.g., distance)?**
- **Optimization Criterion**

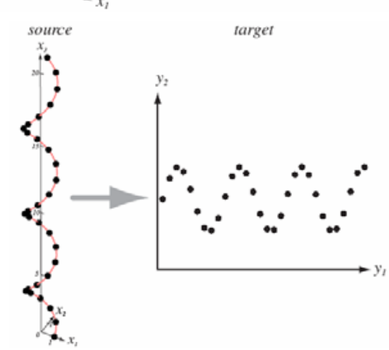
$$J_{ee} = \frac{\sum_{i<j} (d_{ij} - \delta_{ij})^2}{\sum_{i<j} \delta_{ij}^{\,2}} \qquad J_{ff} = \sum_{i<j} \left(\frac{d_{ij} - \delta_{ij}}{\delta_{ij}}\right)^2$$

- **Gradient Decent to find new locations**

$$\nabla_{\mathbf{y}_k} J_{ee} = \frac{2}{\sum_{i<j} \delta_{ij}^{\,2}} \sum_{j\neq k} \left(d_{kj} - \delta_{kj}\right) \frac{\mathbf{y}_k - \mathbf{y}_j}{d_{kj}}$$

- **Sometimes rank order is more important**





(HW #7 P.1) No Free Lunch Theorem

- Without any prior context inform[Target function]lassifi[Hypothesis]thm is better than other algorithms

	x	F	h_1 (Majority)	h_2 (Minority)
Training	000	1		
	001	-1		
	010	1		
Test	011	-1	1	-1
	100	1	1	-1
	101	-1	1	-1
	110	1	1	-1
	111	1	1	-1

- If we don't assume the statistics of labels of the test patterns, then all target functions are equally likely.
- Use this to 'prove' NFL in HW#7 P.1

HW#7 P.2

Algorithm AdaBoost

As in AdaBoost Ref.

Input: set of N labeled examples $\{(1, c(1)), \ldots, (N, c(N))\}$
distribution D over the examples
weak learning algorithm **WeakLearn**
integer T specifying number of iterations

Initialize the weight vector: $w_i^1 = D(i)$ for $i = 1, \ldots, N$

Do for $t = 1, 2, \ldots, T$

1. Set
$$\mathbf{p}^t = \frac{\mathbf{w}^t}{\sum_{i=1}^N w_i^t}$$
2. Call **WeakLearn**, providing it with the distribution $\mathbf{p}^t$; get back a hypothesis h_t.
3. Calculate the error of h_t: $\epsilon_t = \sum_{i=1}^N p_i^t |h_t(i) - c(i)|$.
4. Set $\beta_t = \epsilon_t / (1 - \epsilon_t)$.
5. Set the new weights vector to be
$$w_i^{t+1} = w_i^t \beta_t^{1 - |h_t(i) - c(i)|}$$

Final Classifier h_f

$$h_f(i) = \begin{cases} 1, & \sum_{t=1}^{T}\left(\log\frac{1}{\beta_t}\right)h_t(i) \geq \frac{1}{2}\sum_{t=1}^{T}\log\frac{1}{\beta_t} \\ 0, & \text{otherwise} \end{cases}.$$

- **When will the final classifier be incorrect?**
- **Suppose c(i)=0, then $h_f(i)$ is incorrect if**

$$\sum_{t=1}^{T}(\log\beta_t^{-1})h_t(i) \geq \frac{1}{2}\sum_{t=1}^{T}\log(\beta_t^{-1})$$

$$\text{namely } \prod_{t=1}^{T}\beta_t^{-h_t(i)} \geq \prod_{t=1}^{T}\beta_t^{-1/2} \quad \Rightarrow \quad \prod_{t=1}^{T}\beta_t^{1-h_t(i)} \geq \prod_{t=1}^{T}\beta_t^{1/2}$$

- **In general**

$$h_f(i) \text{ is incorrect if } \prod_{t=1}^{T}\beta_t^{1-|h_t(i)-c(i)|} \geq \left(\prod_{t=1}^{T}\beta_t\right)^{1/2} \qquad \times D(i)$$

$$\Rightarrow D(i)\prod_{t=1}^{T}\beta_t^{1-|h_t(i)-c(i)|} \geq D(i)\left(\prod_{t=1}^{T}\beta_t\right)^{1/2} \quad \Rightarrow \quad w_i^{T+1} \geq D(i)\left(\prod_{t=1}^{T}\beta_t\right)^{1/2}$$

?

$$h_f(i) \text{ is incorrect if } w_i^{T+1} \geq D(i)\left(\prod_{t=1}^{T}\beta_t\right)^{1/2}$$

$$\sum_{i,\, h_f(i)\neq c(i)}^{N} w_i^{T+1} \geq \sum_{i,\, h_f(i)\neq c(i)}^{N} D(i)\left(\prod_{t=1}^{T}\beta_t\right)^{1/2} = E\left(\prod_{t=1}^{T}\beta_t\right)^{1/2}$$

Ref.

Theorem 1 in Ref. $\sum_{i=1}^{N} w_i^{t+1} \leq \sum_{i=1}^{N} w_i^{t}(1-(1-\beta_t)(1-E_t))$

$$\beta_t = \frac{E_t}{1-E_t}$$

$$\sum_{i=1}^{N} w_i^{t+1} \leq \sum_{i=1}^{N} w_i^{t}(2E_t) \Rightarrow \sum_{i=1}^{N} w_i^{T+1} \leq \prod_{t=1}^{T}(2E_t)$$

? ?

$$\therefore\ E \leq \prod_{t=1}^{T}(2E_t)\Big/\left(\prod_{t=1}^{T}\beta_t\right)^{1/2} = \prod_{t=1}^{T}\left(2\sqrt{E_t(1-E_t)}\right)$$

... Fill in details to complete HW7 P.2