

LECTURE 27: Independent Components Analysis

- The “cocktail party” problem
- Definition of ICA
- Independence and uncorrelatedness
- Independence and non-Gaussianity
- Preprocessing for ICA
- The FastICA algorithm
- Non-linear ICA



The cocktail party problem (1)

- **Imagine that you are at a large party where several conversations are being held at the same time**
 - Despite the strong background, you are able to focus your attention on a specific conversation (of your choice), and ignore all others
 - At the same time, if someone were to call our name from the other name of the room, we would immediately be able to respond to it
 - How is it that we can separate different flows of information that occur at the same time and share the very same frequency bands?



The cocktail party problem (2)

■ Let's formalize this scenario

- Two people are speaking simultaneously
 - We will denote their sound pressure waveforms by $s_1(t)$ and $s_2(t)$
- Two microphones are placed at different locations
 - We will denote their recorded signals by $x_1(t)$ and $x_2(t)$
- We will assume that the recorded signals $x_i(t)$ are a linear combination of the sources $s_i(t)$

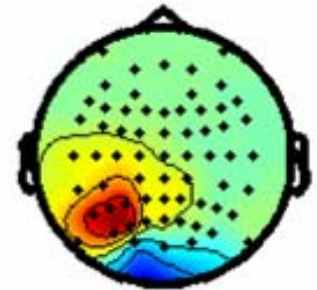
$$\begin{cases} x_1(t) = a_{11}s_1(t) + a_{12}s_2(t) \\ x_2(t) = a_{21}s_1(t) + a_{22}s_2(t) \end{cases}$$

- where the coefficients a_{ij} would depend on the relative distance between the microphones and the speakers
- Note that this is a very oversimplified model, since we are ignoring very basic phenomena such as propagation delays and reverberances in the room



The cocktail party problem (2)

- **Our goal is to find the original sources $s_i(t)$ from mixed signals $x_j(t)$**
 - Obviously, if we knew the mixing coefficients a_{ij} , the problem could be trivially solved through matrix inversion
 - But how about when a_{ij} are unknown?
- **To solve for both a_{ij} (actually, its inverse) and $s_i(t)$, we need to make further assumptions**
 - One such assumption would be that the speech waveforms of the two speakers are statistically independent, which is not that unrealistic
 - Interestingly, this simple assumption is sufficient to solve the problem and, in some cases, the assumption need not be strictly true
- **The same principles can be (and have been) used for a variety of problems**
 - Separating sources of activity in the brain from electrical (EEG) and magnetic (MEG, fMRI) recordings
 - Denoising and detrending of sensor signals
 - Finding “interesting” projections in high-dimensional data (projection pursuit)



Definition of ICA

- Assume that we observe n linear mixtures x_1, x_2, \dots, x_n , from n independent observers

$$x_j(t) = a_{j1}s_1(t) + a_{j2}s_2(t) + \dots + a_{jn}s_n(t)$$

- Or, using matrix notation

$$x = As$$

- Our goal is to find a de-mixing matrix W such that

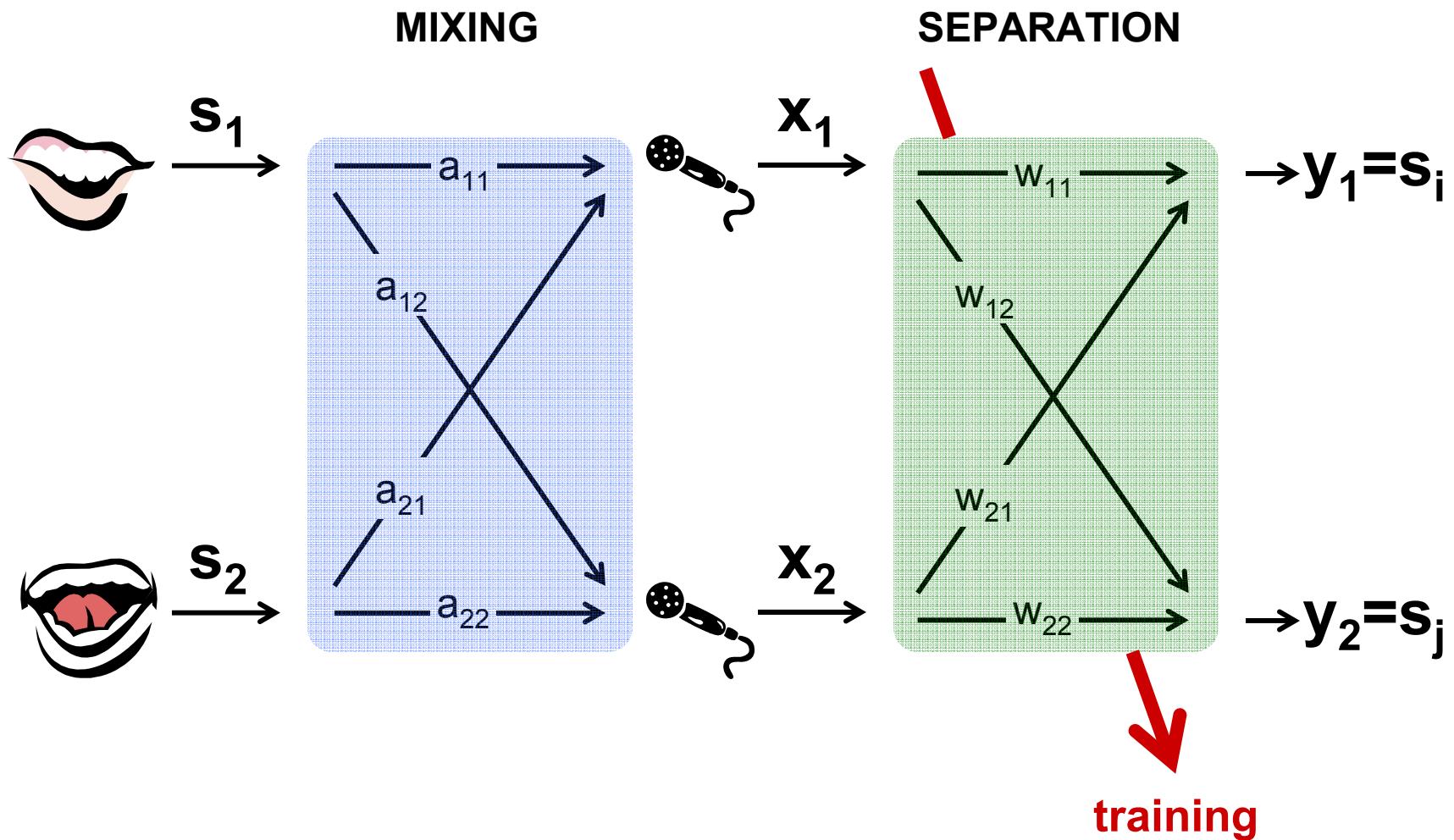
$$s = Wx$$

■ Assumptions

- Both mixture signals and source signals are zero-mean (i.e. $E[x_i] = E[s_j] = 0, \forall i, j$)
 - If not, we simply subtract their means
- The sources have non-Gaussian distributions
 - More on this in a minute
- The mixing matrix is square, i.e., there are as many sources as mixing signals
 - This assumption, however, can sometimes be relaxed

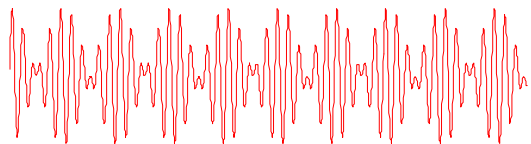


Blind source separation

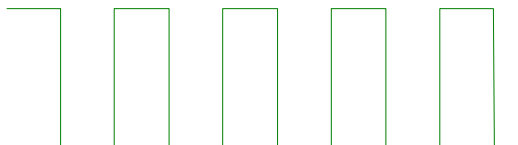


An example

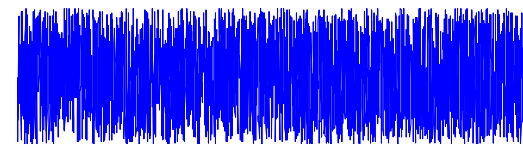
$$s_1(n) = \sin(100n)\cos(10n)$$



$$s_2 = \text{sign}(\sin(10n))$$

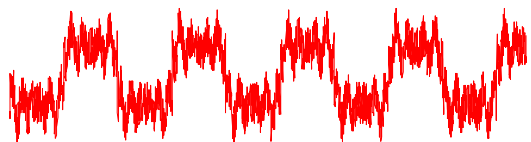


$$s_3 = \text{rand}(n)$$

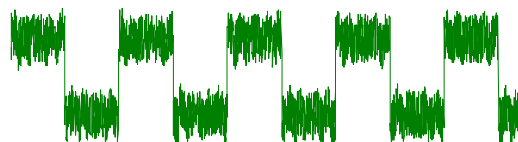


$$x = As$$

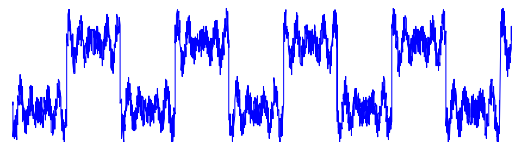
x_1



x_2

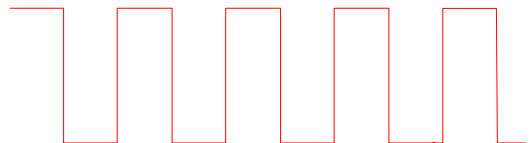


x_3

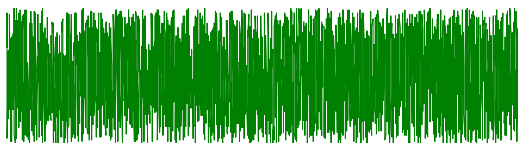


FastICA

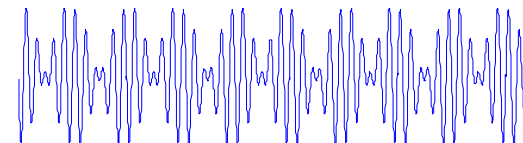
y_1



y_2



y_3



Ambiguities of ICA

■ The variance of the independent components cannot be determined

- Since both s and A are undetermined, any multiplicative factor in s , including a change of sign, could be absorbed by the coefficients of A

$$\begin{aligned}x_j(t) &= (ka_{j1})s_1(t) + (ka_{j2})s_2(t) \\ &= a_{j1}(ks_1(t)) + a_{j2}(ks_2(t))\end{aligned}$$

- To resolve this ambiguity, source signals are assumed to have unit variance

$$E[s_i^2] = 1$$

■ The order of independent components cannot be determined

- Since both s and A are unknown, any permutation of the mixing terms would yield the same result
- Compare this with Principal Components Analysis, where the order of the components can be determined by their eigenvalues (their variance)



Independence vs. uncorrelatedness

■ What is independence?

- Two random variables y_1 and y_2 are said to be independent if knowledge of the value of y_1 does not provide any information about the value of y_2 , and viceversa

$$p(y_1|y_2) = p(y_1) \Leftrightarrow p(y_1, y_2) = p(y_1)p(y_2)$$

■ What is uncorrelatedness?

- Two random variables y_1 and y_2 are said to be uncorrelated if their covariance is zero

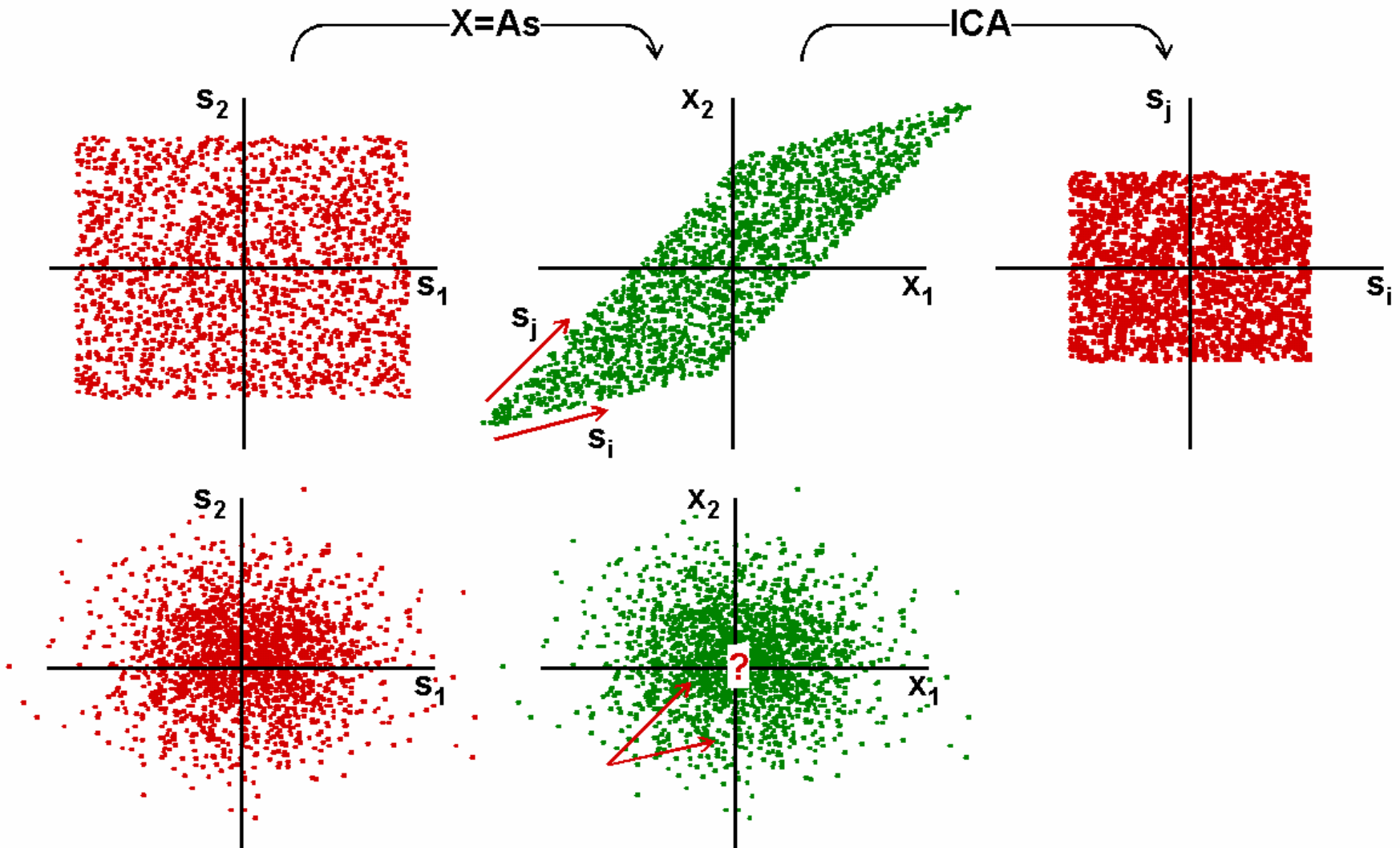
$$E[y_1^2 y_2^2] = 0$$

■ Equivalences

- Independence implies uncorrelatedness
- Uncorrelatedness DOES NOT imply independence...
 - Unless the random variables y_1 and y_2 are Gaussian, in which case uncorrelatedness and independence are equivalent
 - Note that, in this case, the covariance matrix is diagonal, and $p(x_1, x_2)$ can be trivially factorized as the product of the two univariate densities $p(x_1)$ and $p(x_2)$



Why can't Gaussian variables be used with ICA?



Independence and non-Gaussianity

- **As we have just seen, a necessary condition for ICA to work is that the signals be non-Gaussian**
 - Otherwise, ICA cannot resolve the independent directions due to symmetries
 - Besides, if signals are Gaussian, one may just use PCA to solve the problem (!)
- **We will now show that finding the independent components is equivalent to finding the directions of largest non-Gaussianity**
 - For simplicity, let us assume that all the sources have identical distributions
 - Our goal is to find the vector w such that $y=w^T x$ is equal to one of the sources s
 - We make the change of variables $z=A^T w$,
 - This leads to $y=w^T x=w^T A s=z^T s$
 - Thus, y is a linear combination of the sources s
 - According to the CLT, the signal y is more Gaussian than the sources s since it is a linear combination of them, and becomes the least Gaussian when it is equal to one of the sources
 - Therefore, the optimal w is the vector that maximizes the non-Gaussianity of $w^T x$, since this will make y equal to one of the sources
- **The trick is now how to measure “non-Gaussianity” ...**



Measures of non-Gaussianity (1)

■ Kurtosis

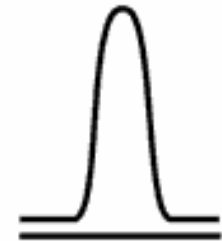
- The classical measure of non-Gaussianity is kurtosis, which is defined as the fourth order cumulant

$$\text{kurt}(y) = E[y^4] - 3(E[y^2])^2$$

- Kurtosis can be both positive or negative
 - When kurtosis is zero, the variable is Gaussian
 - Supergaussians are characterized by a “spiky” pdf with heavy tails, i.e., the Laplace pdf
 - When kurtosis is positive, the variable is said to be supergaussian or leptokurtic
 - Subgaussians are characterized by a rather “flat” pdf
- Thus, the absolute value of the kurtosis can be used as a measure of non-Gaussianity
 - Kurtosis has the advantage of being computationally cheap
 - Unfortunately, kurtosis is rather sensitive to outliers



Mesokurtic
(Normal)
 $K = 0$



Leptokurtic
 $K > 0$



Platykurtic
 $K < 0$



Measures of non-Gaussianity (2)

■ Negentropy

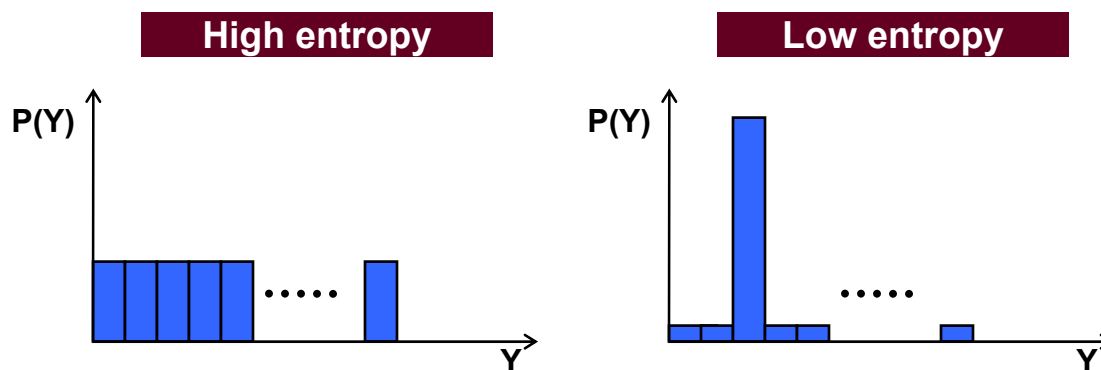
- An information-theoretic quantity of differential entropy
- The entropy of a variable can be thought of as a measure of randomness
 - For a discrete-valued variable, the entropy $H(Y)$ is defined as

$$H(Y) = -\sum_i P(Y = a_i) \log P(Y = a_i)$$

- Whereas for a continuous-valued variable, the (differential) entropy $H(y)$ is

$$H(y) = -\int p(y) \log p(y) dy$$

- A uniform variable has the largest entropy among discrete-valued variables, whereas a Gaussian has the largest entropy among continuous valued variables



Measures of non-Gaussianity (3)

■ Negentropy (continued)

- To obtain a measure of non-Gaussianity, one may then take a differential measurement of entropy relative to a Gaussian; this is known as negentropy

$$J(y) = H(y_{\text{gauss}}) - H(y)$$

- where y_{gauss} is a Gaussian variable with the same variance as y
- Note that $J(y)$ is always non-negative, and only equal to zero for a Gaussian
- Properties
 - Negentropy is statistically robust
 - Unfortunately, it is also computationally intensive, since it requires density estimation, possibly non-parametric



Measures of non-Gaussianity (4)

■ Approximations of negentropy

- Since the estimation of negentropy is difficult, one typically uses approximations proposed by Hyvarinen, which have the form

$$J(y) \propto [E\{G(y)\} - E\{G(v)\}]^2$$

- where v is a Gaussian variable $N(0,1)$, y is assumed to be zero-mean, and G is a nonquadratic function
- Several choices of G have been shown to work well, including:

$$G_1(u) = \frac{1}{a_1} \log \cosh(a_1 u); \quad 1 \leq a_1 \leq 2$$

$$G_2(u) = -\exp(-u^2/2)$$

- These approximations have several advantages
 - Even in cases when they are not accurate, they are consistent (always non-negative and zero if y is Gaussian)
 - They provide a tradeoff between the properties of the two classical measures (kurtosis and negentropy)
 - They are computationally simple, fast to compute, and have good statistical properties (robustness)



Preprocessing for ICA (1)

- The computation of independent components can be made simpler and better conditioned if the data is preprocessed prior to the analysis

- **Centering**

- This step consists of subtracting the mean of the observation vector

$$\mathbf{x}' = \mathbf{x} - \mathbf{E}[\mathbf{x}]$$

- The mean vector can be added to the estimates of the sources afterwards

$$\mathbf{s} = \mathbf{s}' + \mathbf{A}^{-1}\mathbf{E}[\mathbf{x}]$$

- **Whitening**

- Whitening consists of applying a linear transform to the observations so that its components are uncorrelated and have unit variance

$$\tilde{\mathbf{x}} = \mathbf{W}\mathbf{x} \Rightarrow \mathbf{E}[\tilde{\mathbf{x}}\tilde{\mathbf{x}}^T] = \mathbf{I}$$

- This can be achieved through principal components

$$\tilde{\mathbf{x}} = \mathbf{E}\mathbf{D}^{-1/2}\mathbf{E}^T\mathbf{x}$$

- where (the columns of) \mathbf{E} and (the diagonal of) \mathbf{D} are the are eigenvector and eigenvalues of $\mathbf{E}[\mathbf{x}\mathbf{x}^T]$, respectively

$$\mathbf{E}[\mathbf{x}\mathbf{x}^T] = \mathbf{E}\mathbf{D}\mathbf{E}^T$$



Preprocessing for ICA (2)

■ Whitening (continued)

- Note that whitening makes the mixing matrix orthogonal

$$\tilde{\mathbf{x}} = \mathbf{E}\mathbf{D}^{-1/2}\mathbf{E}^T \mathbf{x}$$

⇓

$$\tilde{\mathbf{x}} = \mathbf{E}\mathbf{D}^{-1/2}\mathbf{E}^T \mathbf{A}\mathbf{s} = \tilde{\mathbf{A}}\mathbf{s}$$

⇓

$$\underbrace{\mathbf{E}[\tilde{\mathbf{x}}\tilde{\mathbf{x}}^T]} = \tilde{\mathbf{A}}\underbrace{\mathbf{E}[\mathbf{s}\mathbf{s}^T]} \tilde{\mathbf{A}} = \tilde{\mathbf{A}}\tilde{\mathbf{A}}^T = \mathbf{I}$$

- Which has the advantage of halving the number of parameters that need to be estimated , since an orthogonal matrix only has $n(n-1)/2$ free parameters
 - For a 2 source problem, there will only be one parameter (an angle) to be estimated!
- Since computing PCA is straightforward, it is then worthwhile to reduce the complexity of the ICA problem by whitening the data



The FastICA algorithm for one unit (1)

- **FastICA is a very efficient method of maximizing the measures of non-Gaussianity mentioned earlier**

- In what follows, we will assume that the data has been centered and whitened
- We will first start with a single-unit problem, then generalize to multiple units

- **FastICA for one unit**

- The goal is to find a weight vector w that maximizes the negentropy estimate

$$J(w^T x) \propto (E[G(w^T x)] - E[G(v)])^2$$

- Note that the maxima of $J(w^T x)$ occurs at a certain optima of $E\{G(w^T x)\}$, since the second part of the estimate is independent of w
- According to the Kuhn-Tucker conditions, the optima of $E\{G(w^T x)\}$ under the constraint $E\{(w^T x)^2\} = \|w\|^2 = 1$ occurs at points where

$$F(w) = E[xg(w^T x)] - \beta w = 0$$

- where $g(u) = dG(u)/du$
 - The constraint $E\{(w^T x)^2\} = \|w\|^2 = 1$ occurs because the variance of $w^T x$ must be equal to unity (by design): if the data is pre-whitened, then the norm of w must be equal to one
- The problem can be solved as an approximation of Newton's method
 - To find a zero of $f(x)$, apply the iteration $x_{n+1} = x_n - f(x_n)/f'(x_n)$



The FastICA algorithm for one unit (2)

- Computing the Jacobian of $F(w)$ yields

$$JF(w) = E\{xx^T g'(w^T x)\} - \beta I$$

- To simplify inversion of this matrix, we approximate the first term of the expression by noting that the data is sphered

$$E[xx^T g'(w^T x)] \approx E[xx^T] E[g'(w^T x)] = \underbrace{E[g'(w^T x)]}_{\text{a scalar}} I$$

- So the Jacobian is diagonal, which simplifies the inversion
- Thus, the (approximate) Newton's iteration becomes

$$w^+ = w - (E[xg(w^T x)] - \beta w) / (E[g'(w^T x)] - \beta)$$

- This algorithm can be further simplified by multiplying both sides by $\beta - E\{g'(w^T x)\}$, which yields the FastICA iteration

- (1) Choose an initial (e.g., random) weight vector w
- (2) Let $w^+ = E\{xg(w^T x)\} - E\{g'(w^T x)\}w$
- (3) Let $w = w^+ / \|w^+\|$
- (4) If not converged, go back to (2)



The FastICA algorithm for several units (1)

- To estimate several independent components, we run the one-unit FastICA with several units w_1, w_2, \dots, w_n
 - To prevent several of these vectors from converging to the same solution, we decorrelate outputs $w_1^T x, w_2^T x, \dots, w_n^T x$ at each iteration
 - This can be done using a deflation scheme based on Gram-Schmidt
 - We estimate each independent component one by one
 - With p estimated components w_1, w_2, \dots, w_p , we run the one-unit ICA iteration for w_{p+1}
 - After each iteration, we subtract from w_{p+1} its projections $(w_{p+1}^T w_j)w_j$ on the previous vectors w_j
 - Then, we renormalize w_{p+1}

$$(1) \text{ Let } w_{p+1} = w_{p+1} - \sum_{j=1}^p w_{p+1}^T w_j w_j$$

$$(2) \text{ Let } w_{p+1} = w_{p+1} / \sqrt{w_{p+1}^T w_{p+1}}$$

- Or, if all components must be computed simultaneously (to avoid asymmetries), the following iteration proposed by Hyvarinen can be used

$$(1) \text{ Let } W = W / \sqrt{\|WW^T\|}$$

$$(2) \text{ Repeat until convergence } W = \frac{3}{2}W - \frac{1}{2}WW^T W$$

