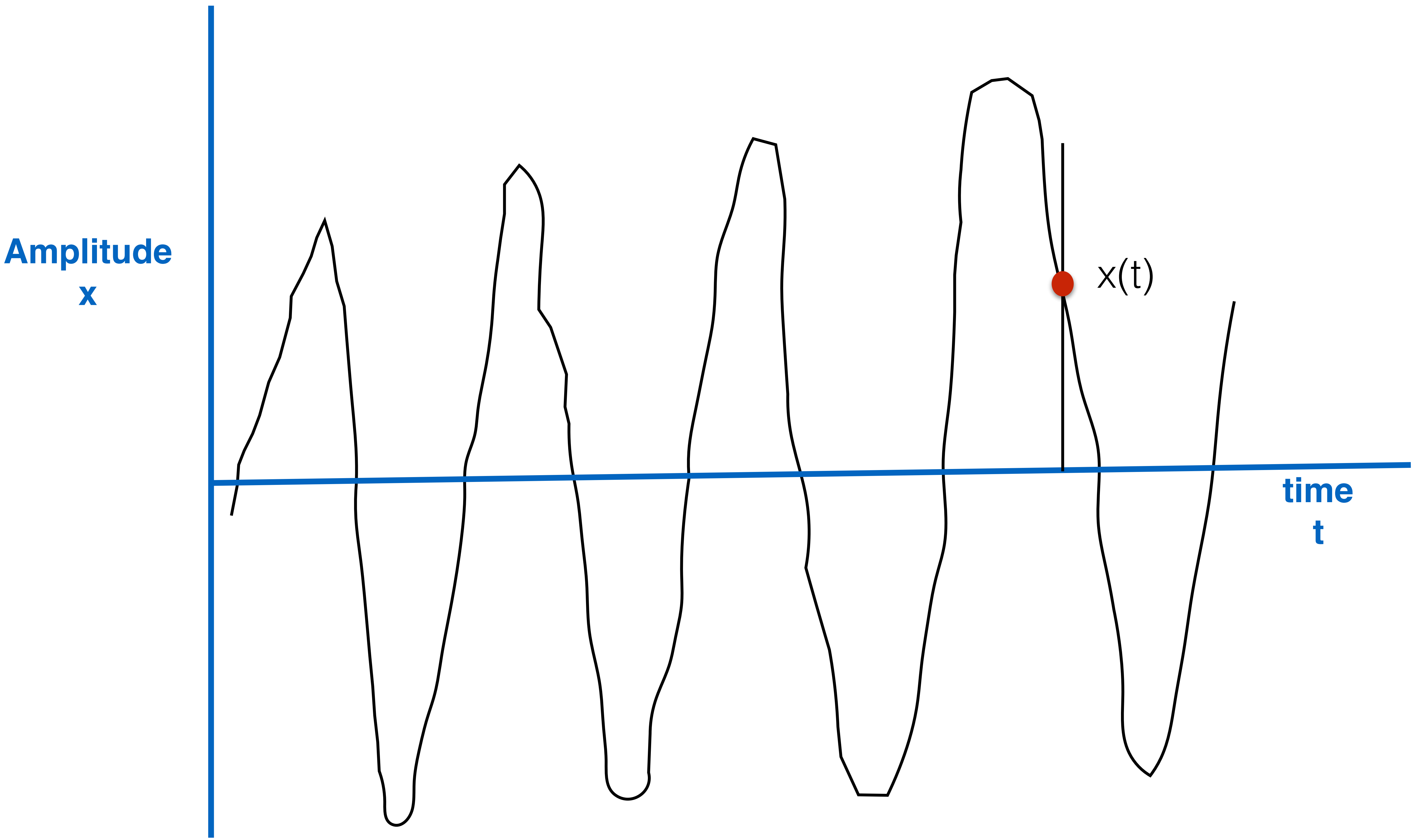
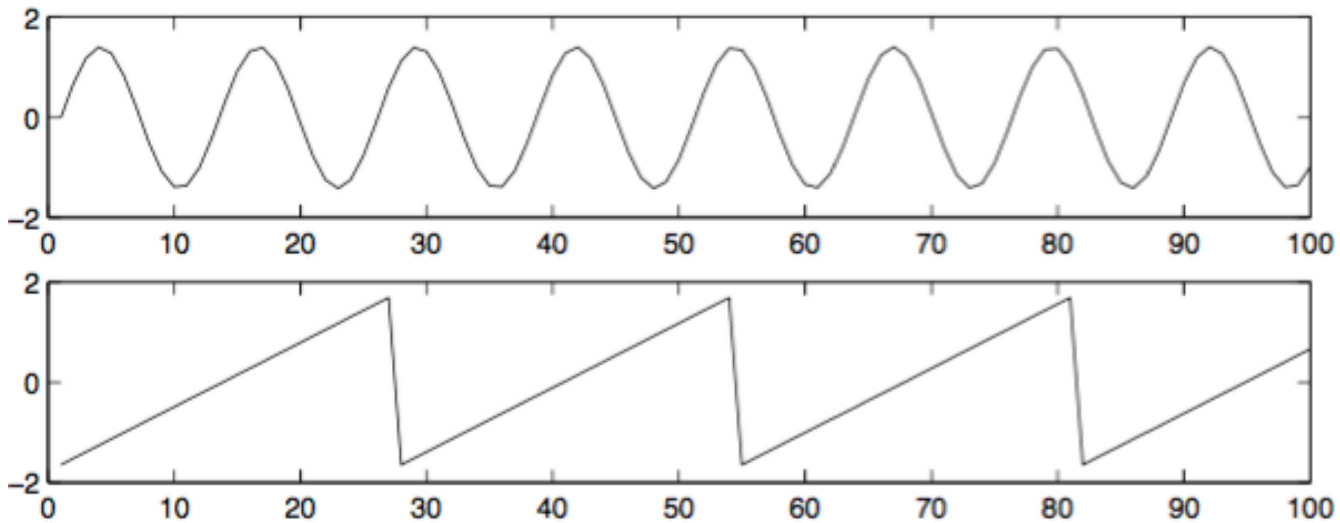
Lecture 6: Independent Components Analysis

Problem Setting: "Cocktail Party Problem"

Sound sources	Sound sources are mixed in microphones	Want to un-mix them	Desired answer
s	$x = As$	$\hat{s} = Wx$	$W = A^{-1}$


Amplitude
x
x(t)
time
t

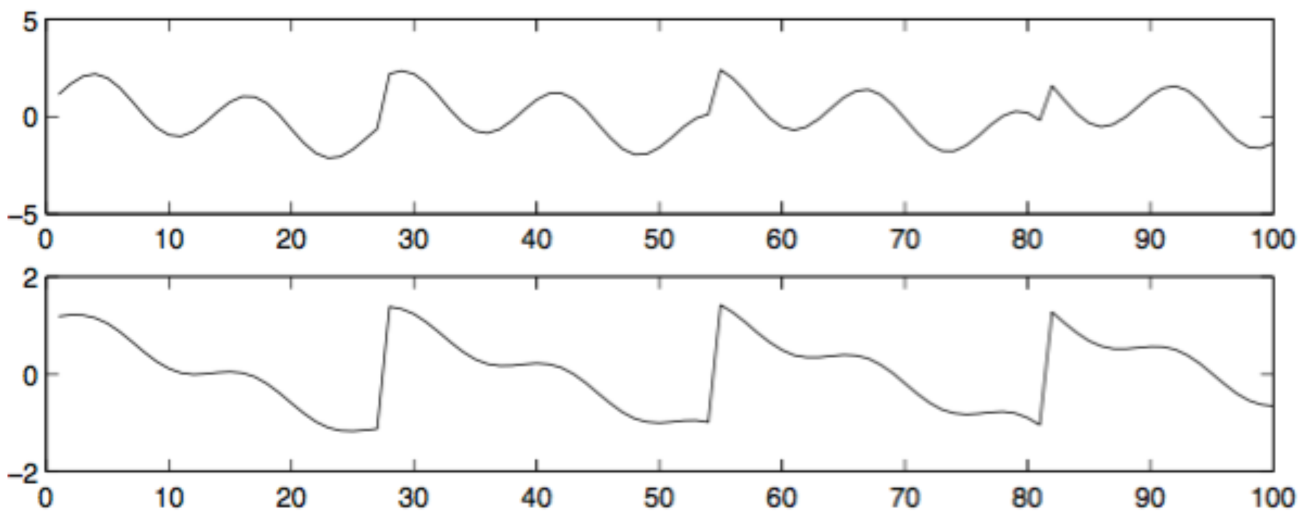
An example with two signals



S1

S2

Figure 1: The original signals.



X1

X2

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

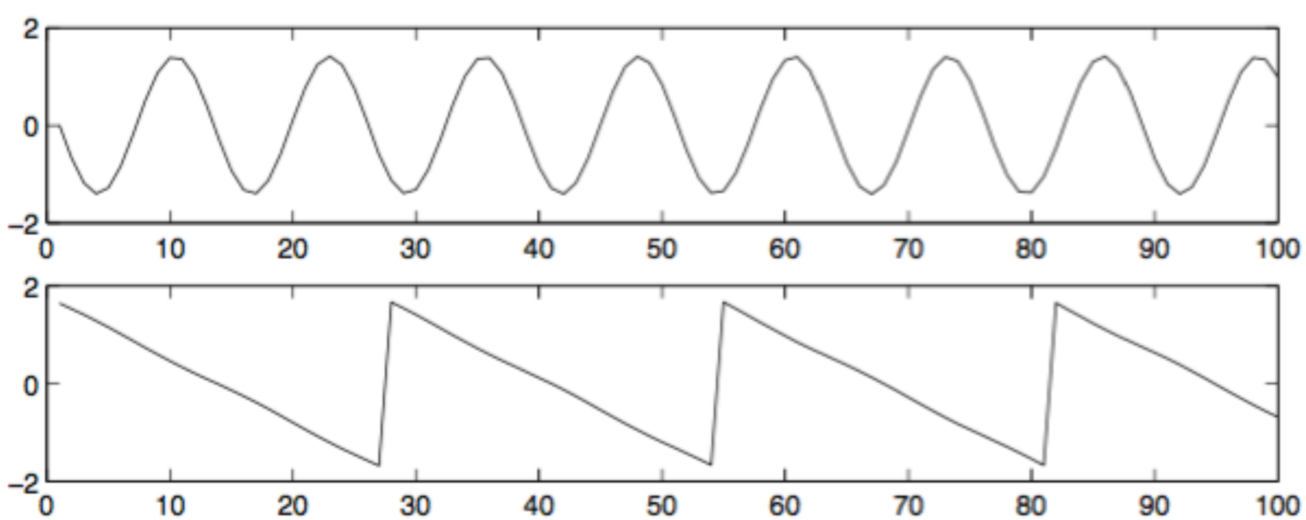


Figure 3: The estimates of the original source signals, estimated using only the observed signals in Fig. 2. The original signals were very accurately estimated, up to multiplicative signs.

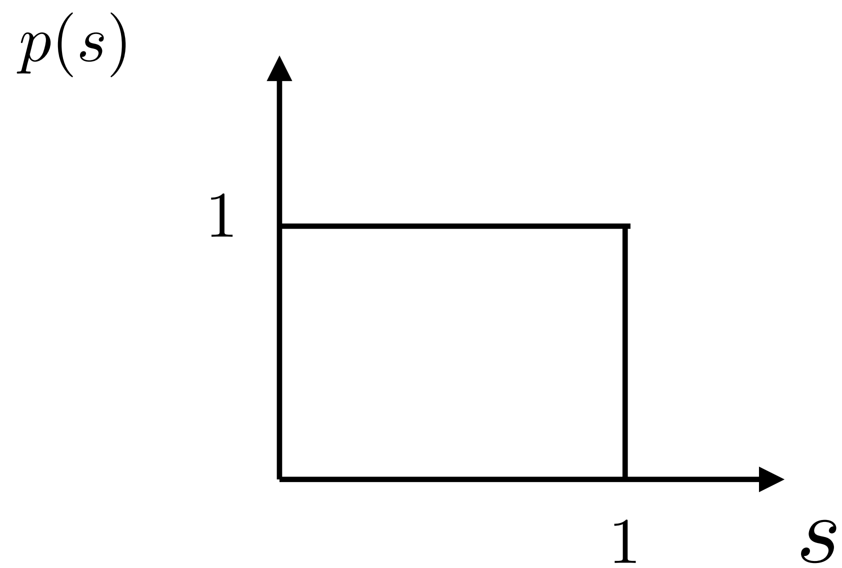
Notational convenience

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

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

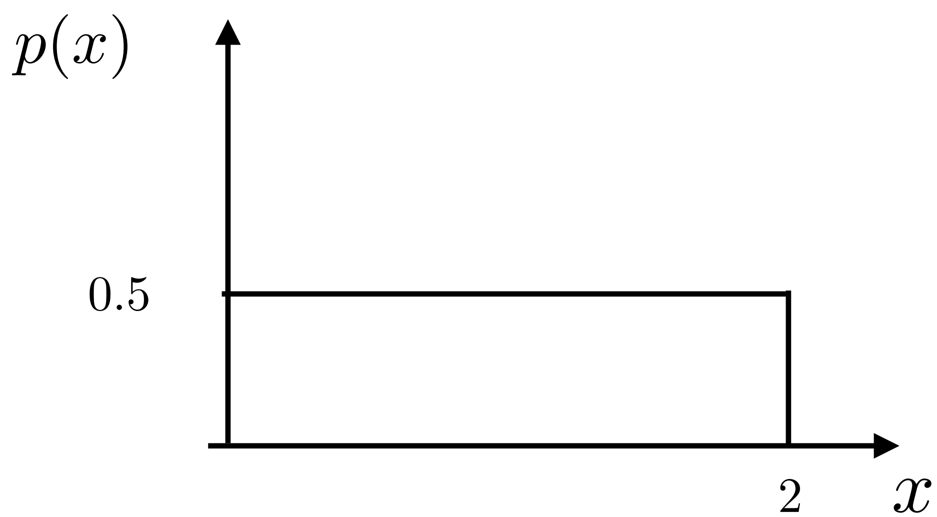
$$P = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}; \quad P = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}; \quad P = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$



$$\cancel{p(x) = p_s(Wx) \quad ?}$$

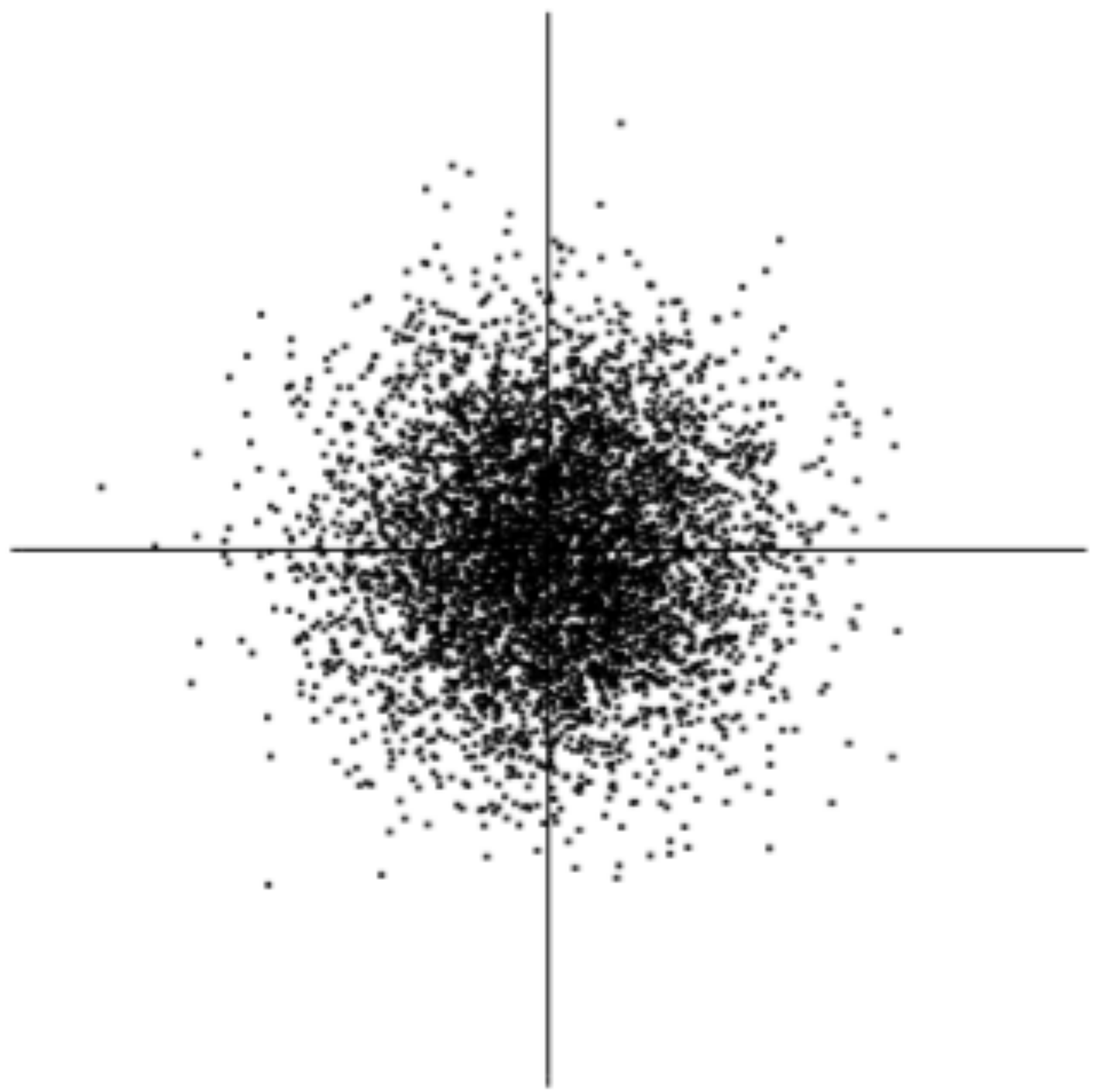
$$A = 2$$

$$x = 2s$$

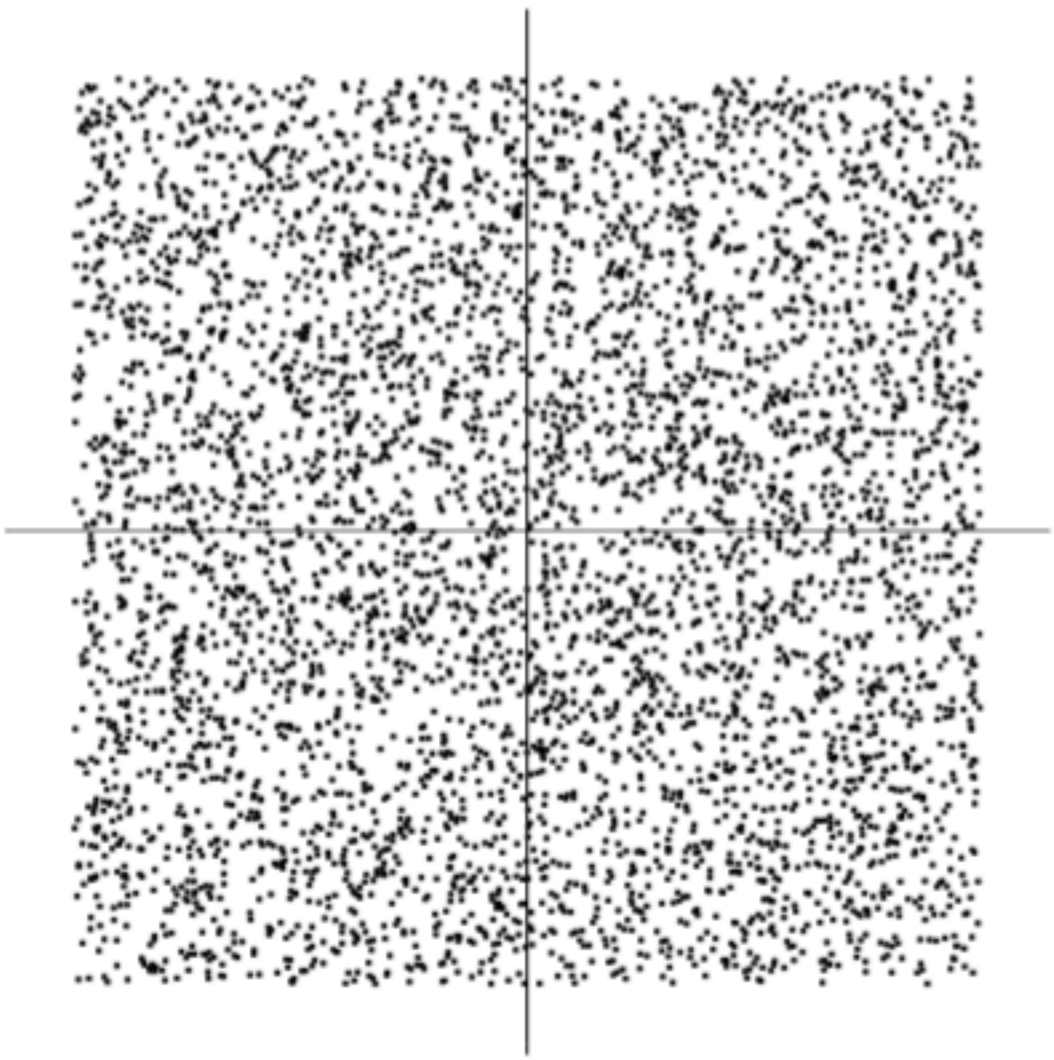


$$p(x) = p_s(Wx)|W|$$

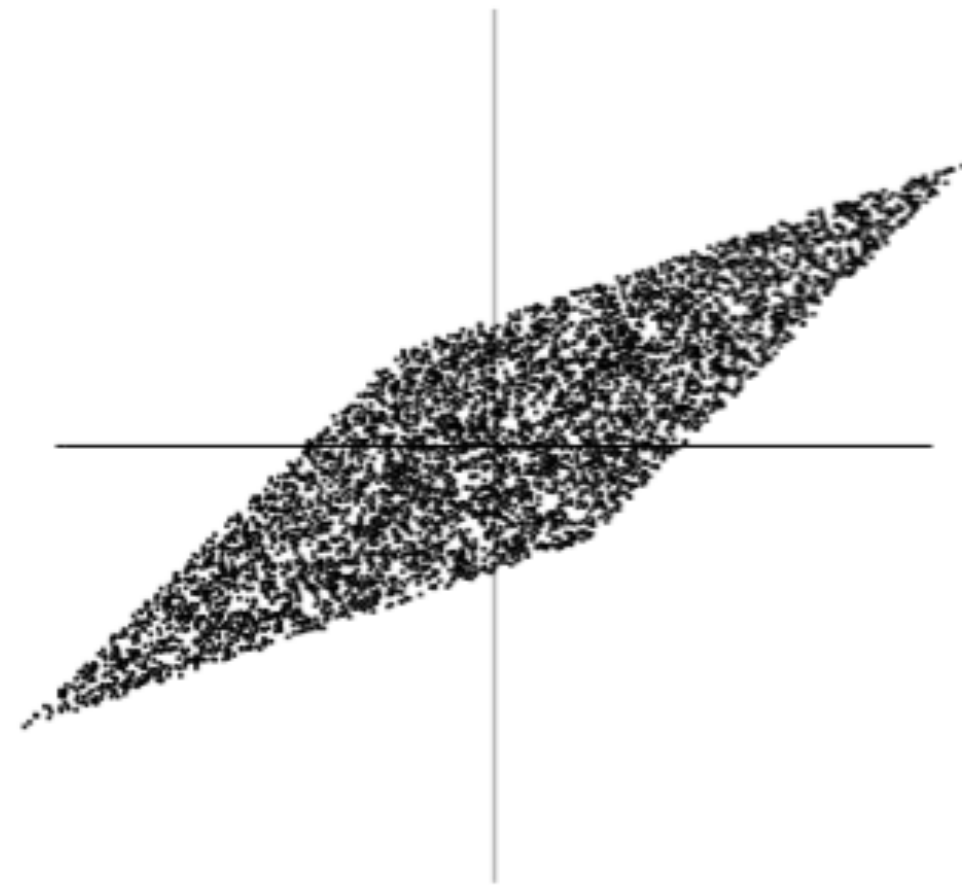
Symmetries in Gaussians preclude recovery of mixed signals



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

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

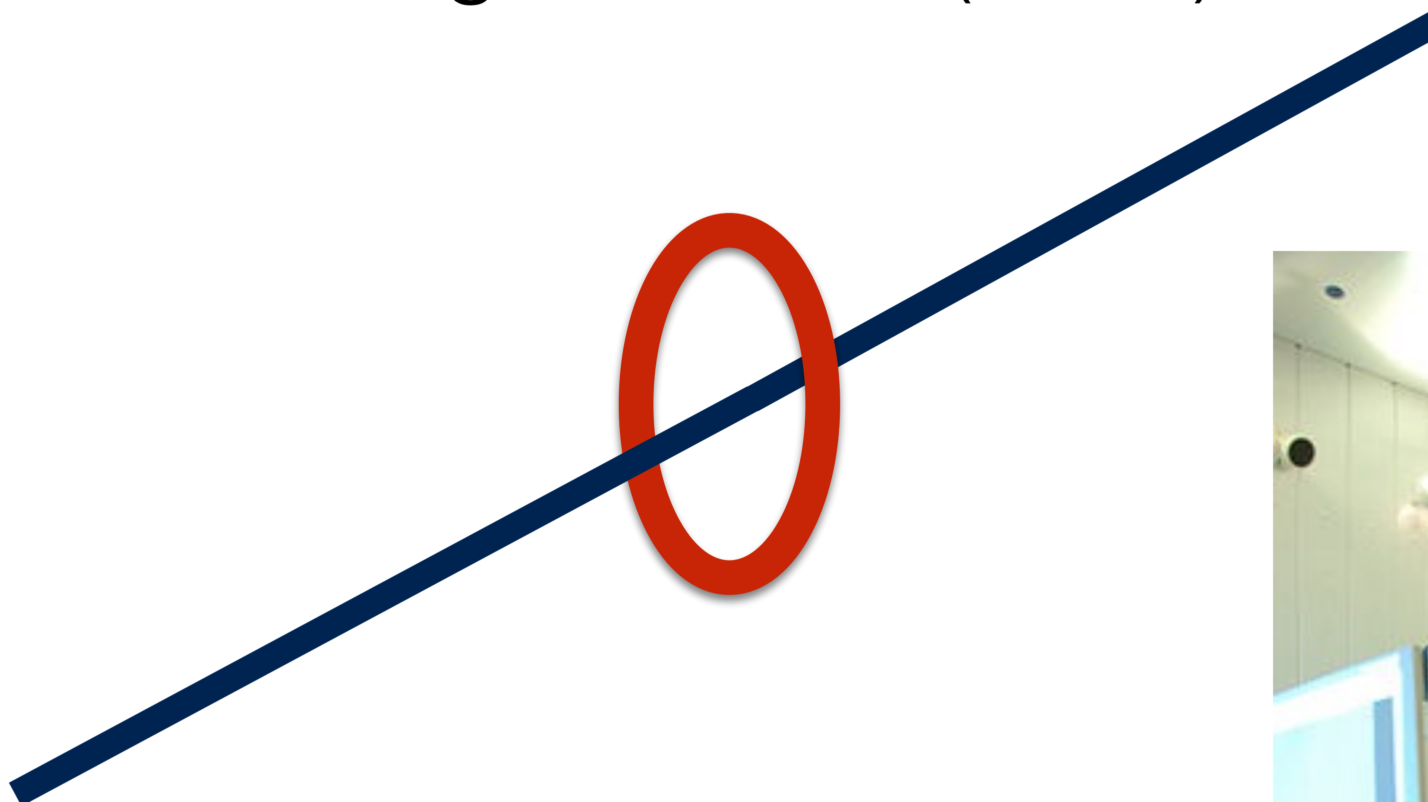
To produce the algorithm
differentiate the log likelihood function wrt W

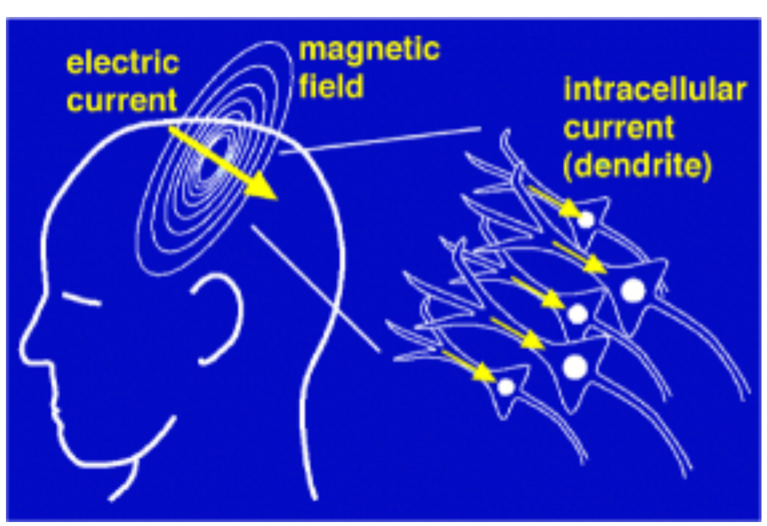
$$W := W + \alpha \left(\begin{bmatrix} 1 - 2g(w_1^T x^{(i)}) \\ 1 - 2g(w_2^T x^{(i)}) \\ \vdots \\ 1 - 2g(w_n^T x^{(i)}) \end{bmatrix} {x^{(i)}}^T + (W^T)^{-1} \right)$$

Taking the derivative

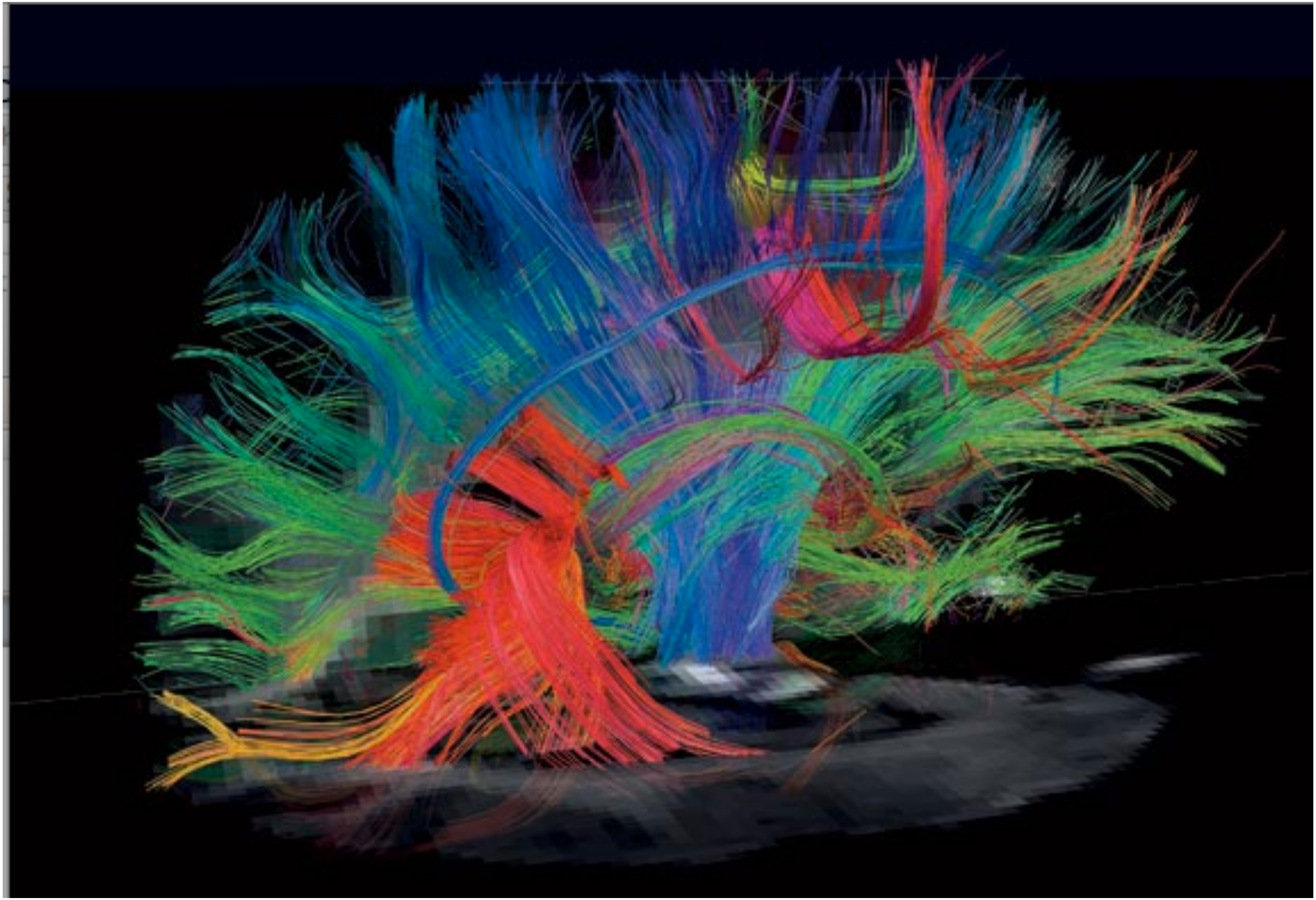
$$\frac{\partial}{\partial W}\ell(W) = \sum_{i=1}^{m}\left(\sum_{j=1}^{n}\log g'(w_j^T x^{(i)}) + \log|W|\right) \longrightarrow W := W + \alpha\left(\begin{bmatrix} 1-2g(w_1^T x^{(i)}) \\ 1-2g(w_2^T x^{(i)}) \\ \vdots \\ 1-2g(w_n^T x^{(i)}) \end{bmatrix} x^{(i)^T} + (W^T)^{-1}\right)$$

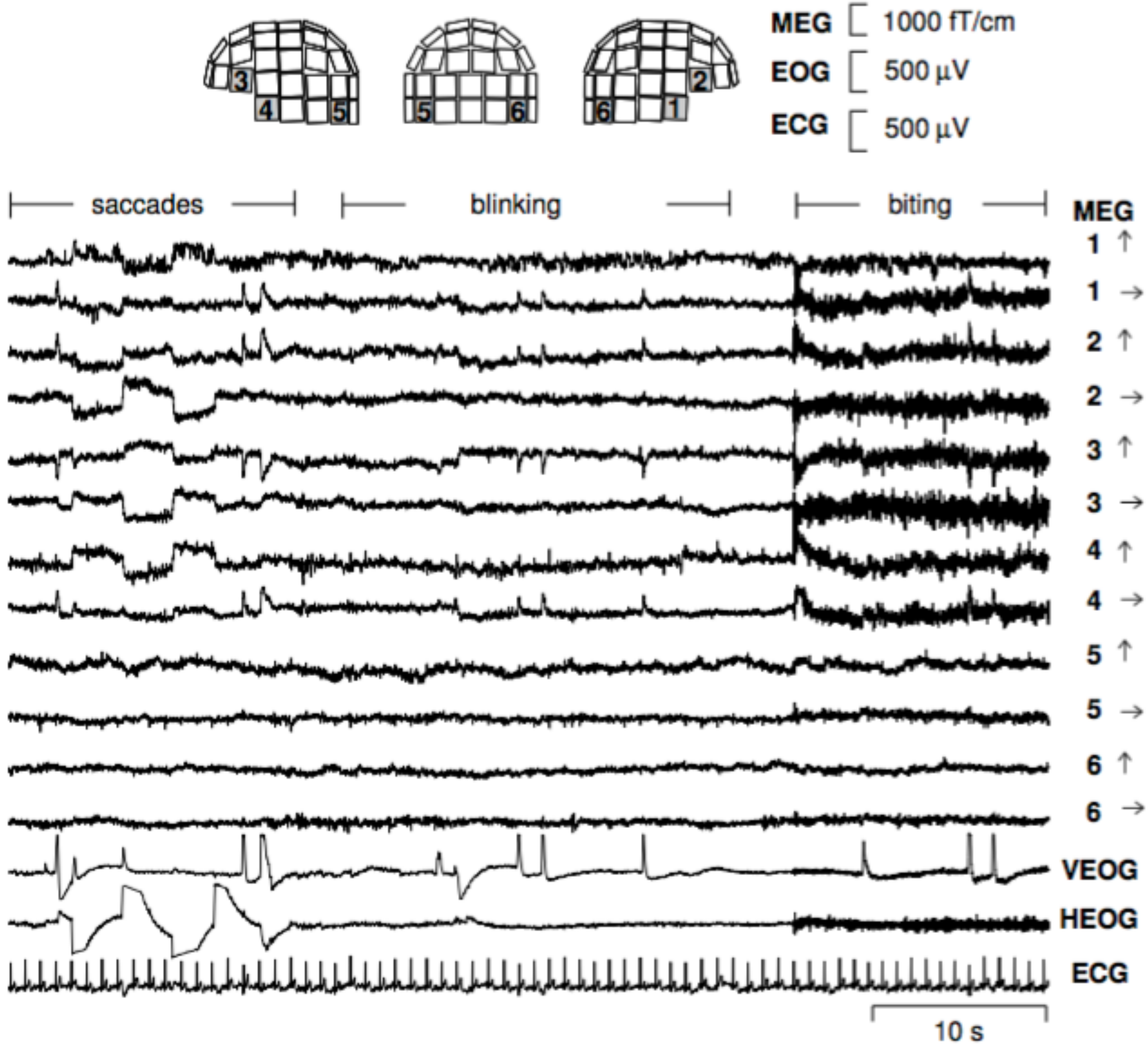
Current along neural circuits produces a magnetic field (MEG)





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

sources s

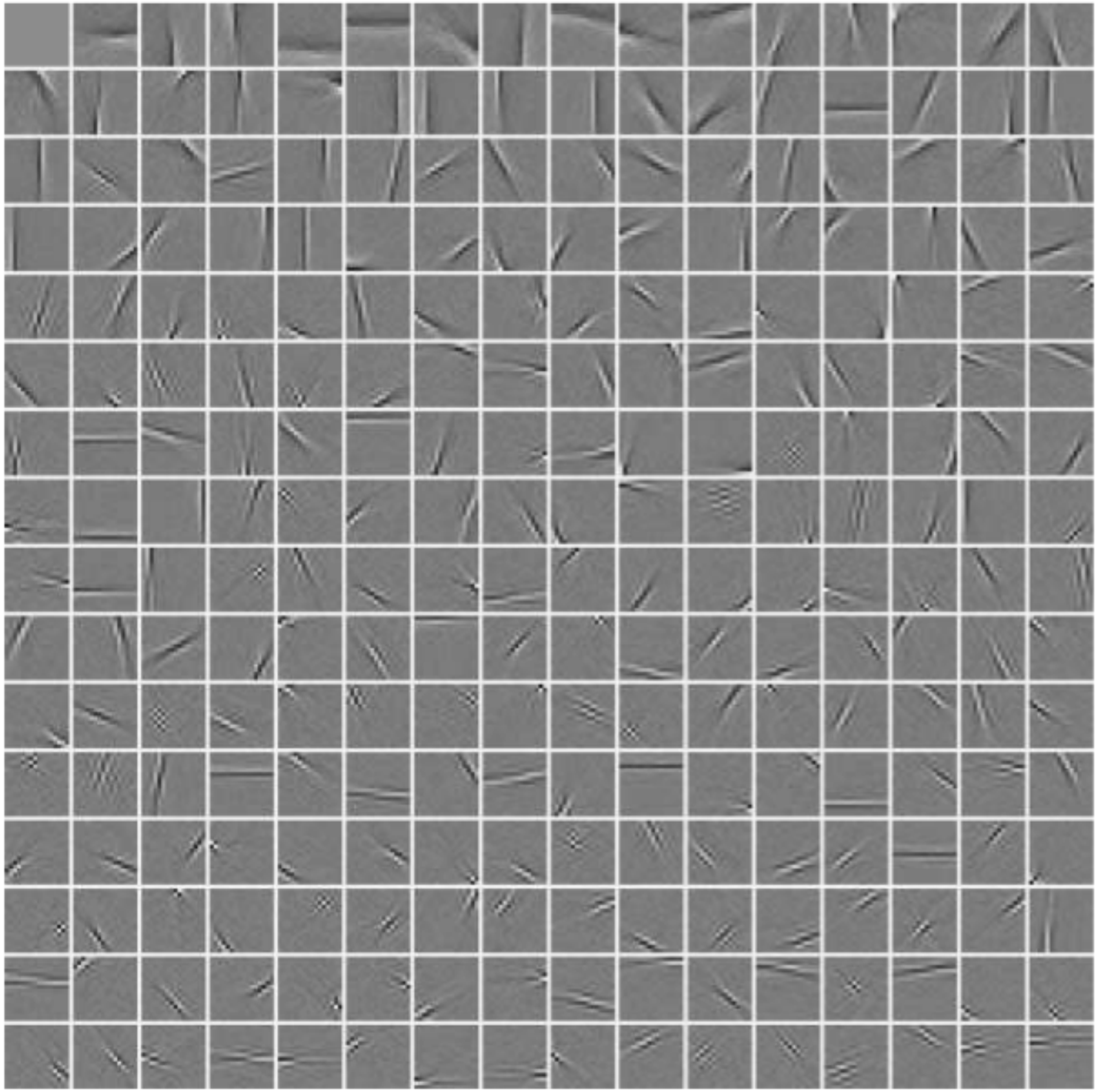
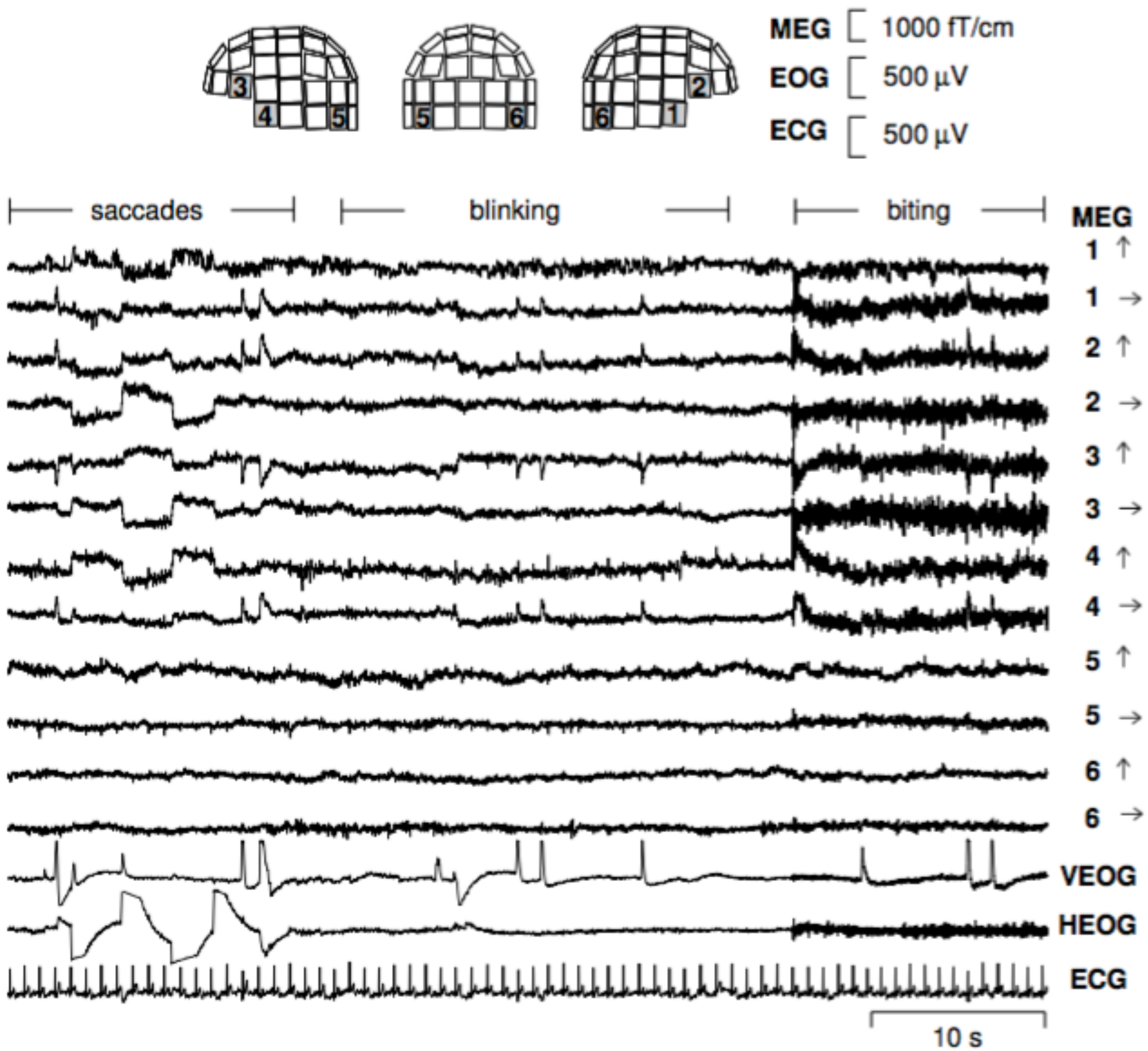
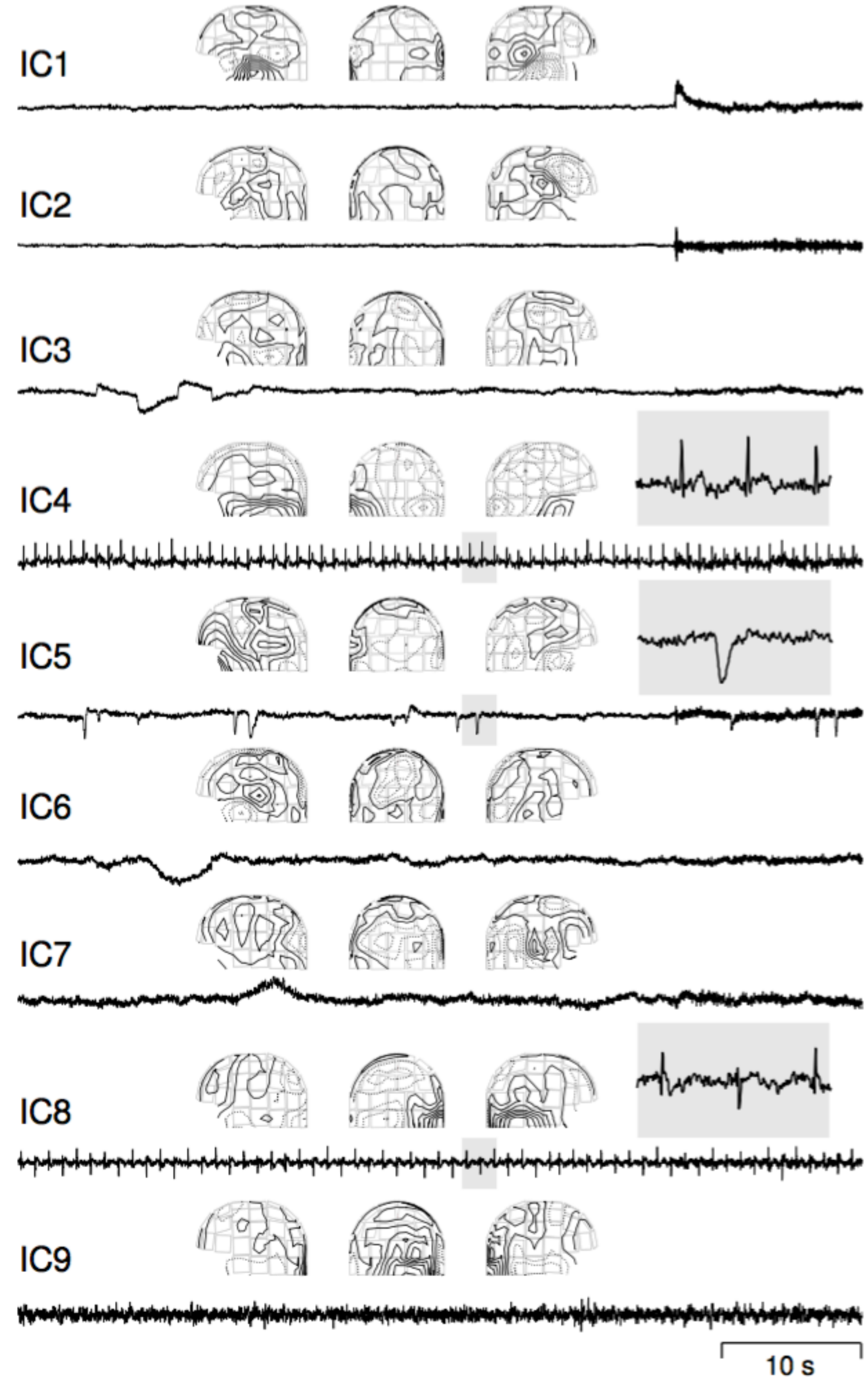


image x

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.



3
4
5
5
6
6
1
2
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

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 $\mathbf{y} = \mathbf{W}\mathbf{x}$:

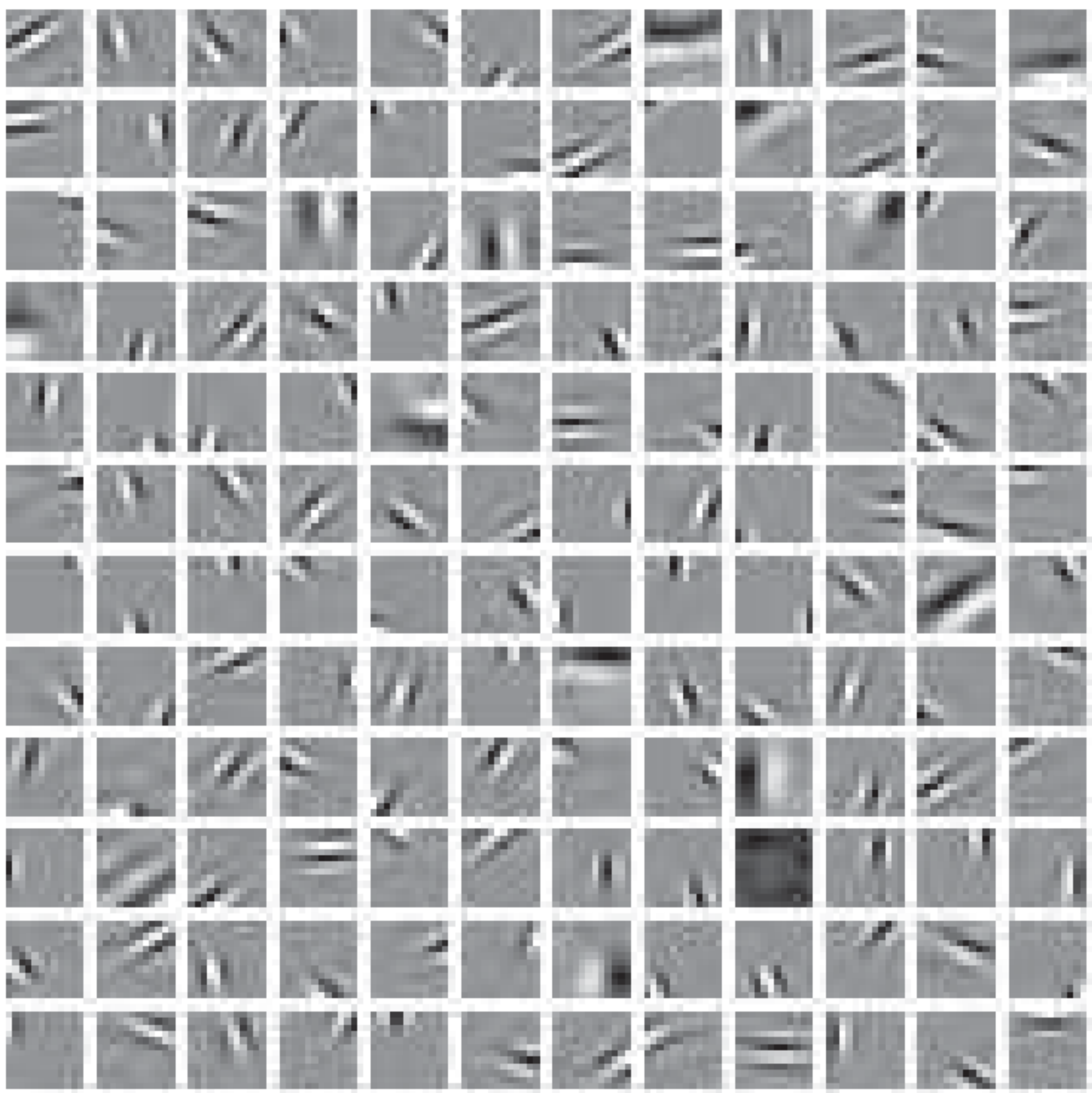
$$I(y_1, y_2, \ldots, y_n) = \sum_i H(y_i) - H(\mathbf{x}) - \log|\det \mathbf{W}|. \tag{28}$$

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

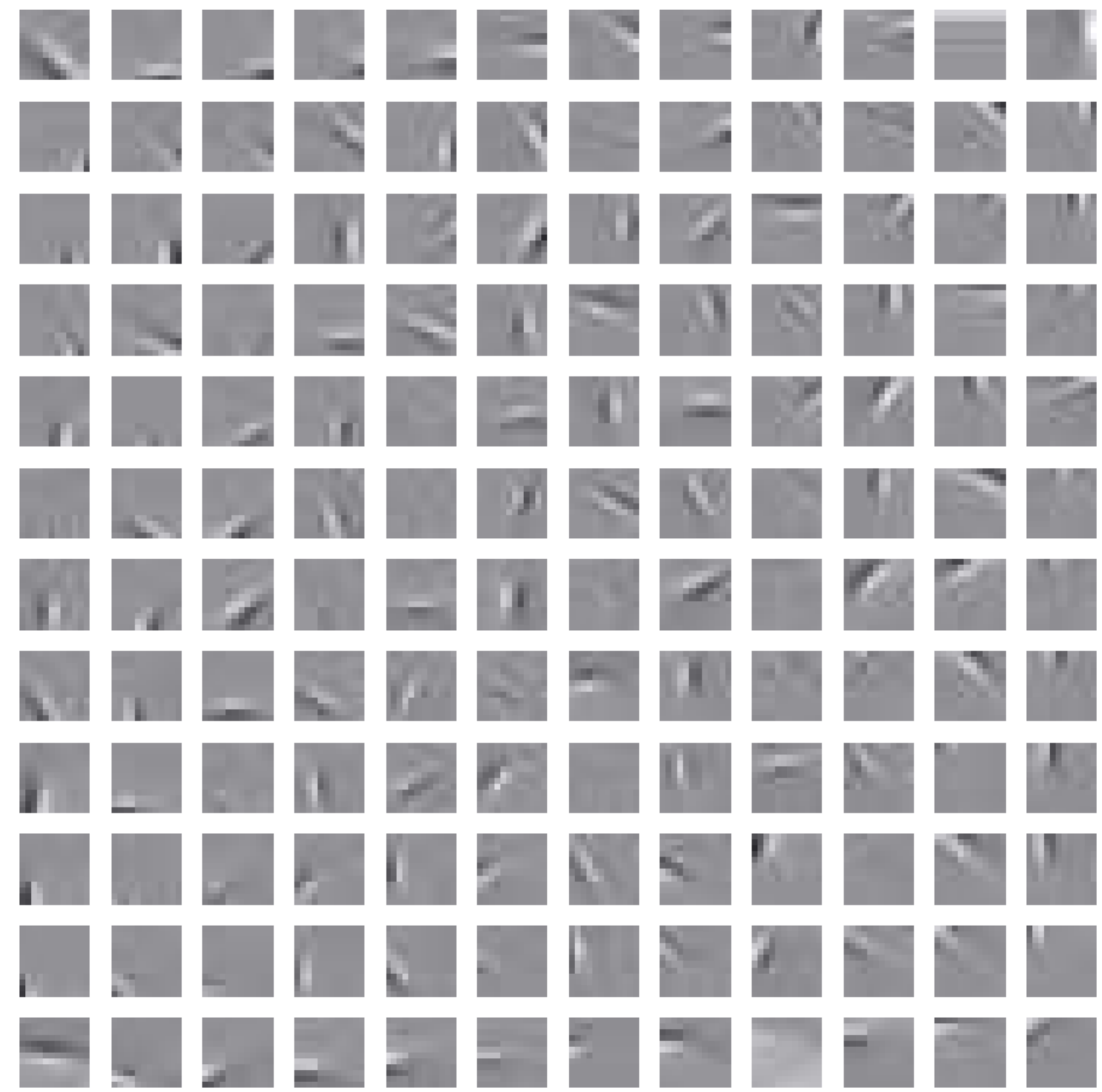
Actually, if the f_i were equal to the actual distributions of $\mathbf{w}_i^T \mathbf{x}$, the first term would be equal to $-\sum_i H(\mathbf{w}_i^T \mathbf{x})$. Thus the likelihood would be equal, up to an additive constant, to the negative of mutual information as given in Eq. (28).

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



Introducing ICA

Aapo Hyvärinen and Erkki Oja
Neural Networks, 13(4-5):411-430, 2000

To rigorously define ICA (Jutten and Hérault, 1991; Comon, 1994), we can use a statistical "latent variables" model. Assume that we observe n linear mixtures $x_1, \ldots, x_n$ of n independent components

$$x_j = a_{j1}s_1 + a_{j2}s_2 + \ldots + a_{jn}s_n, \text{ for all } j. \tag{3}$$

We have now dropped the time index t; in the ICA model, we assume that each mixture x_j as well as each independent component s_k is a random variable, instead of a proper time signal. The observed values $x_j(t)$, e.g., the microphone signals in the cocktail party problem, are then a sample of this random variable. Without loss of generality, we can assume that both the mixture variables and the independent components have zero mean: If this is not true, then the observable variables x_i can always be centered by subtracting the sample mean, which makes the model zero-mean.

$s_1, \ldots, s_n$. Let us denote by $\mathbf{A}$ the matrix with elements a_{ij}. Generally, bold lower case letters indicate vectors and bold upper-case letters denote matrices. All vectors are understood as column vectors; thus $\mathbf{x}^T$, or the transpose of $\mathbf{x}$, is a row vector. Using this vector-matrix notation, the above mixing model is written as

$$\mathbf{x} = \mathbf{A}\mathbf{s}. \tag{4}$$

The starting point for ICA is the very simple assumption that the components s_i are statistically *independent*. Statistical independence will be rigorously defined in Section 3. It will be seen below that we must also assume that the independent component must have *nongaussian* distributions. However, in the basic model we do *not* assume these distributions known (if they are known, the problem is considerably simplified.) For simplicity, we are also assuming that the unknown mixing matrix is square, but this assumption can be sometimes relaxed, as explained in Section 4.5. Then, after estimating the matrix $\mathbf{A}$, we can compute its inverse, say $\mathbf{W}$, and obtain the independent component simply by:

$$\mathbf{s} = \mathbf{W}\mathbf{x}. \tag{6}$$