Lecture 5: Independent Components Analysis

Problem Setting: “Cocktail Party Problem”

Sound sources

$$s$$

Sound sources are mixed in microphones

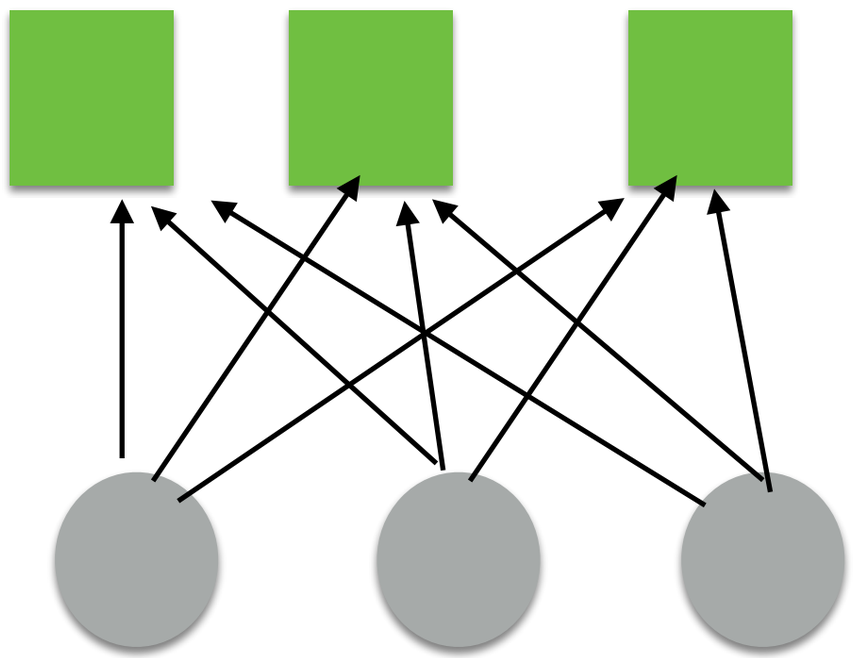
$$x = As$$

Want to un-mix them

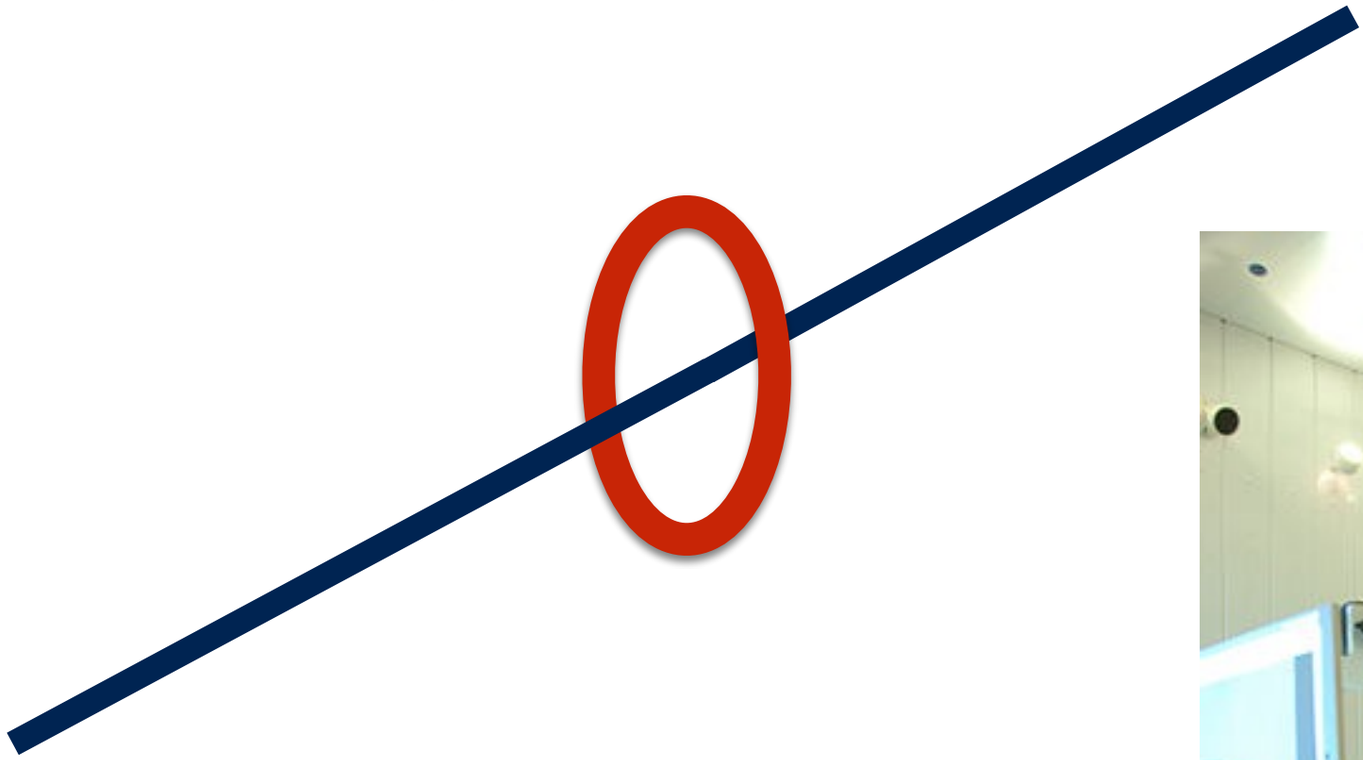
$$\hat{s} = Wx$$

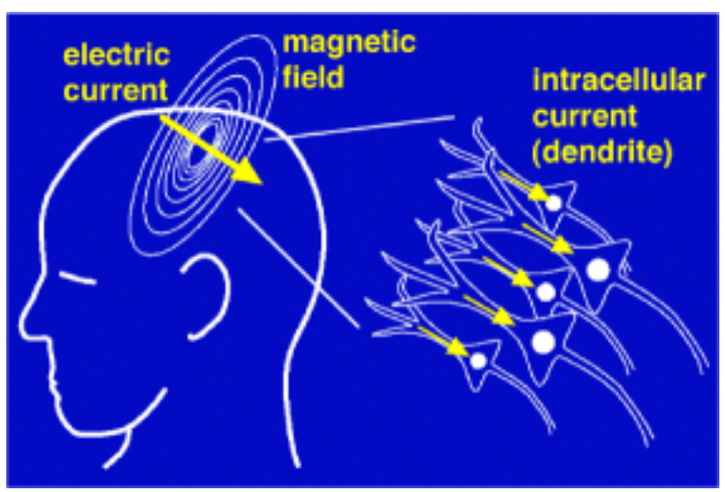
Desired answer

$$W = A^{-1}$$

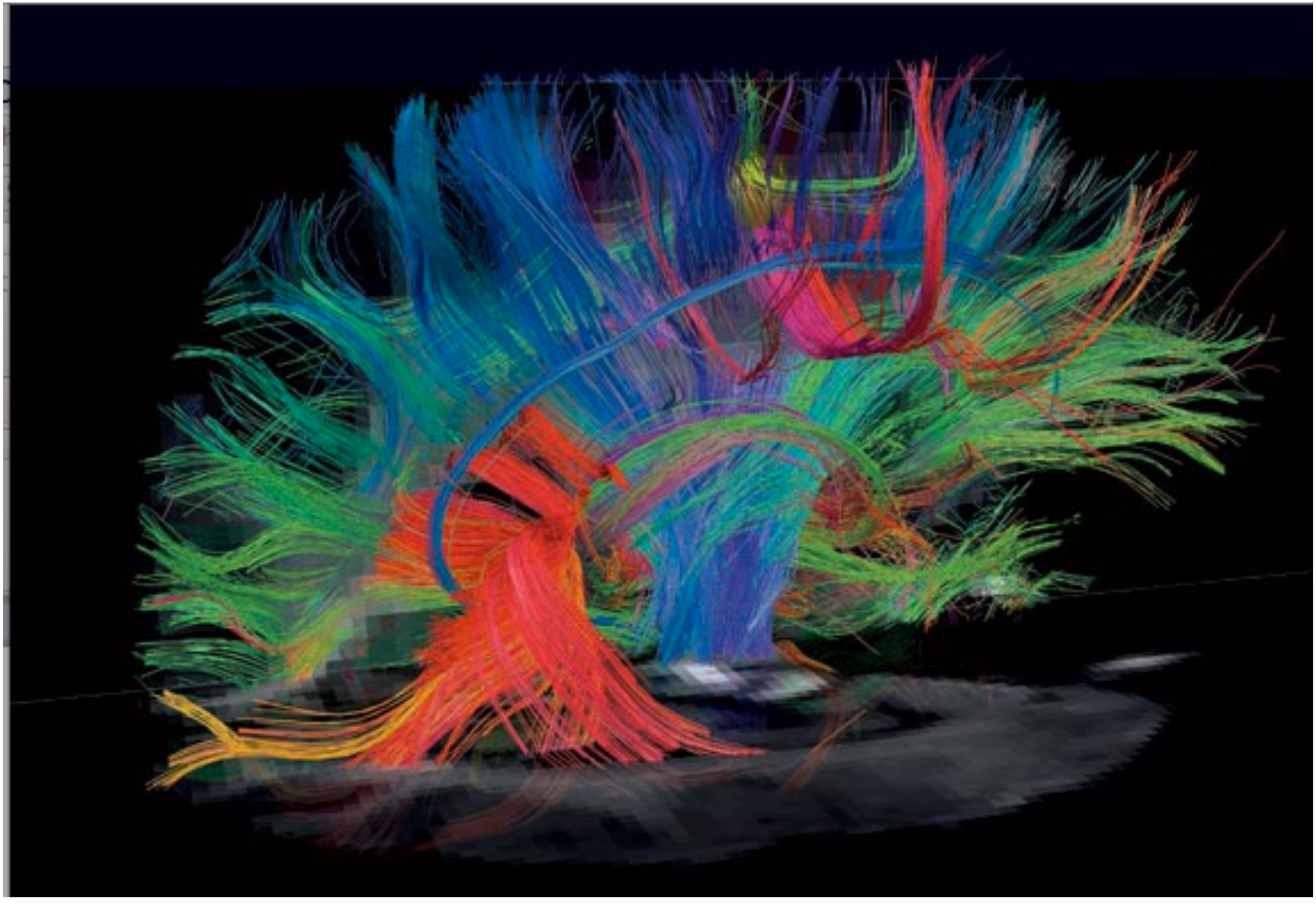


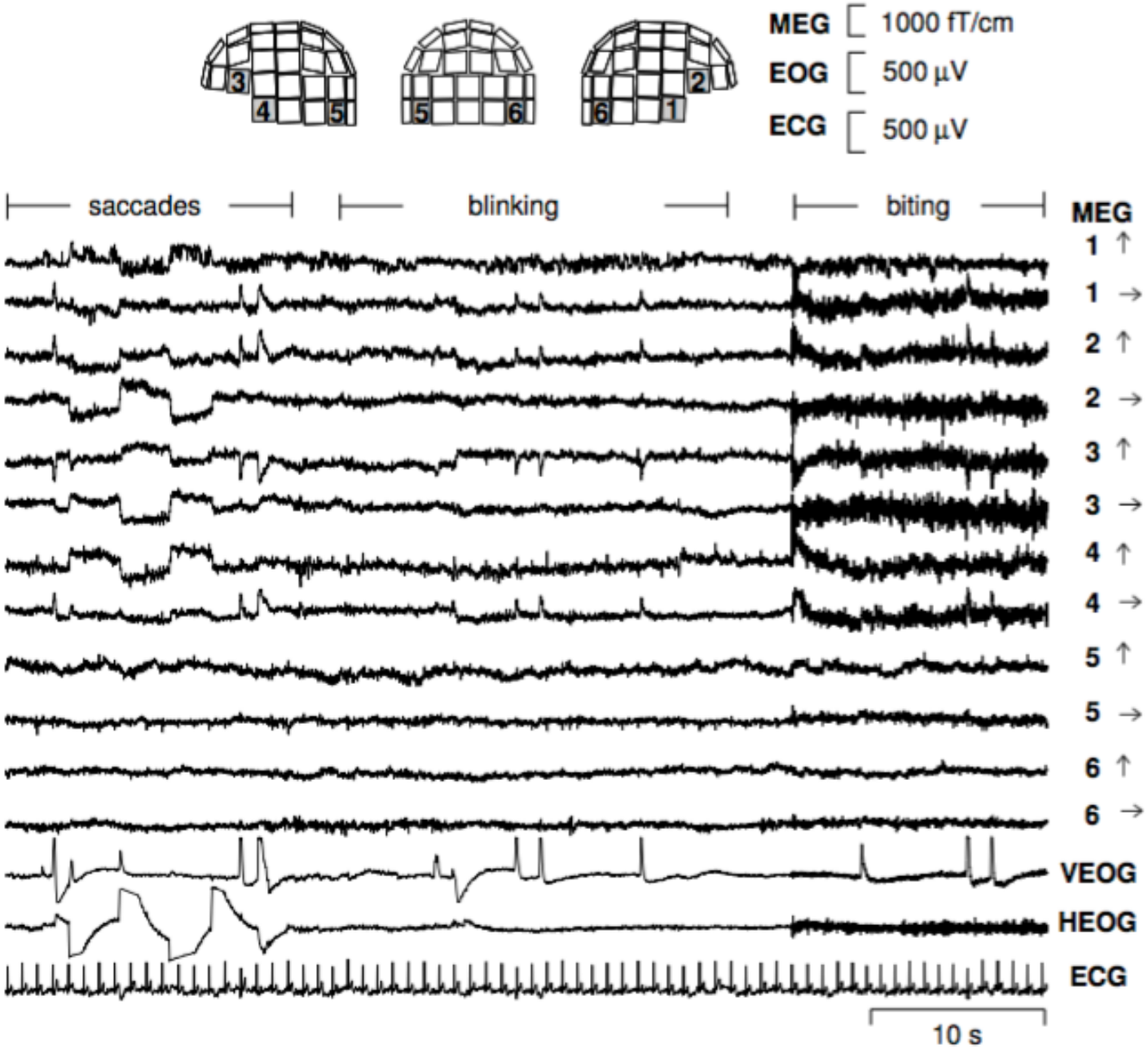
Current along neural circuits produces a magnetic field (MEG)

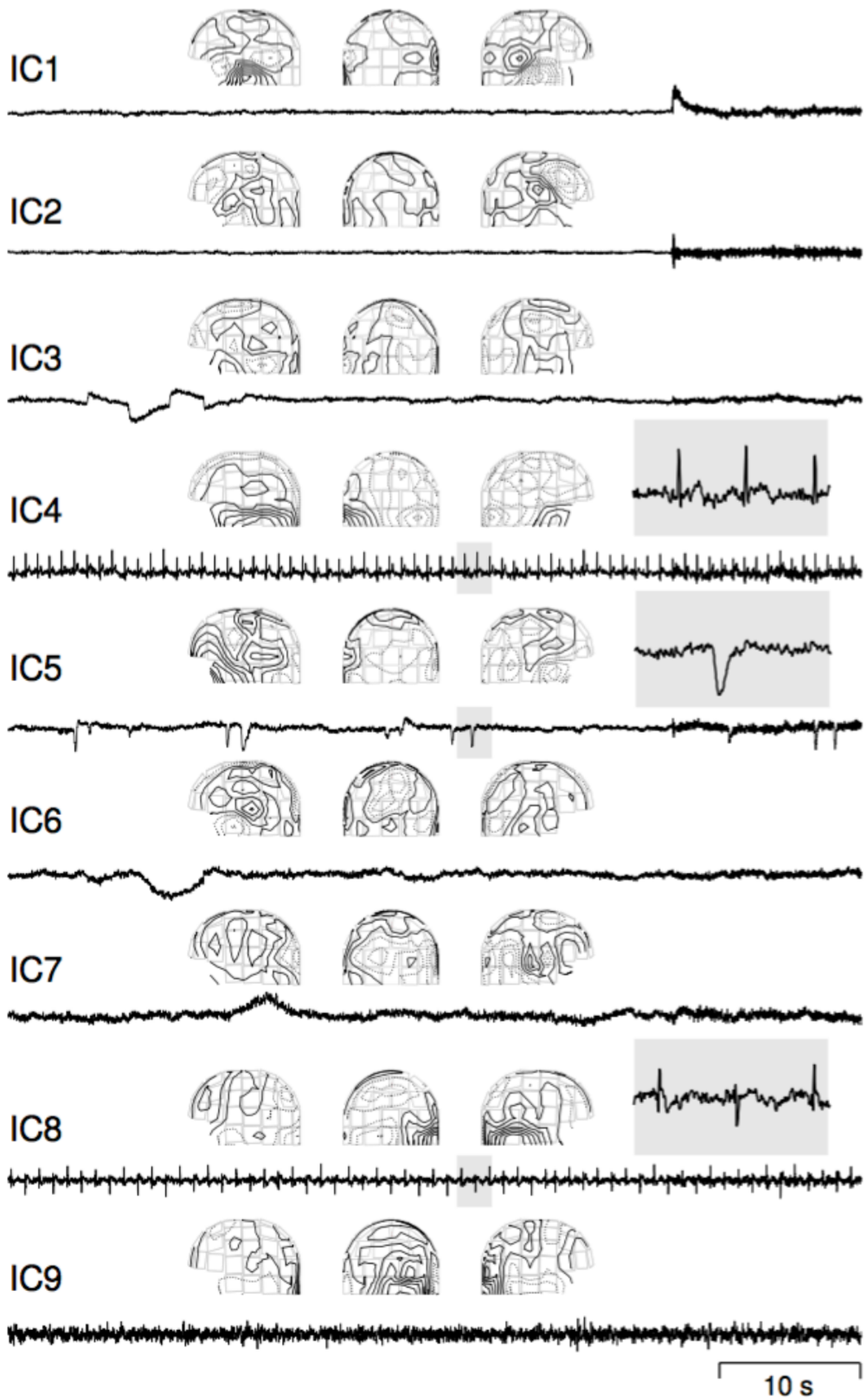




Diffusion tensor imaging of the main brain axonal conduits




MEG
1000 fT/cm
EOG
500 µV
ECG
500 µV
saccades
blinking
biting
MEG
1 ↑
1 →
2 ↑
2 →
3 ↑
3 →
4 ↑
4 →
5 ↑
5 →
6 ↑
6 →
VEOG
HEOG
ECG
10 s


IC1
IC2
IC3
IC4
IC5
IC6
IC7
IC8
IC9
10 s

sources s

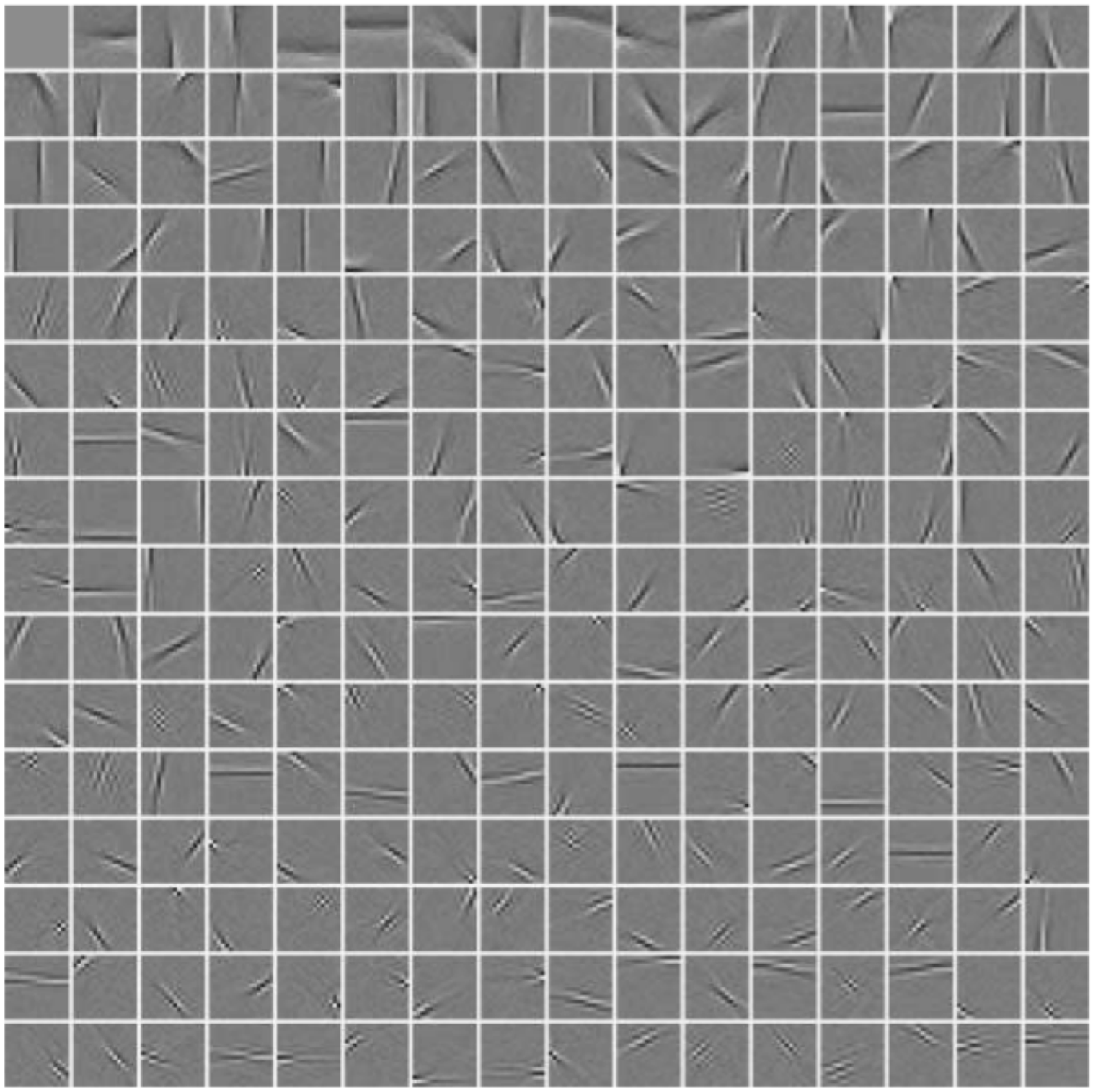
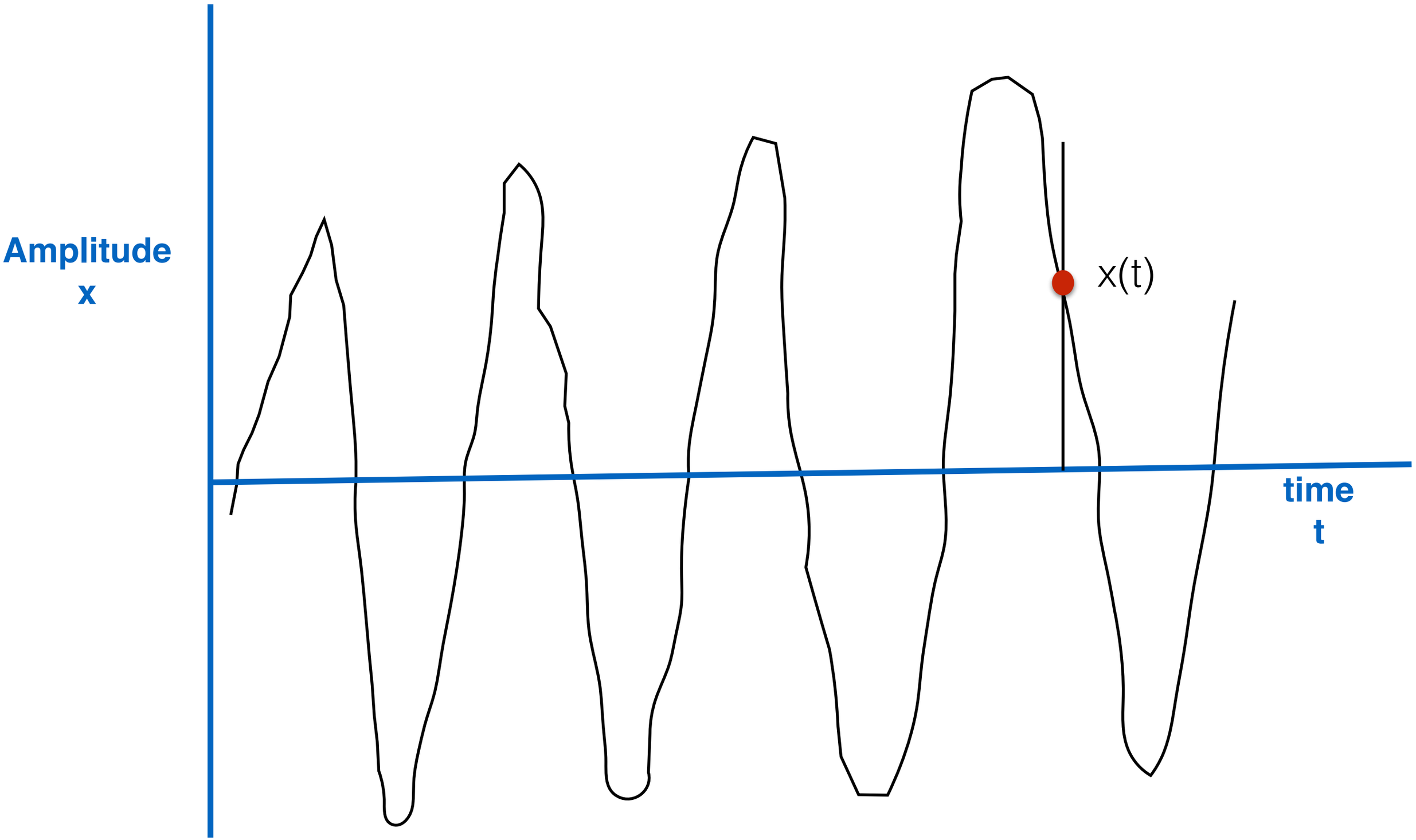
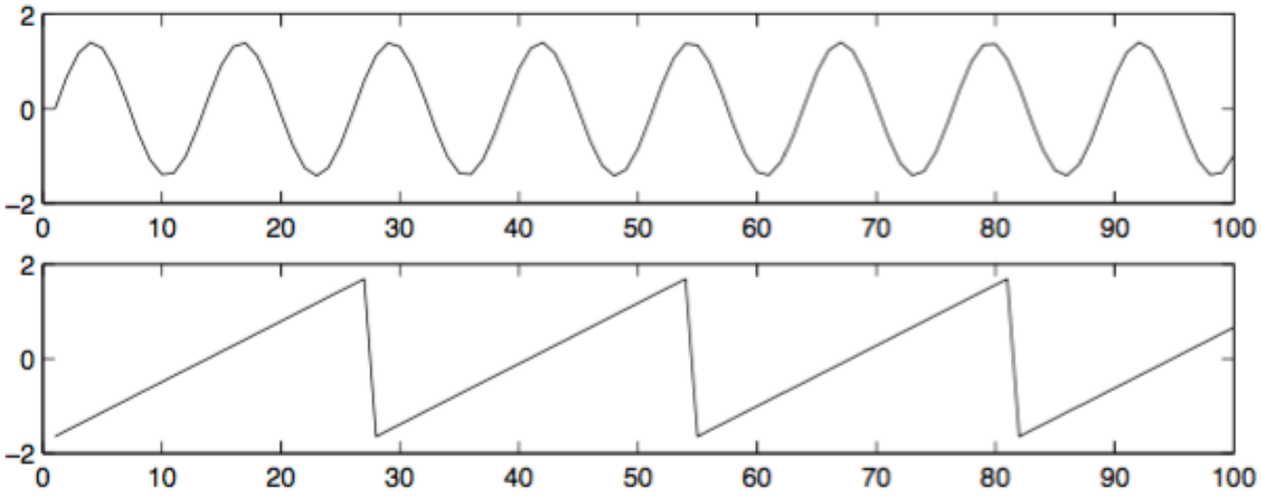


image x

Figure 4: Basis functions in ICA of natural images. The input window size was 16×16 pixels. These basis functions can be considered as the independent features of images.

Audio signals are sampled waveforms



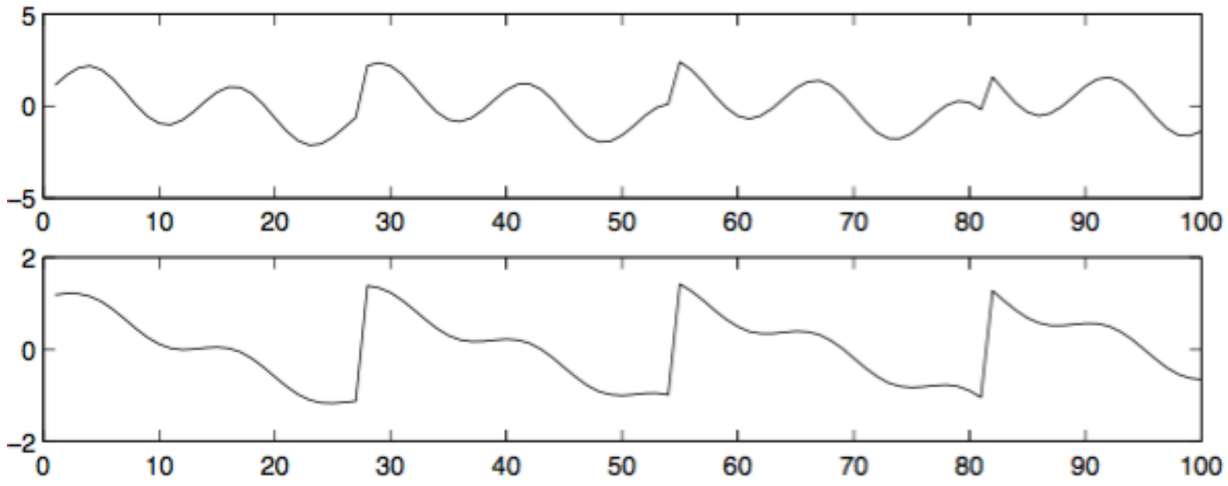


S_1

S_2

An example with two signals

Figure 1: The original signals.



X_1

X_2

Figure 2: The observed mixtures of the source signals in Fig. 1.

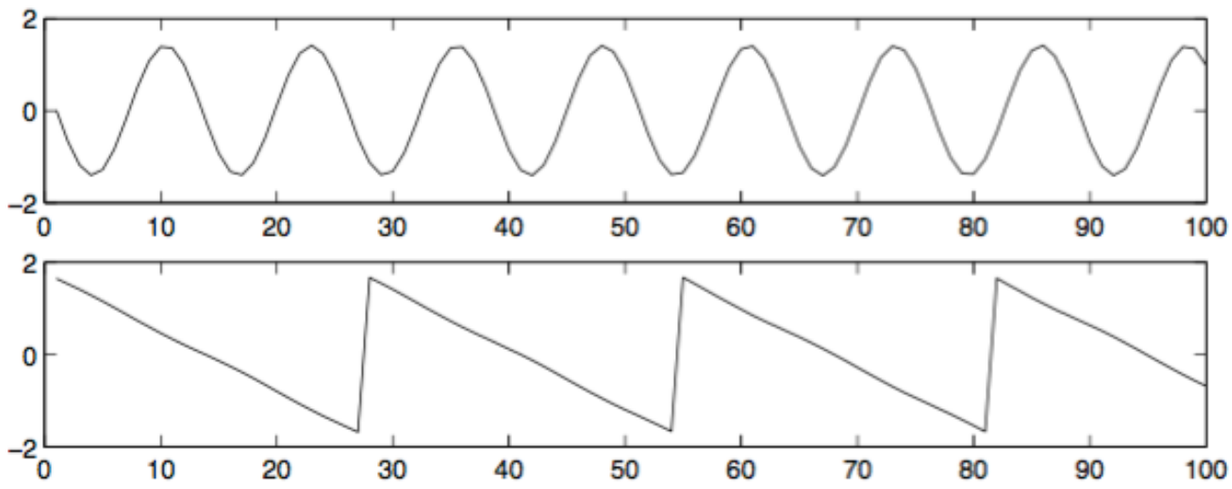


Figure 3: The estimates of the original source signals, estimated using only the observed signals in Fig. 2. The original signals were very accurately estimated, up to multiplicative signs.

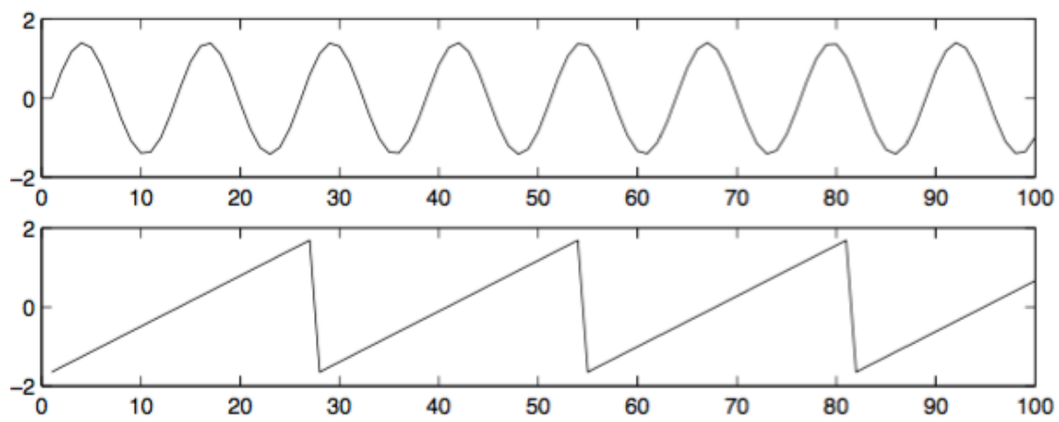
Notational convenience

$$W = \begin{bmatrix} — & w_1^T & — \\ & \vdots & \\ — & w_n^T & — \end{bmatrix}.$$

Thus, $w_i \in \mathbb{R}^n$, and the j-th source can be recovered by computing $s_j^{(i)} = w_j^T x^{(i)}$.

Cannot distinguish between permutations

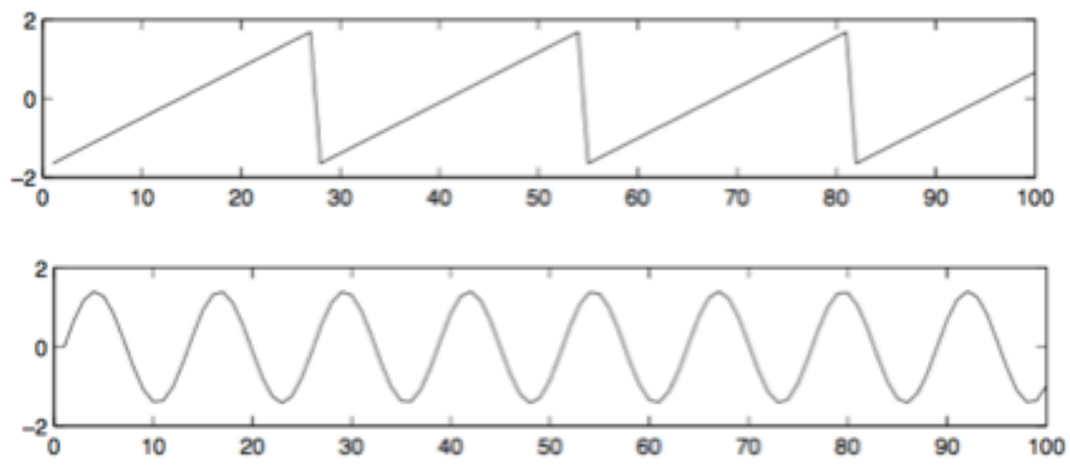
$$P = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}; \quad P = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}; \quad P = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$



s_1

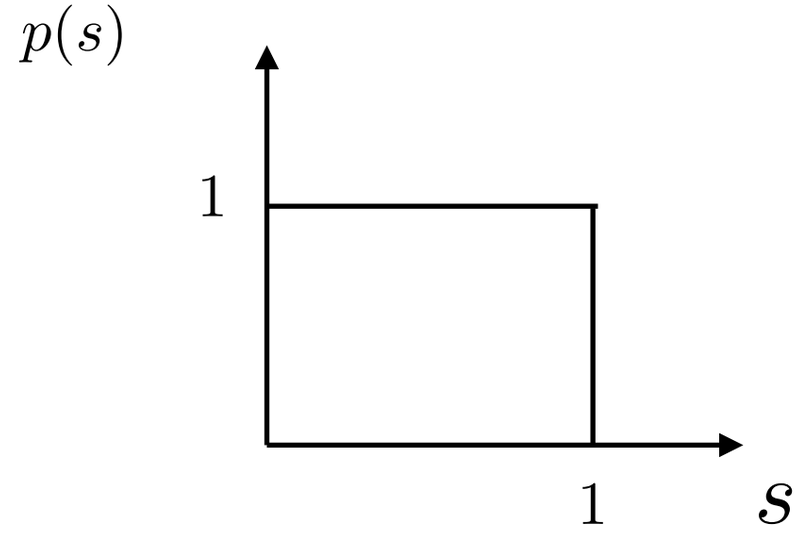
s_2

Figure 1: The original signals.



$\hat{s}_1$

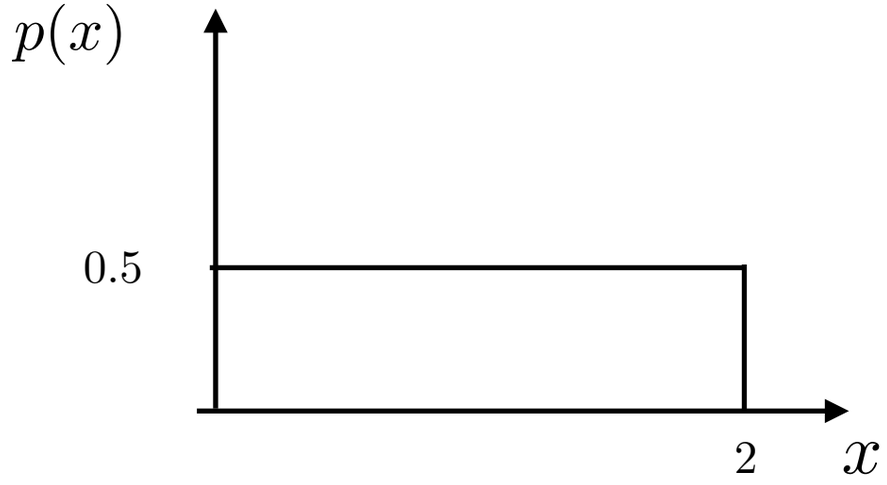
$\hat{s}_2$



~~$p(x) = p_s(Wx)$?~~

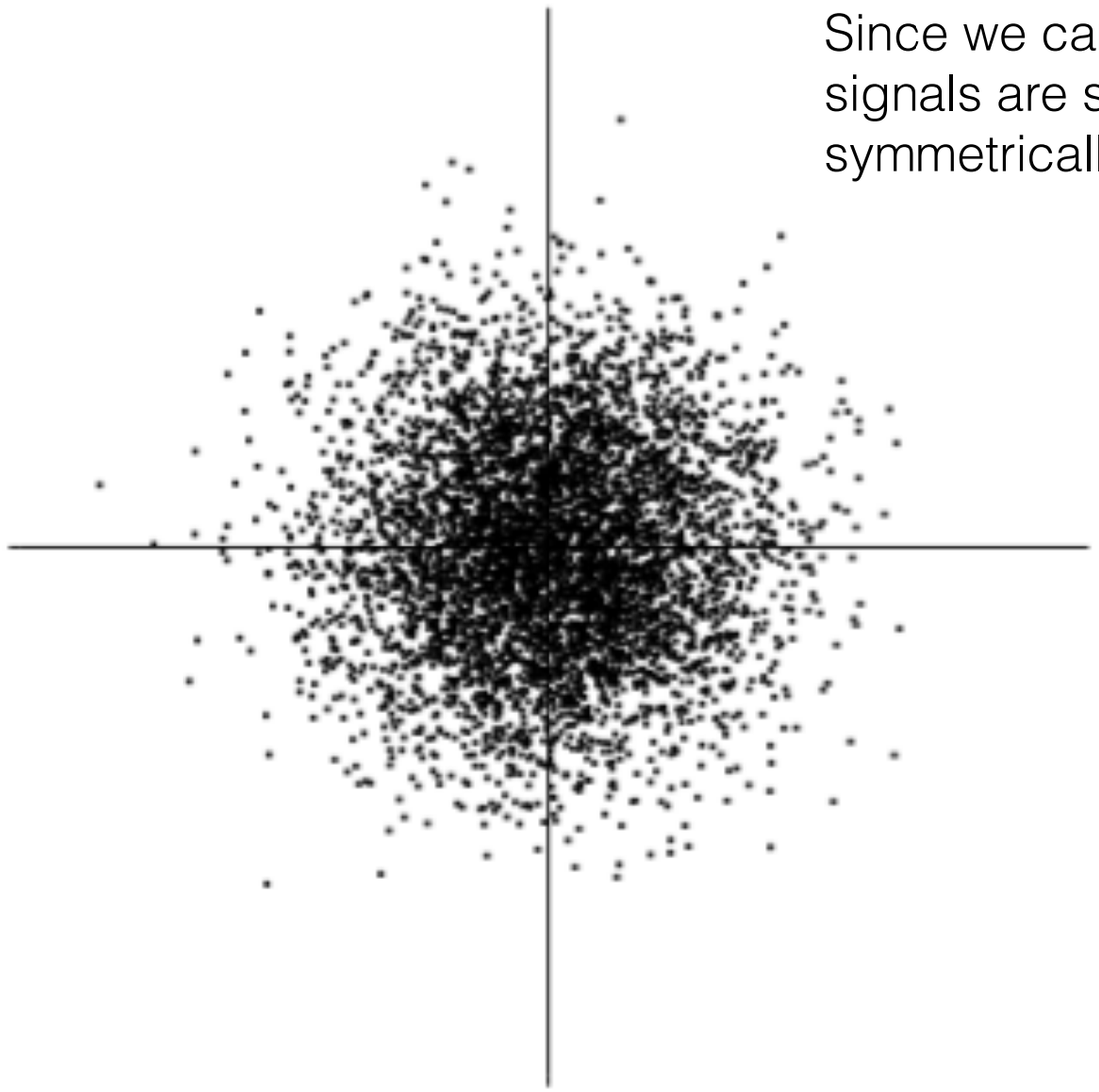
$A = 2$

$x = 2s$



$$p(x) = p_s(Ws)|W|$$

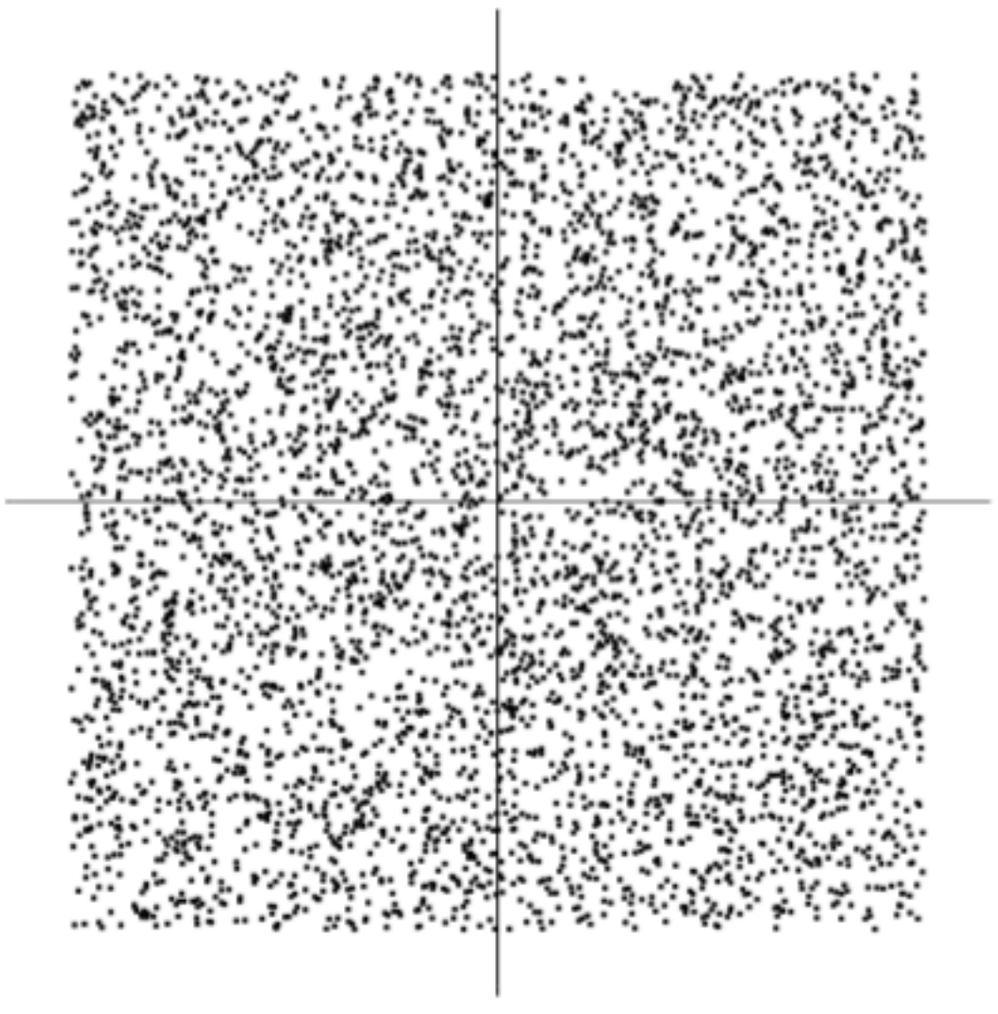
Symmetries in Gaussians preclude recovery of mixed signals



Since we can't tell how the original
signals are stalled, if they are Gaussian, and they are scaled
symmetrically, we can't any original coordinate system

Distributions with higher-order statistics don’t have this problem

Original



Transformed

Source samples are treated as independent

$$p(s) = \prod_{i=1}^{n} p_s(s_i)$$

We can write mixed signal pdf interns of source pdf, but we have to include correction factor |W|

$$p(x) = \prod_{i=1}^{n} p_s(w_i^T x) \cdot |W|$$

Don’t really know pdf,
but we can use the approximate **cdf** g(s):

$$g(s) = 1/(1 + e^{-s})$$

so that the pdf is given by g’(s)

Now write the **log** of the likelihood function for
m samples of n mixed signals

$$\ell(W) = \sum_{i=1}^{m} \left(\sum_{j=1}^{n} \log g'(w_j^T x^{(i)}) + \log |W| \right)$$

To produce the algorithm
differentiate the log likelihood function wrt W

$$W := W + \alpha \left(\begin{bmatrix} 1 - 2g(w_1^T x^{(i)}) \\ 1 - 2g(w_2^T x^{(i)}) \\ \vdots \\ 1 - 2g(w_n^T x^{(i)}) \end{bmatrix} x^{(i)^T} + (W^T)^{-1} \right)$$

Taking the derivative

$$\frac{\partial\, \ell(W)}{\partial W} = \sum_{i=1}^{m} \left(\sum_{j=1}^{n} \log g'(w_j^T x^{(i)}) + \log |W| \right)$$

$$\rightarrow$$

$$W := W + \alpha \left(\begin{bmatrix} 1 - 2g(w_1^T x^{(i)}) \\ 1 - 2g(w_2^T x^{(i)}) \\ \vdots \\ 1 - 2g(w_n^T x^{(i)}) \end{bmatrix} x^{(i)^T} + (W^T)^{-1} \right)$$

Relationship of the algorithm to information theory

An important property of mutual information (Papoulis, 1991; Cover and Thomas, 1991) is that we have for an invertible linear transformation $\mathbf{y} = \mathbf{W}\mathbf{x}$:

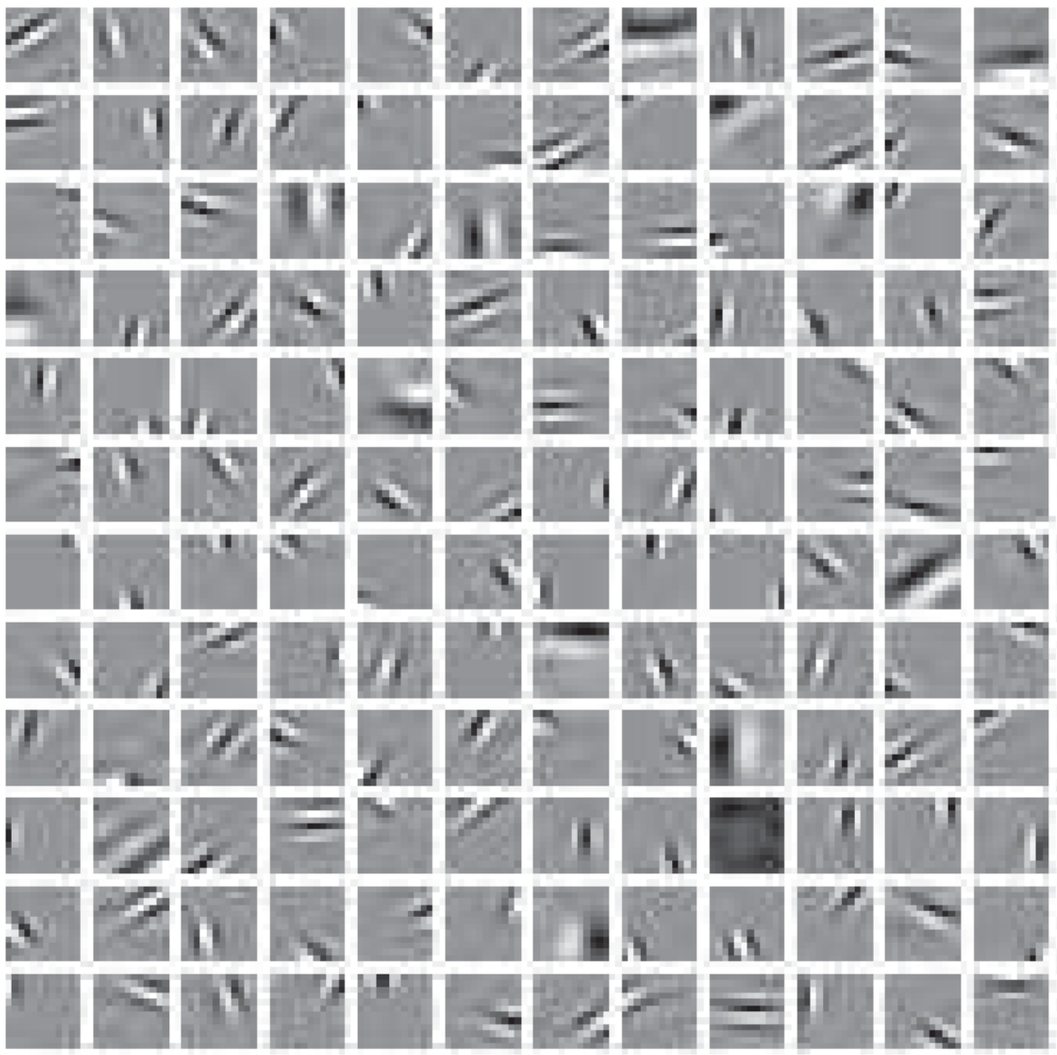
$$I(y_1, y_2, \ldots, y_n) = \sum_i H(y_i) - H(\mathbf{x}) - \log|\det \mathbf{W}|. \tag{28}$$

$$\frac{1}{T} E\{L\} = \sum_{i=1}^{n} E\{\log f_i(\mathbf{w}_i^T \mathbf{x})\} + \log|\det \mathbf{W}|.$$

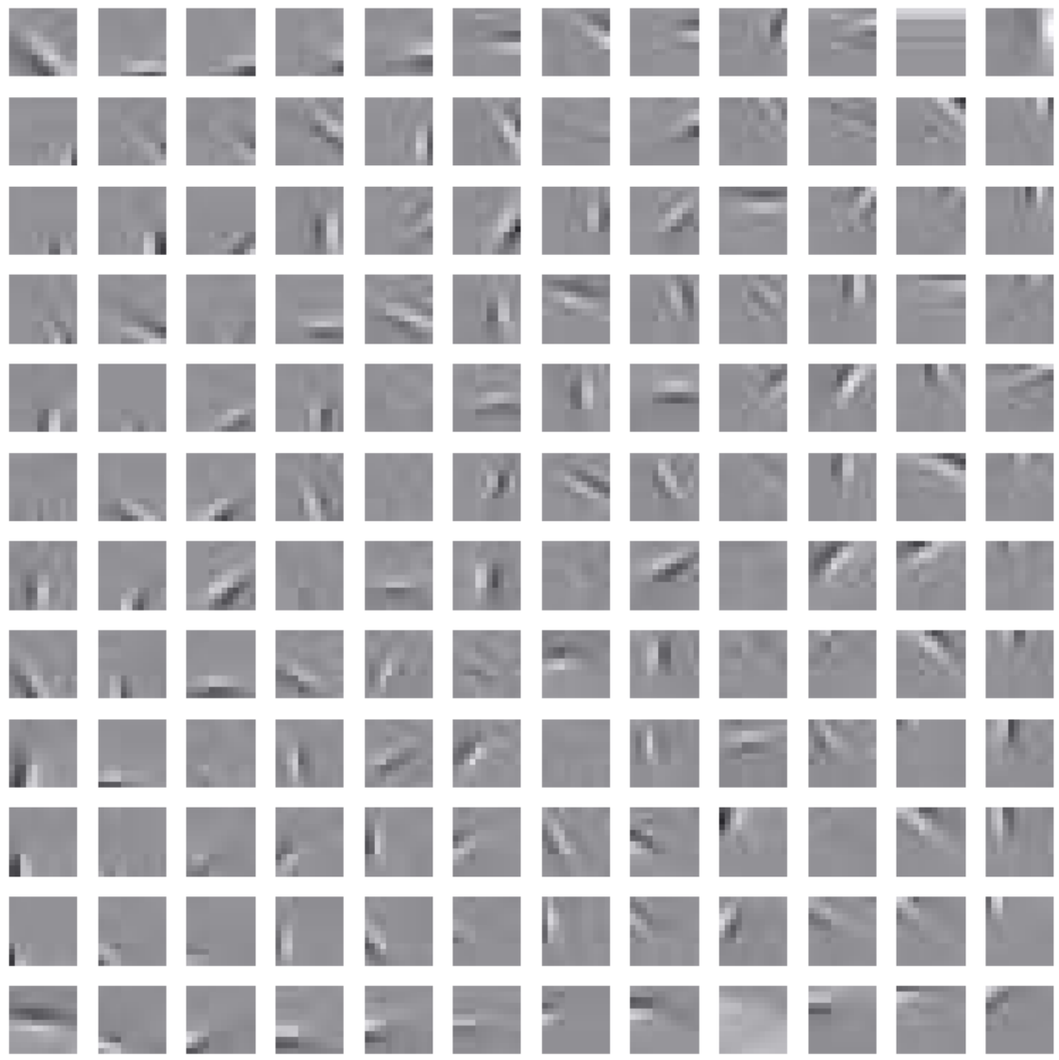
Actually, if the f_i were equal to the actual distributions of $\mathbf{w}_i^T \mathbf{x}$, the first term would be equal to $-\sum_i H(\mathbf{w}_i^T \mathbf{x})$. Thus the likelihood would be equal, up to an additive constant, to the negative of mutual information as given in Eq. (28).

Actually, in practice the connection is even stronger. This is because in practice we don't know the distributions of the independent components. A reasonable approach would be to estimate the density of $\mathbf{w}_i^T \mathbf{x}$ as part of the ML estimation method, and use this as an approximation of the density of s_i. In this case, likelihood and mutual information are, for all practical purposes, equivalent.

Sparse Coding



ICA



Information theory and ICA

A fundamental result of information theory is that *a gaussian variable has the largest entropy among all random variables of equal variance*. For a proof, see e.g. (Cover and Thomas, 1991; Papoulis, 1991). This means that entropy could be used as a measure of nongaussianity. In fact, this shows that the gaussian distribution is the "most random" or the least structured of all distributions. Entropy is small for distributions that are clearly concentrated on certain values, i.e., when the variable is clearly clustered, or has a pdf that is very "spiky".

To obtain a measure of nongaussianity that is zero for a gaussian variable and always nonnegative, one often uses a slightly modified version of the definition of differential entropy, called negentropy. Negentropy J is defined as follows

$$J(\mathbf{y}) = H(\mathbf{y}_{gauss}) - H(\mathbf{y}) \tag{22}$$

where $\mathbf{y}_{gauss}$ is a Gaussian random variable of the same covariance matrix as $\mathbf{y}$. Due to the above-mentioned properties, negentropy is always non-negative, and it is zero if and only if $\mathbf{y}$ has a Gaussian distribution. Negentropy has the additional interesting property that it is invariant for invertible linear transformations (Comon, 1994; Hyvärinen, 1999e).

$$\ell(W) = \sum_{i=1}^{m} \left(\sum_{j=1}^{n} \log g'(w_j^T x^{(i)}) + \log |W| \right)$$