Part 2: Coding Natural Images

A general approach to coding: redundancy reduction

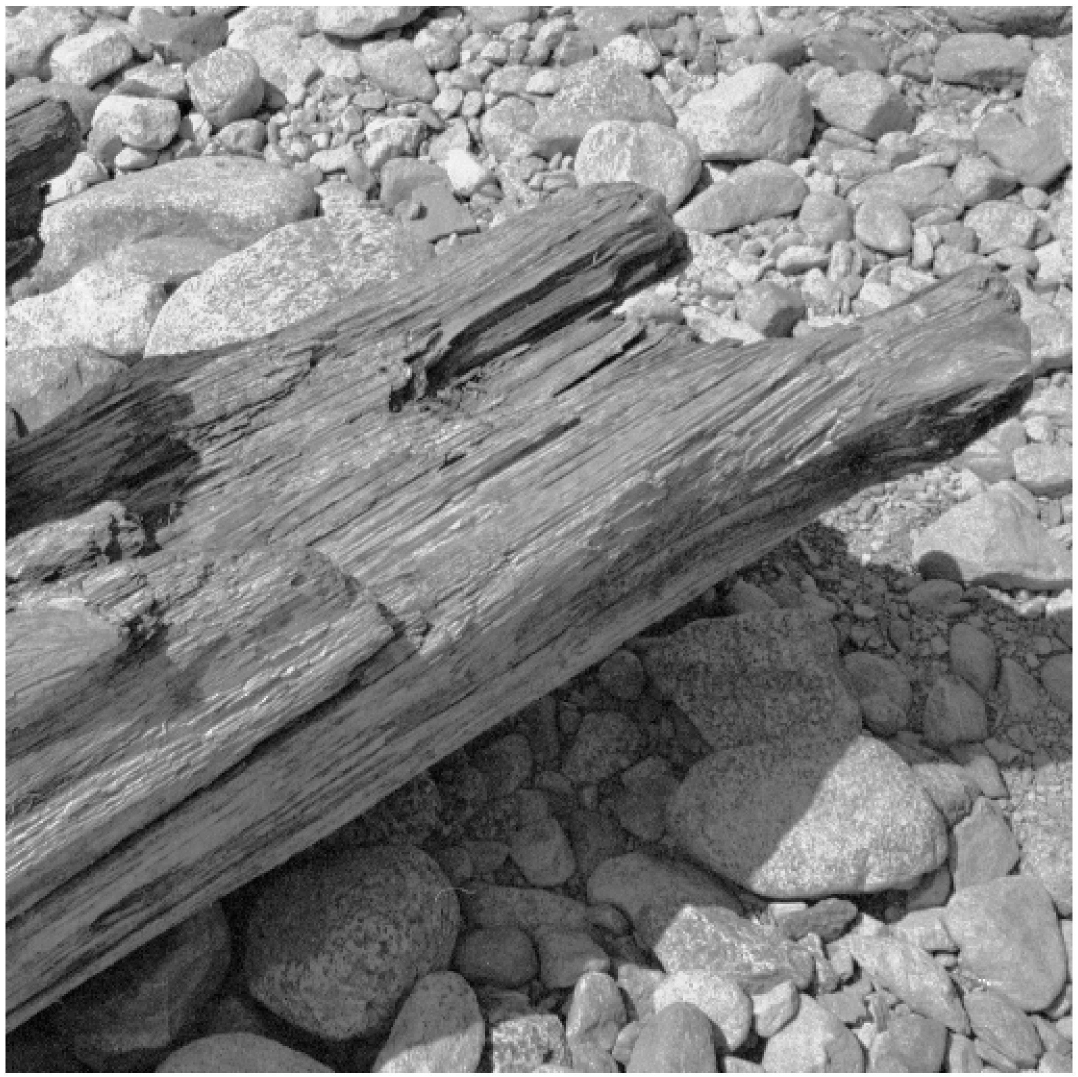
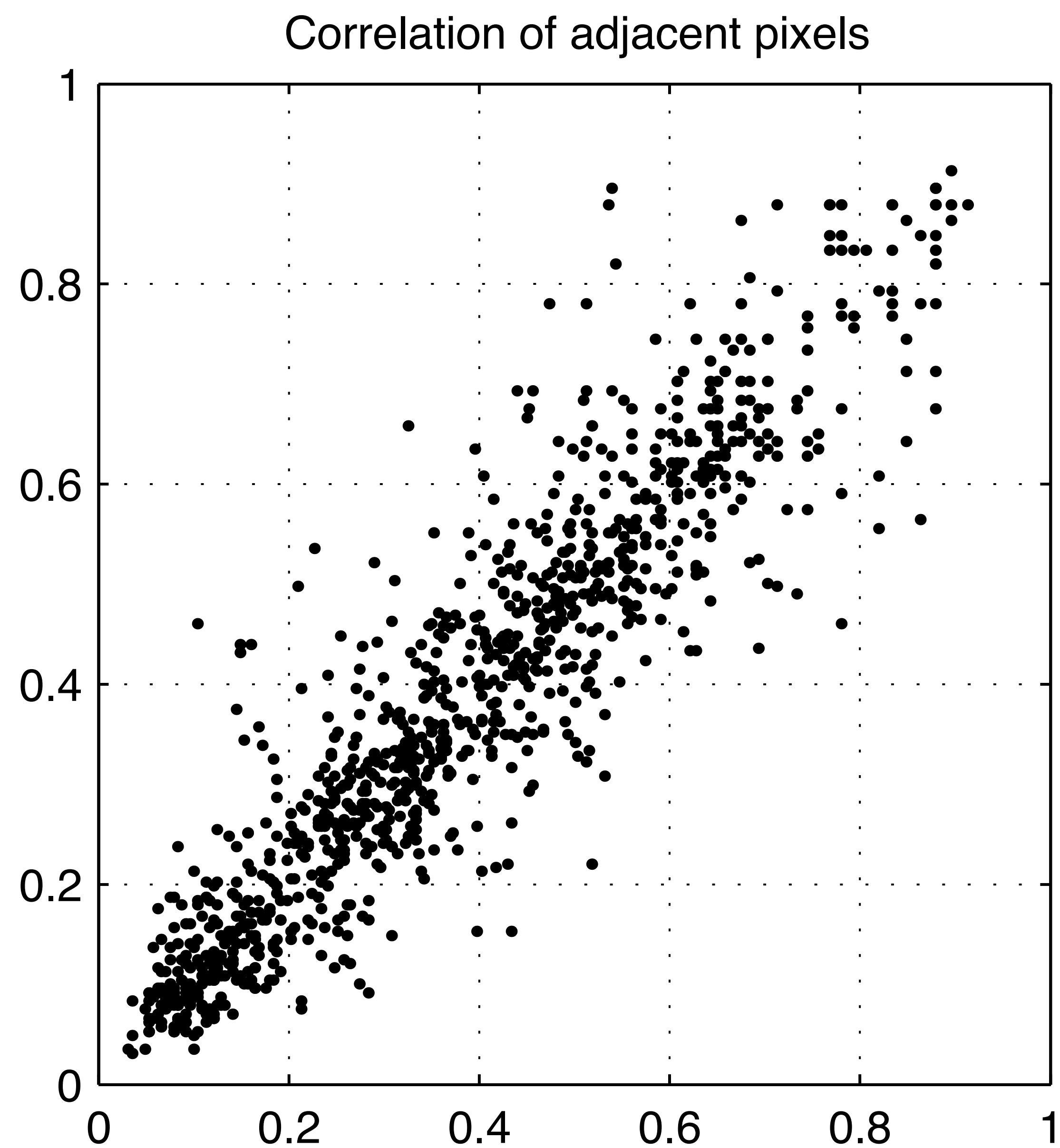


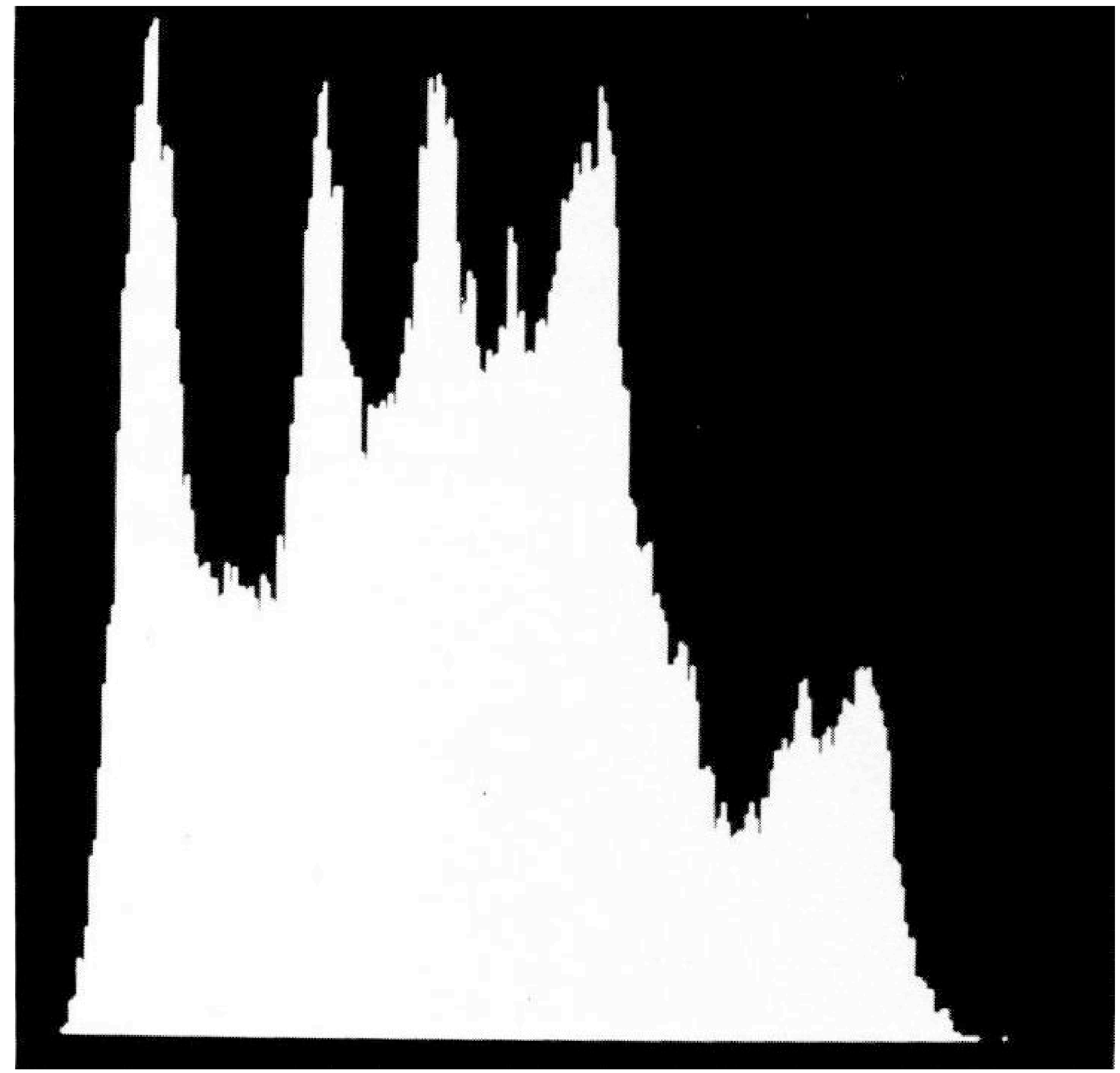
image from Field (1994)



Redundancy reduction is equivalent to efficient coding.

Reducing pixel redundancy

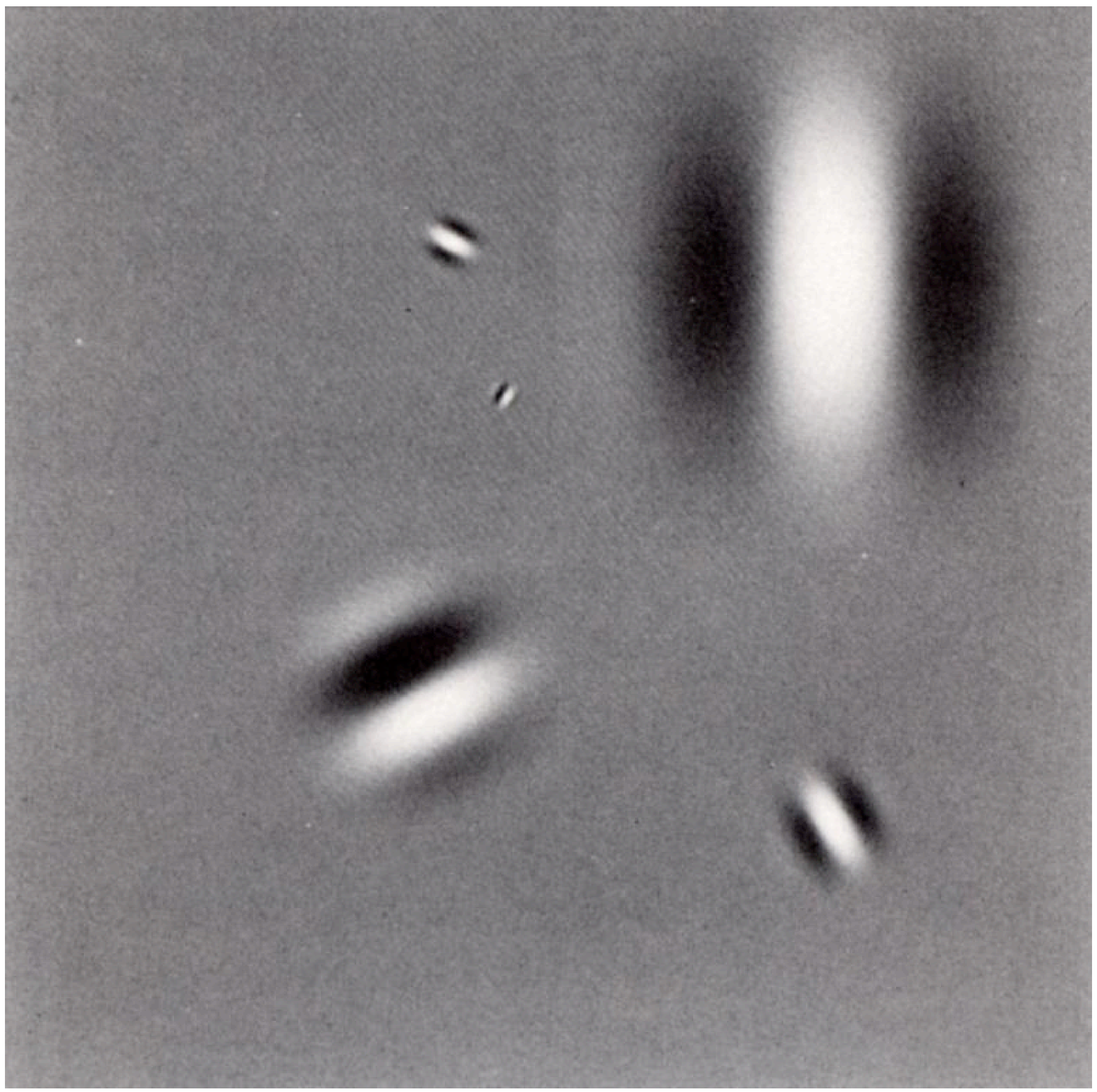
from Daugman (1990)



Lena: a standard 8 bit 256x256 gray scale image

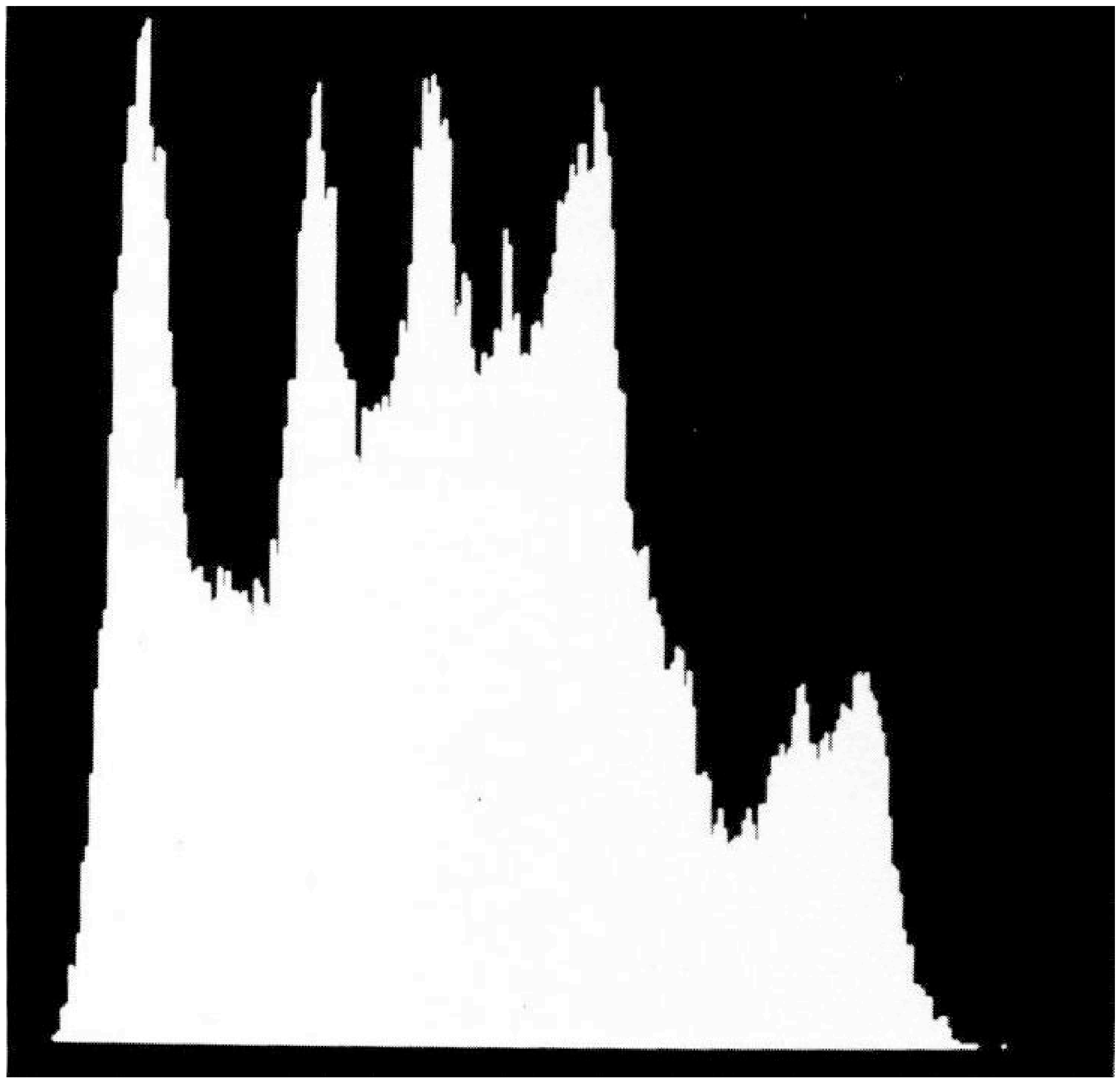
histogram of pixel values
Entropy = 7.57 bits

2D Gabor wavelet captures spatial structure

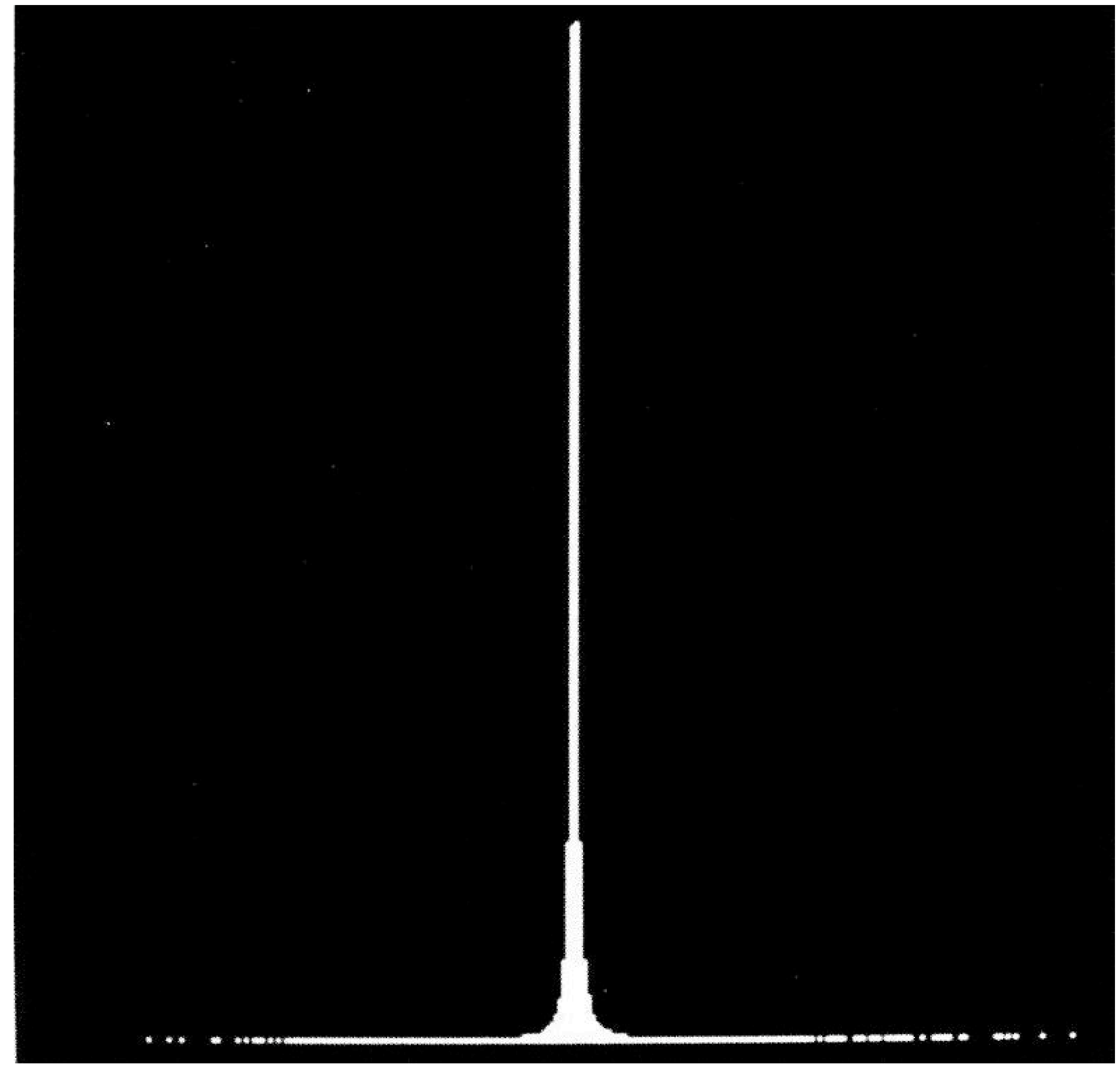


- 2D Gabor functions
- Wavelet basis generated by dilations, translations, and rotations of a single basis function
- Can also control phase and aspect ratio
- (drifting) Gabor functions are what the eye “sees best”

Recoding with Gabor functions



Pixel entropy = 7.57 bits



Recoding with 2D Gabor functions
Coefficient entropy = 2.55 bits

Describing signals with a simple statistical model

Principle

Good codes capture the statistical distribution of sensory patterns.

How do we describe the distribution?

- Goal is to encode the data to desired precision

$$
\begin{aligned}
\mathbf{x} &= \vec{a}_1 s_1 + \vec{a}_2 s_2 + \cdots + \vec{a}_L s_L + \vec{\epsilon} \\
&= \mathbf{A}\mathbf{s} + \boldsymbol{\epsilon}
\end{aligned}
$$

- Can solve for the coefficients in the no noise case

$$
\hat{\mathbf{s}} = \mathbf{A}^{-1}\mathbf{x}
$$

An algorithm for deriving efficient linear codes: ICA

Learning objective:

maximize coding efficiency

⇒ maximize $P(\mathbf{x}|\mathbf{A})$ over $\mathbf{A}$.

Probability of the pattern ensemble is:

$$P(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N|\mathbf{A}) = \prod_k P(\mathbf{x}_k|\mathbf{A})$$

To obtain $P(\mathbf{x}|\mathbf{A})$ marginalize over $\mathbf{s}$:

$$\begin{aligned} P(\mathbf{x}|\mathbf{A}) &= \int d\mathbf{s}\, P(\mathbf{x}|\mathbf{A}, \mathbf{s}) P(\mathbf{s}) \\ &= \frac{P(\mathbf{s})}{|\det \mathbf{A}|} \end{aligned}$$

Using independent component analysis (ICA) to optimize $\mathbf{A}$:

$$\begin{aligned} \Delta\mathbf{A} &\propto \mathbf{A}\mathbf{A}^T \frac{\partial}{\partial\mathbf{A}} \log P(\mathbf{x}|\mathbf{A}) \\ &= -\mathbf{A}(\mathbf{z}\mathbf{s}^T - \mathbf{I}) \end{aligned}$$

where z = (log P(s))’.

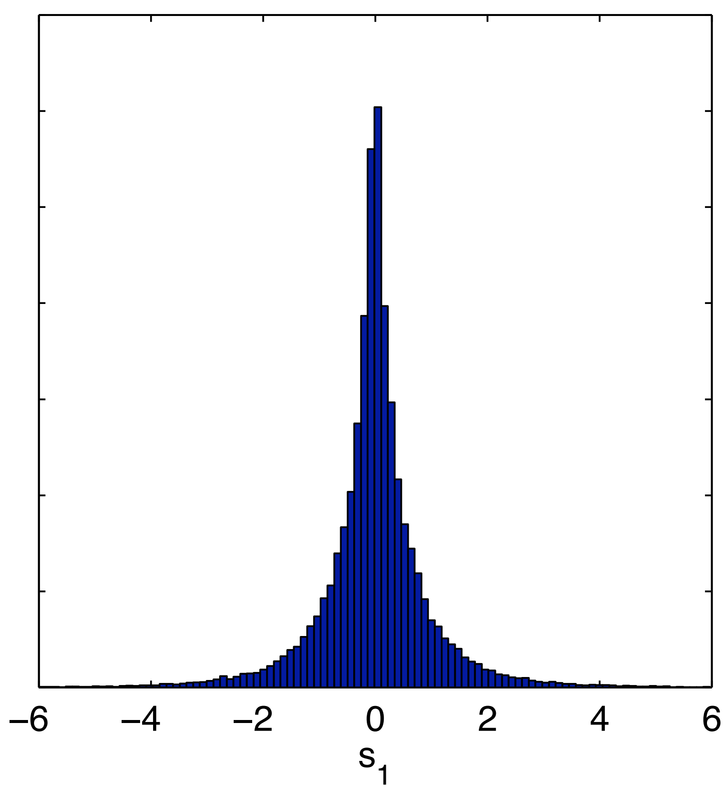
This learning rule:

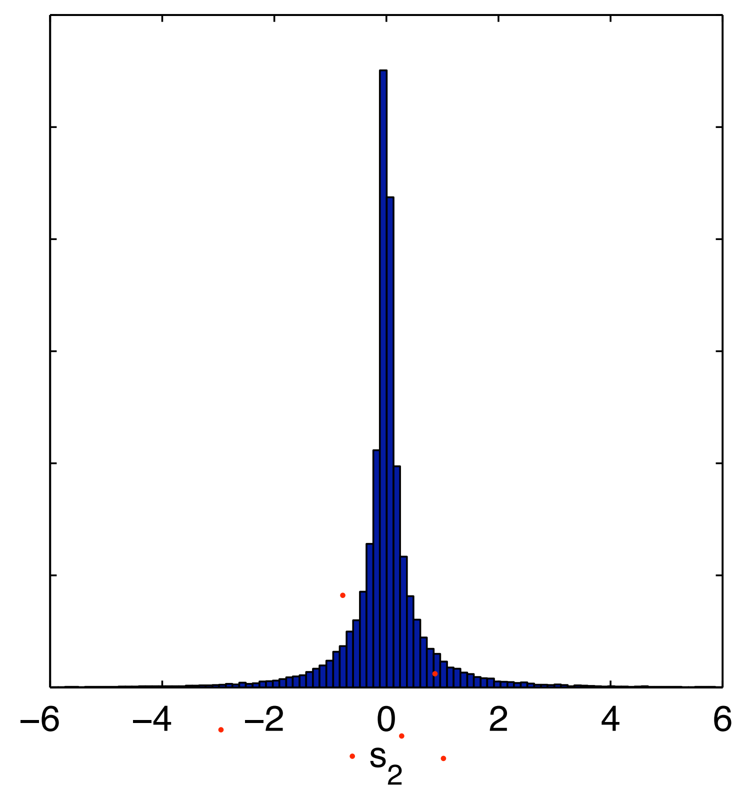
- learns the features that capture the most structure
- optimizes the efficiency of the code

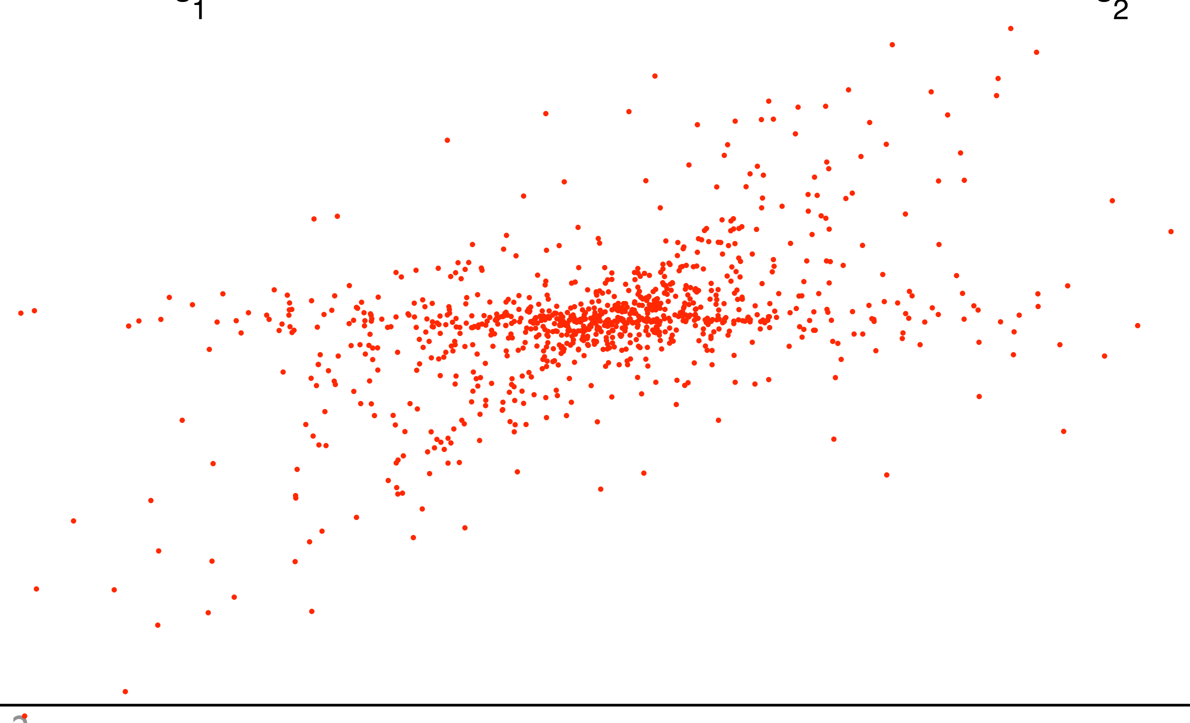
What should we use for $P(\mathbf{s})$?

Modeling Non-Gaussian distributions

- Typical coeff. distributions of natural signals are *non-Gaussian*.







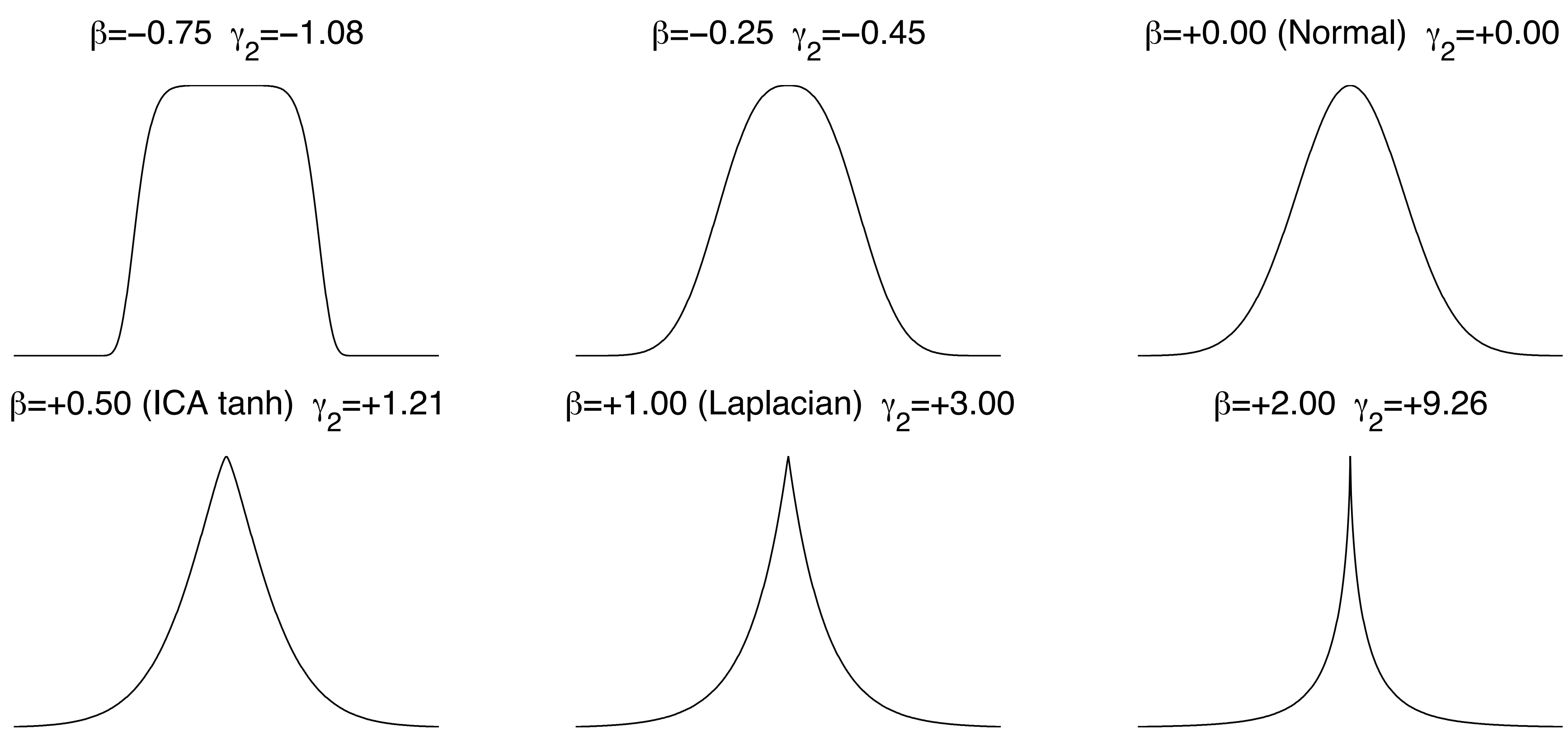
The generalized Gaussian distribution

$$P(x|q) \propto \exp(-\frac{1}{2}|x|^q)$$

- Or equivalently, and exponential power distribution (Box and Tiao, 1973):

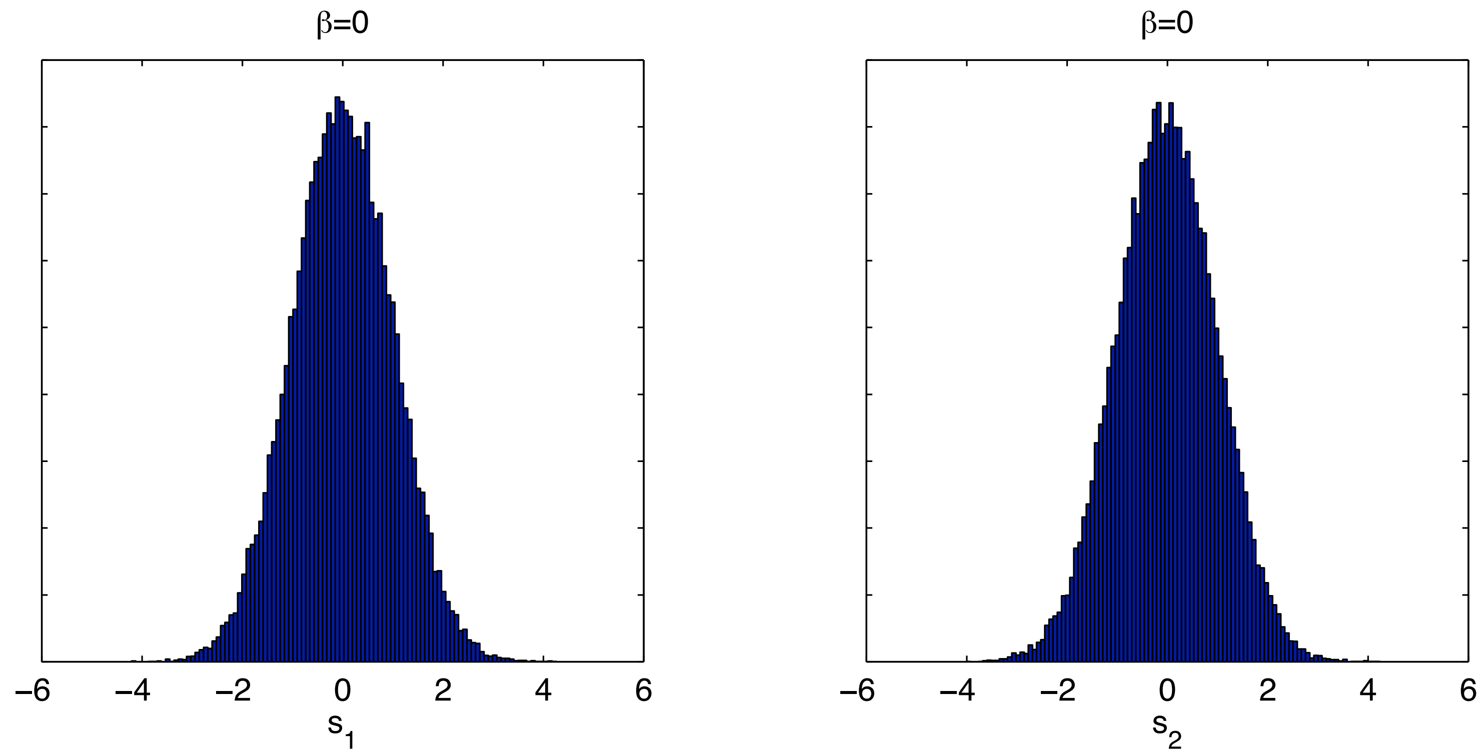
$$P(x|\mu,\sigma,\beta) = \frac{\omega(\beta)}{\sigma} \exp\left[-c(\beta)\left|\frac{x-\mu}{\sigma}\right|^{2/(1+\beta)}\right]$$

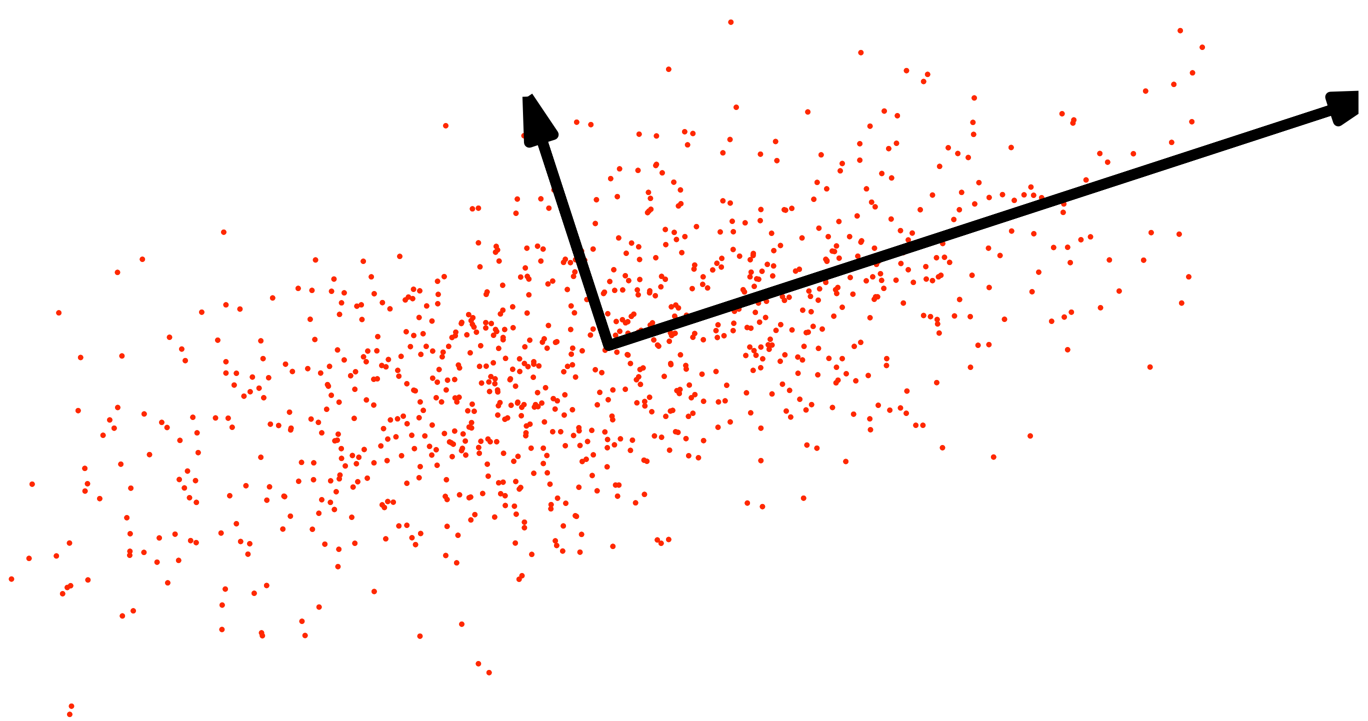
- β varies monotonically with the kurtosis, γ_2:



Modeling Gaussian distributions with PCA

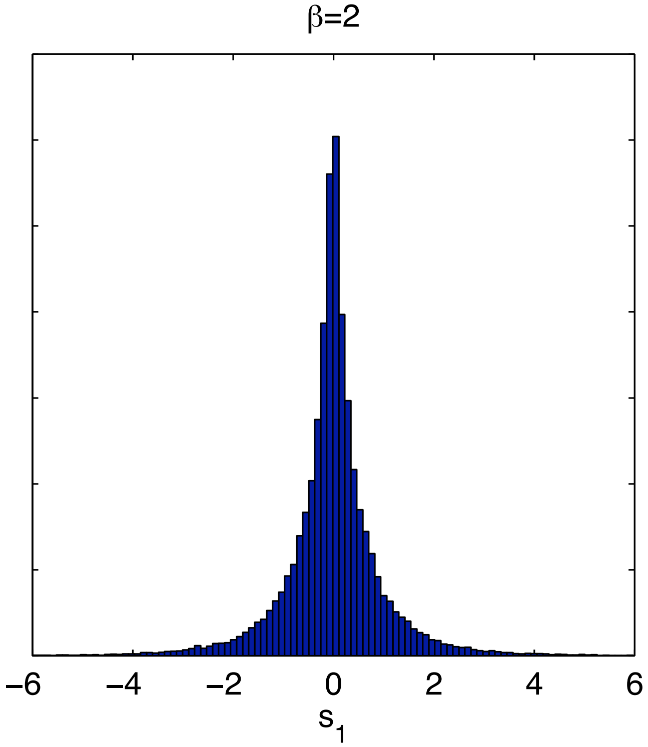
- Principal component analysis (PCA) describes the principal axes of variation in the data distribution.
- This is equivalent to fitting the data with a multivariate Gaussian.

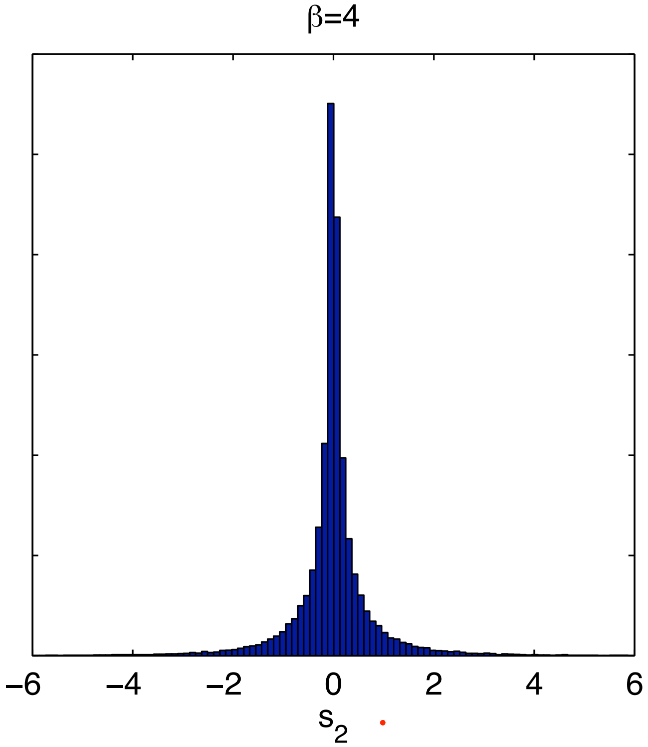




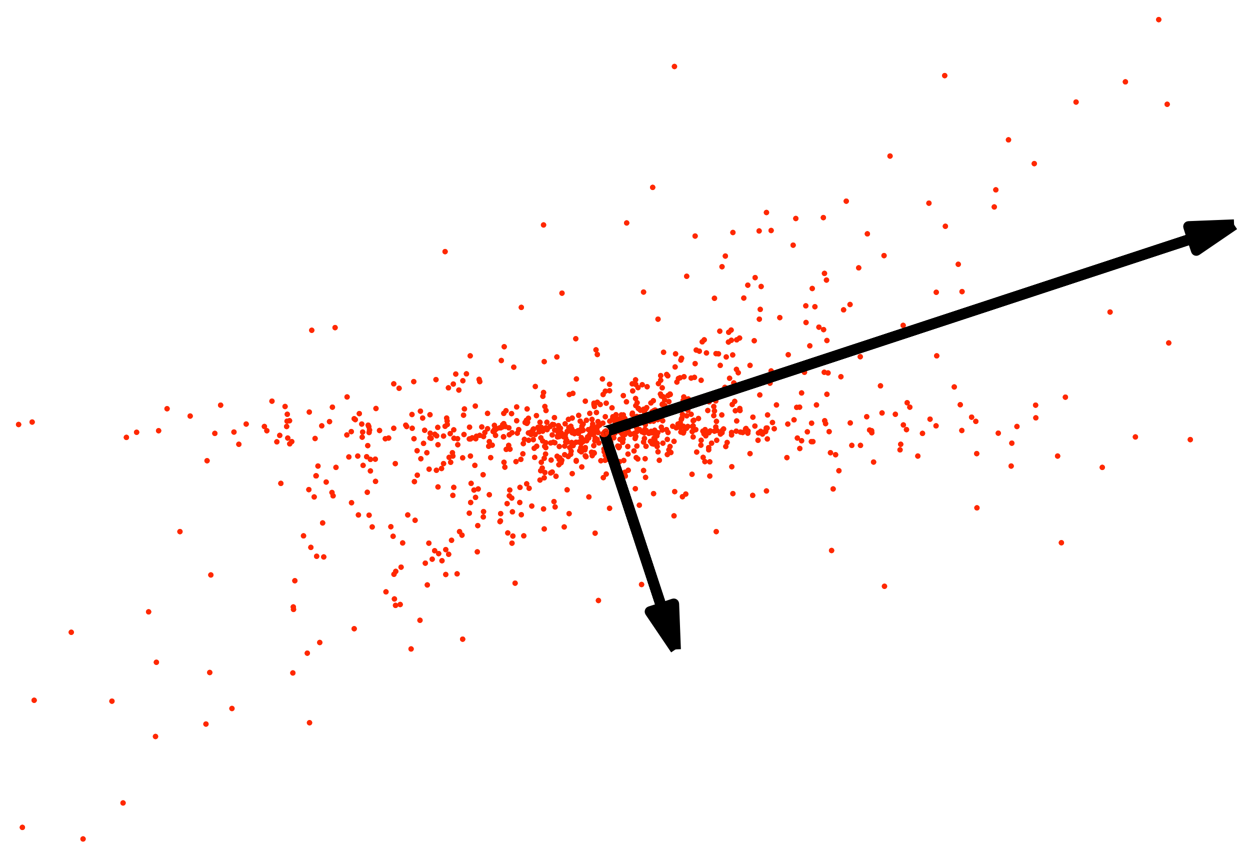
Modeling non-Gaussian distributions

- What about non-Gaussian marginals?



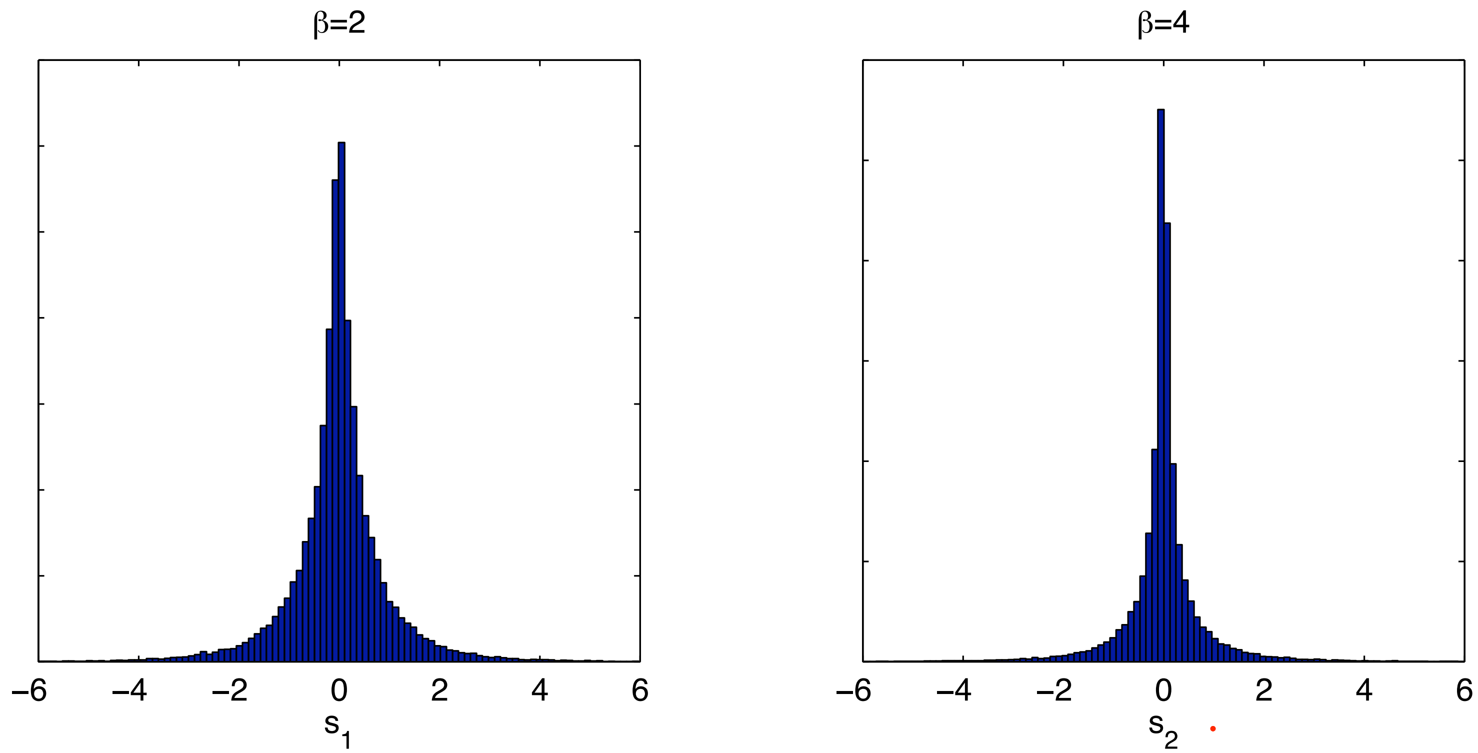


- How would this distribution be modeled by PCA?



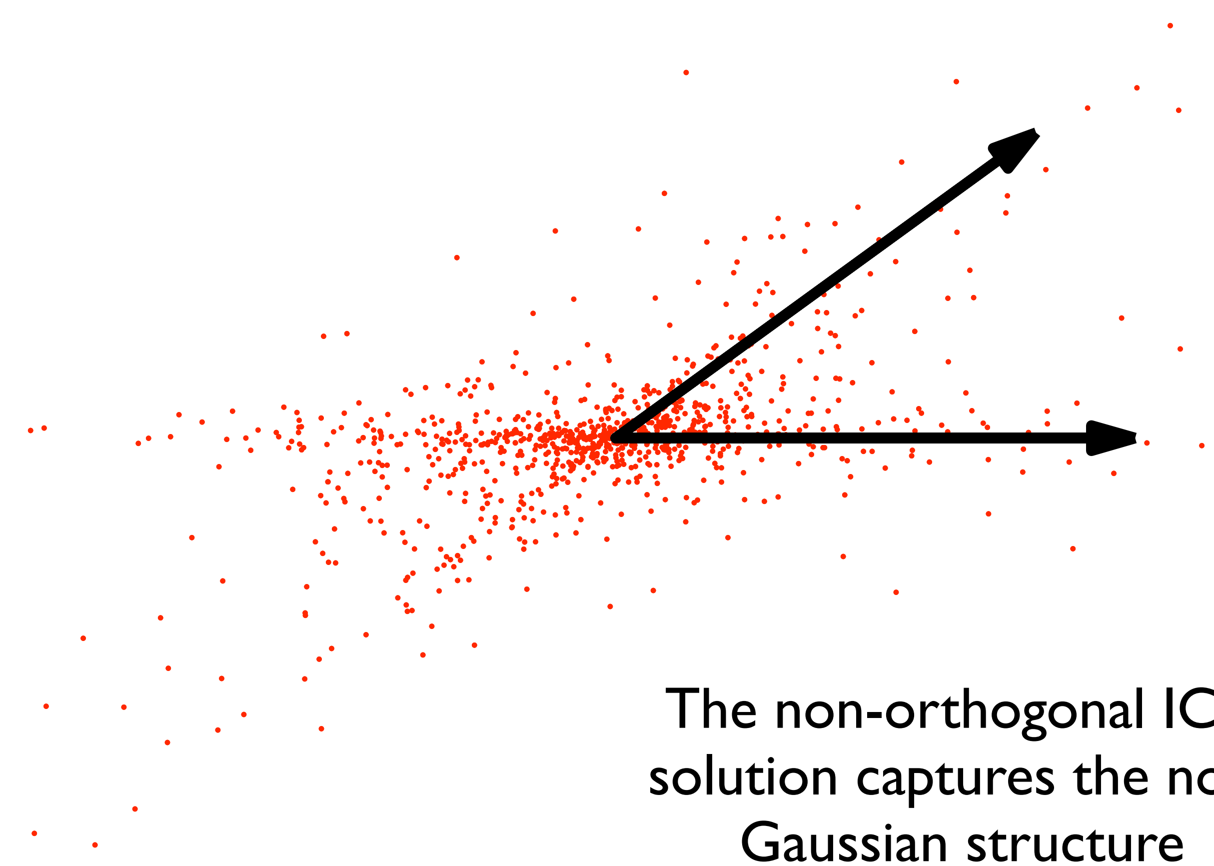
Modeling non-Gaussian distributions

- What about non-Gaussian marginals?



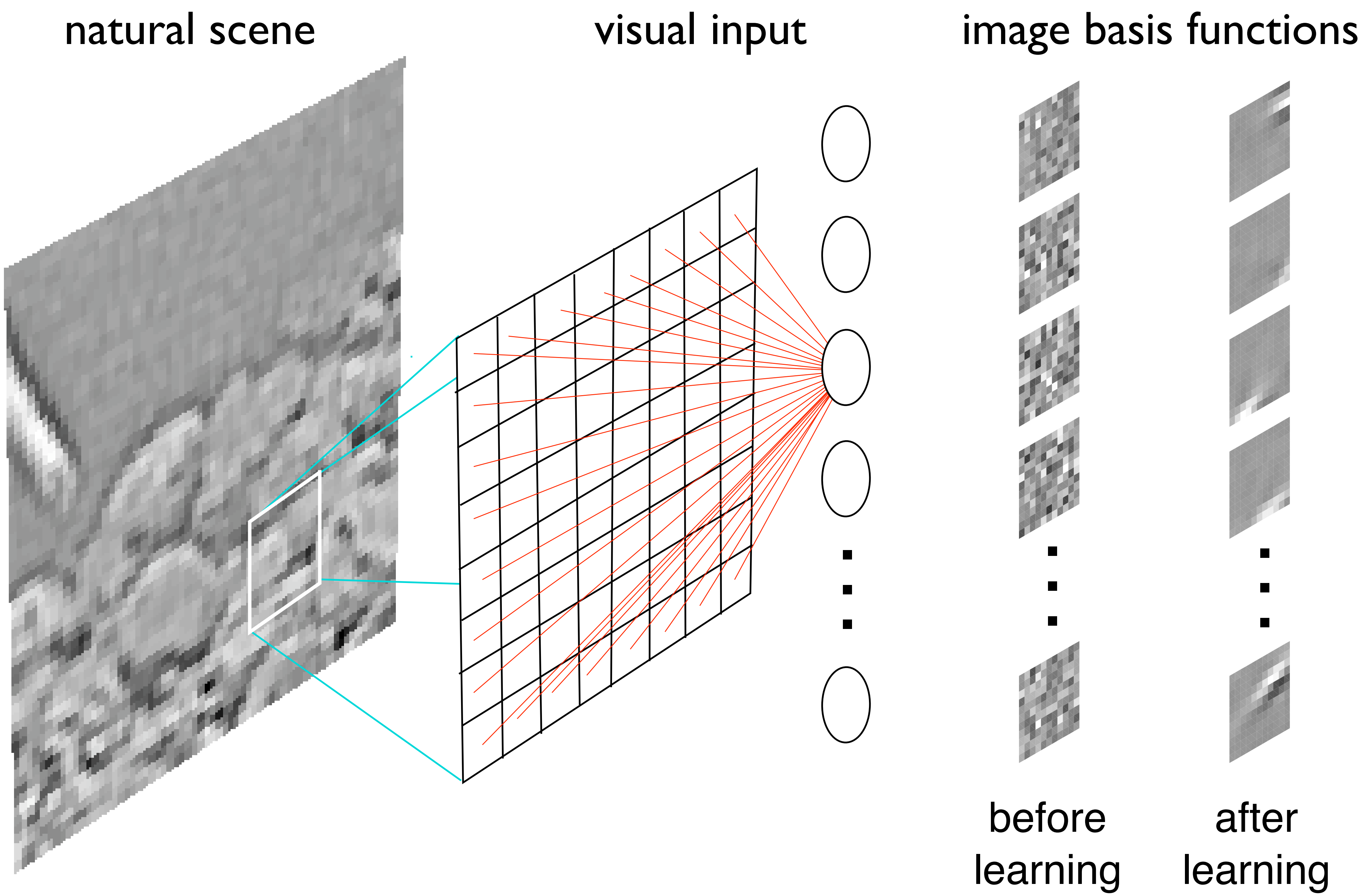
- How would this distribution be modeled by PCA?

- How should the distribution be described?



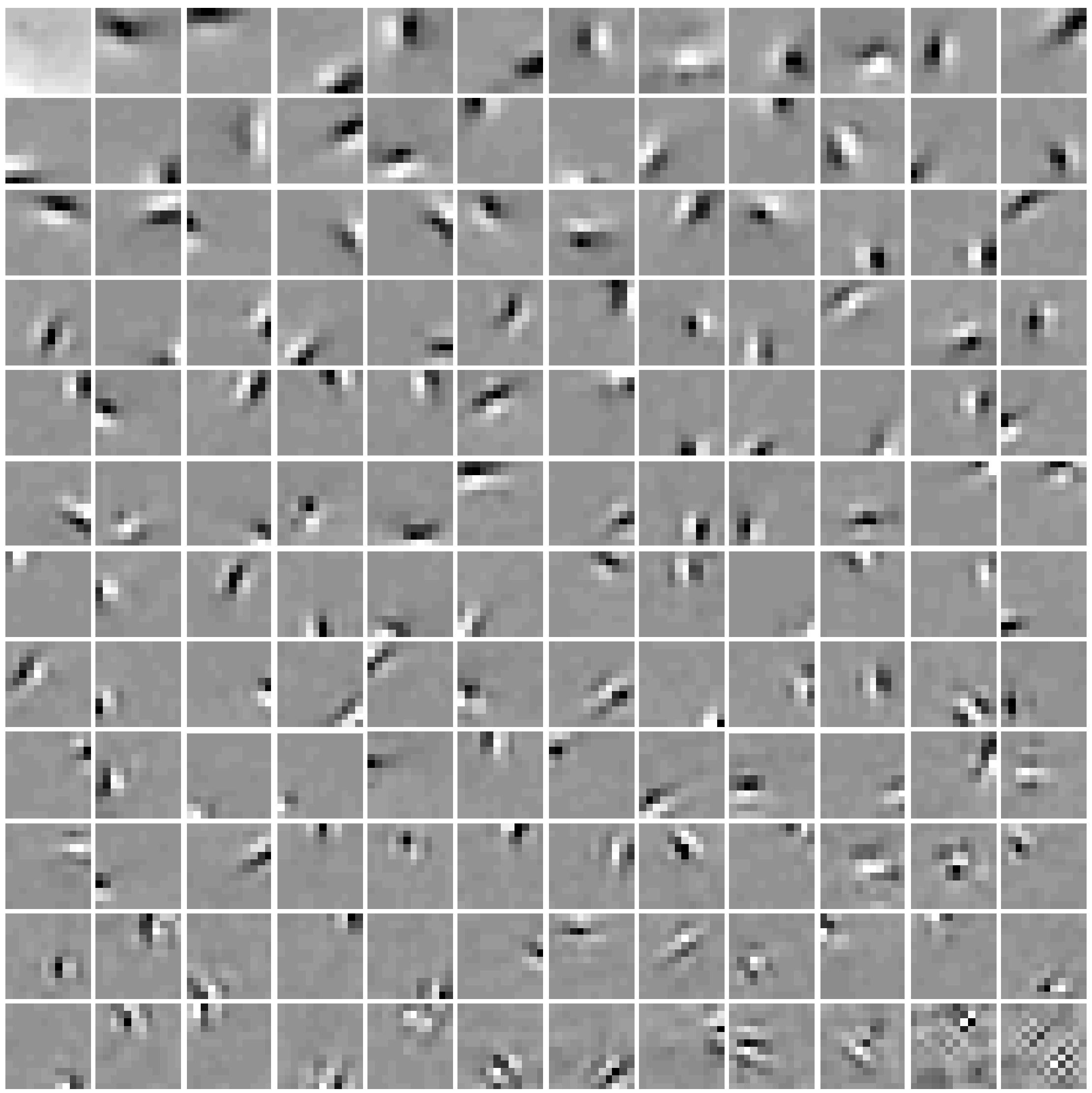
The non-orthogonal ICA solution captures the non-Gaussian structure

Efficient coding of natural images

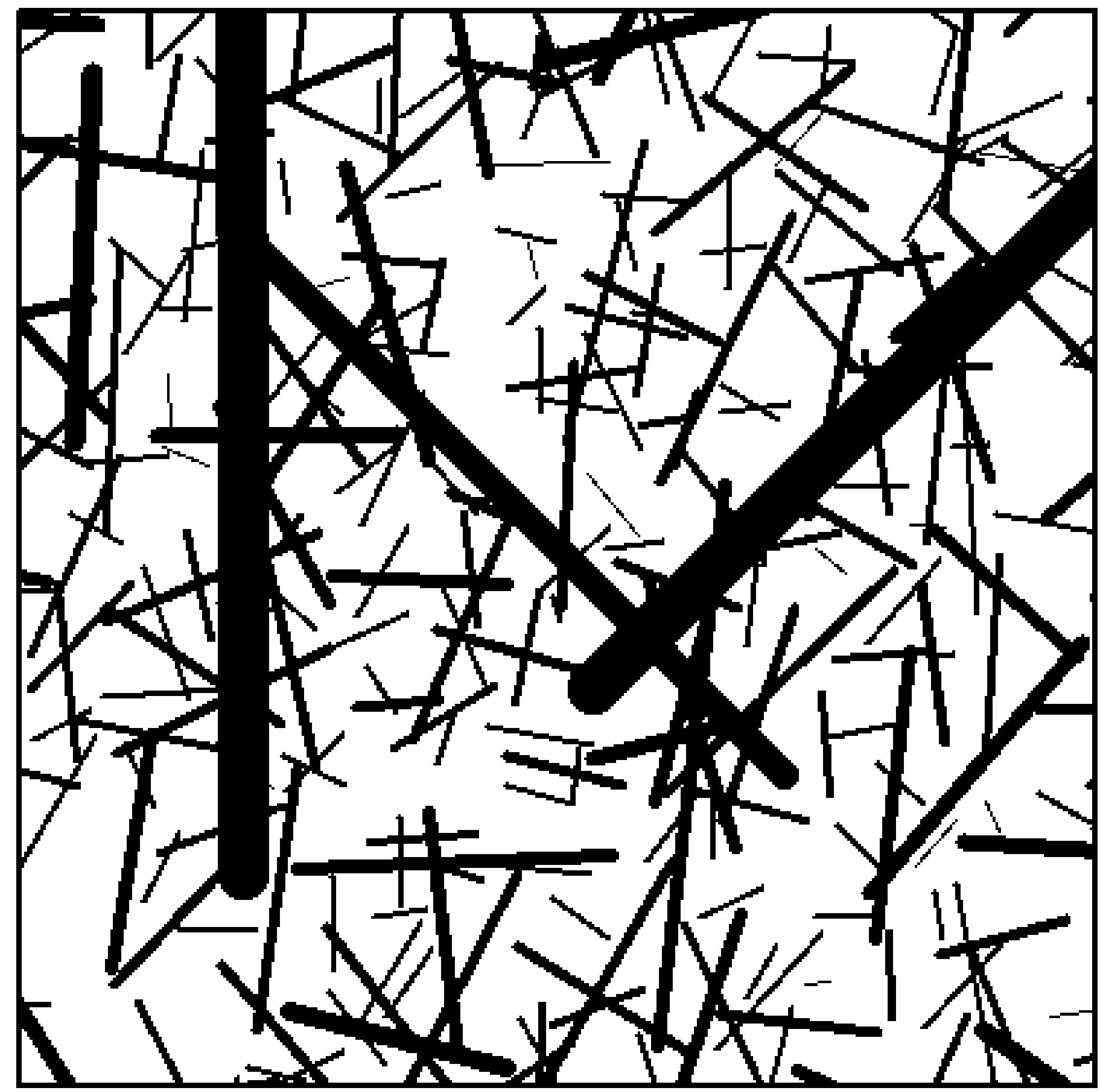


Network weights are adapted to maximize coding efficiency:
minimizes redundancy and maximizes the independence of the outputs

Model predicts local and global receptive field properties



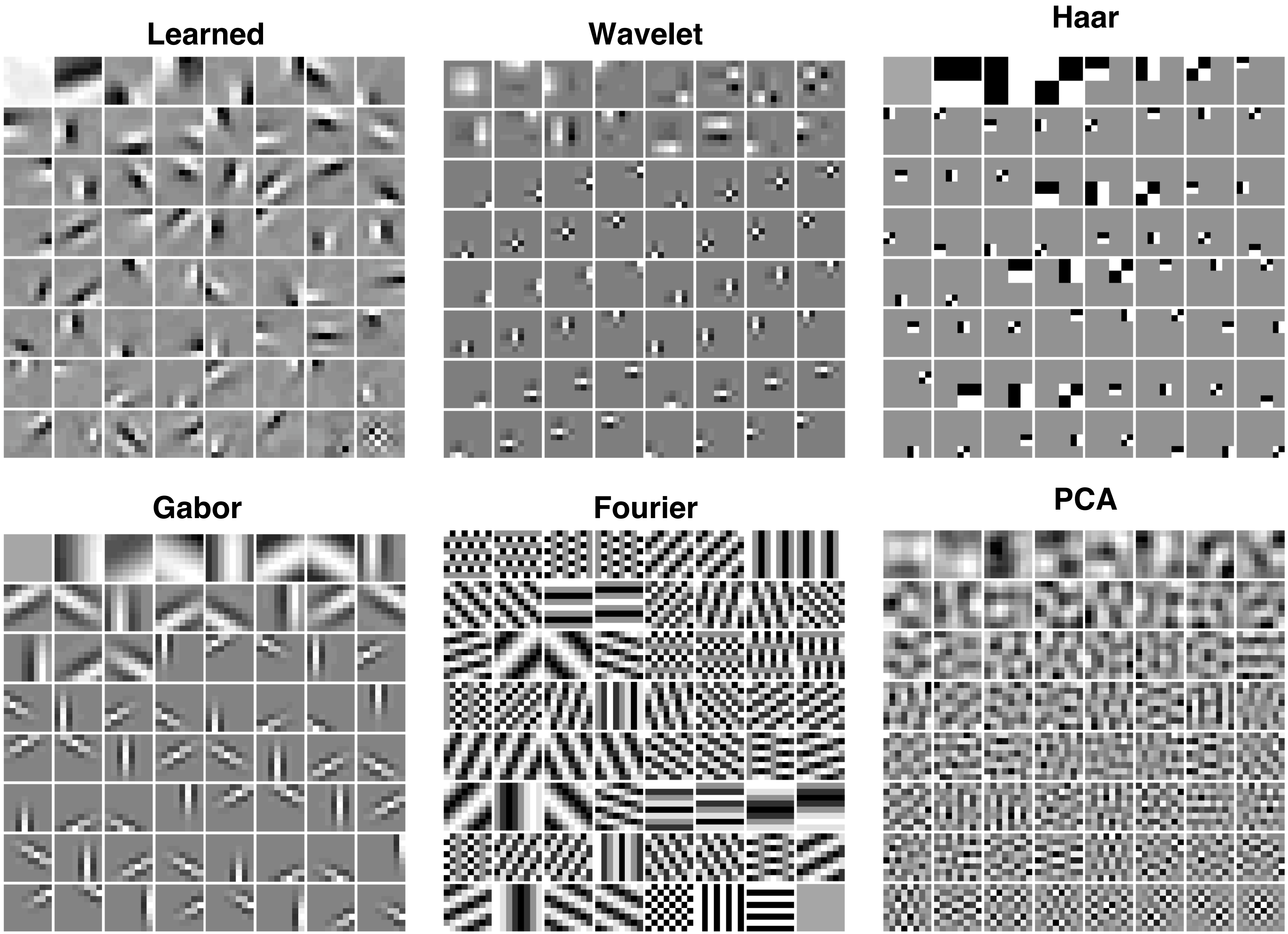
Learned basis for natural images



Overlaid basis function properties

from Lewicki and Olshausen, 1999

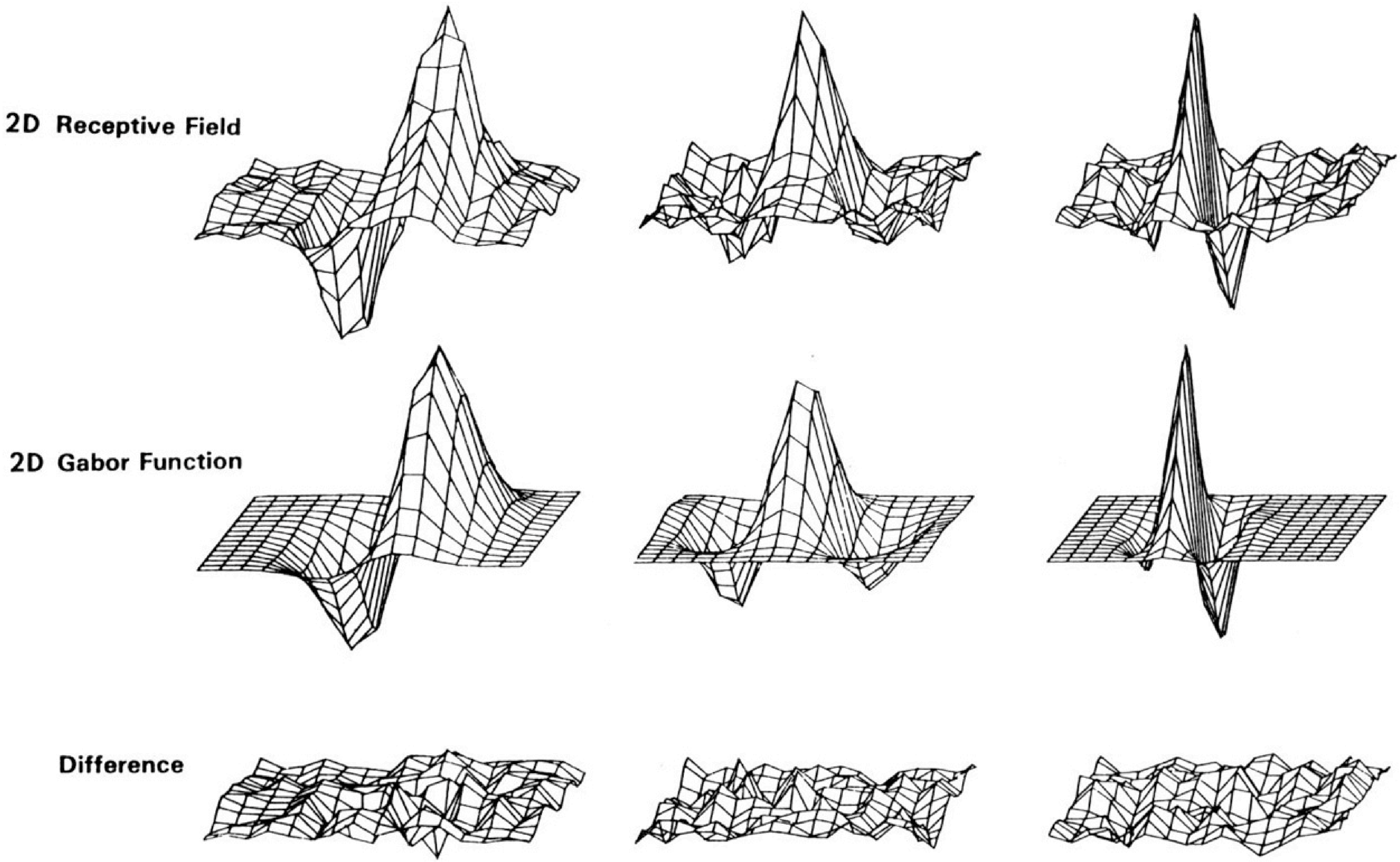
Algorithm selects best of many possible sensory codes



from Lewicki and Olshausen, 1999

Theoretical perspective: Not edge "detectors."
An efficient code for natural images.

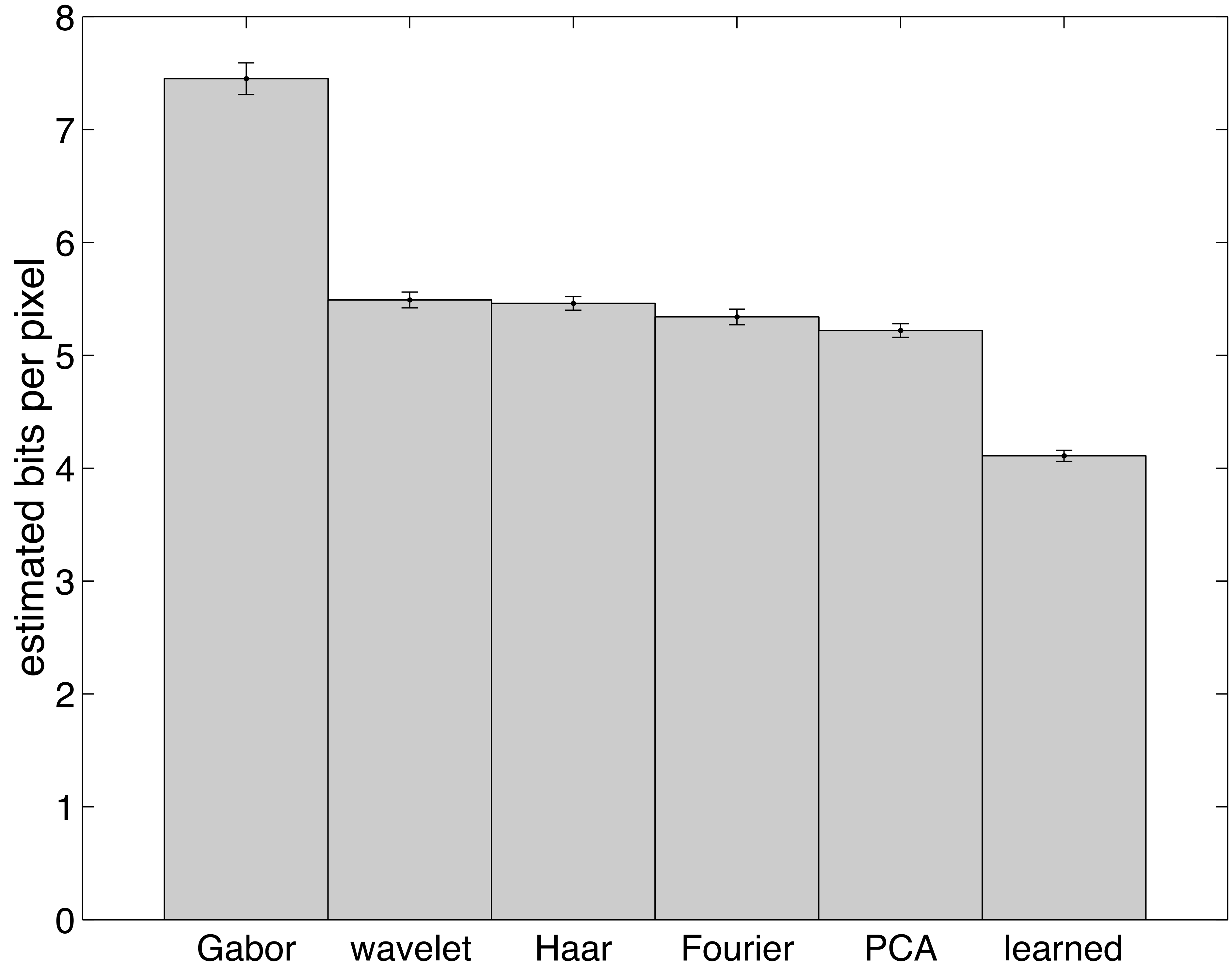
2D Receptive fields in primary visual cortex



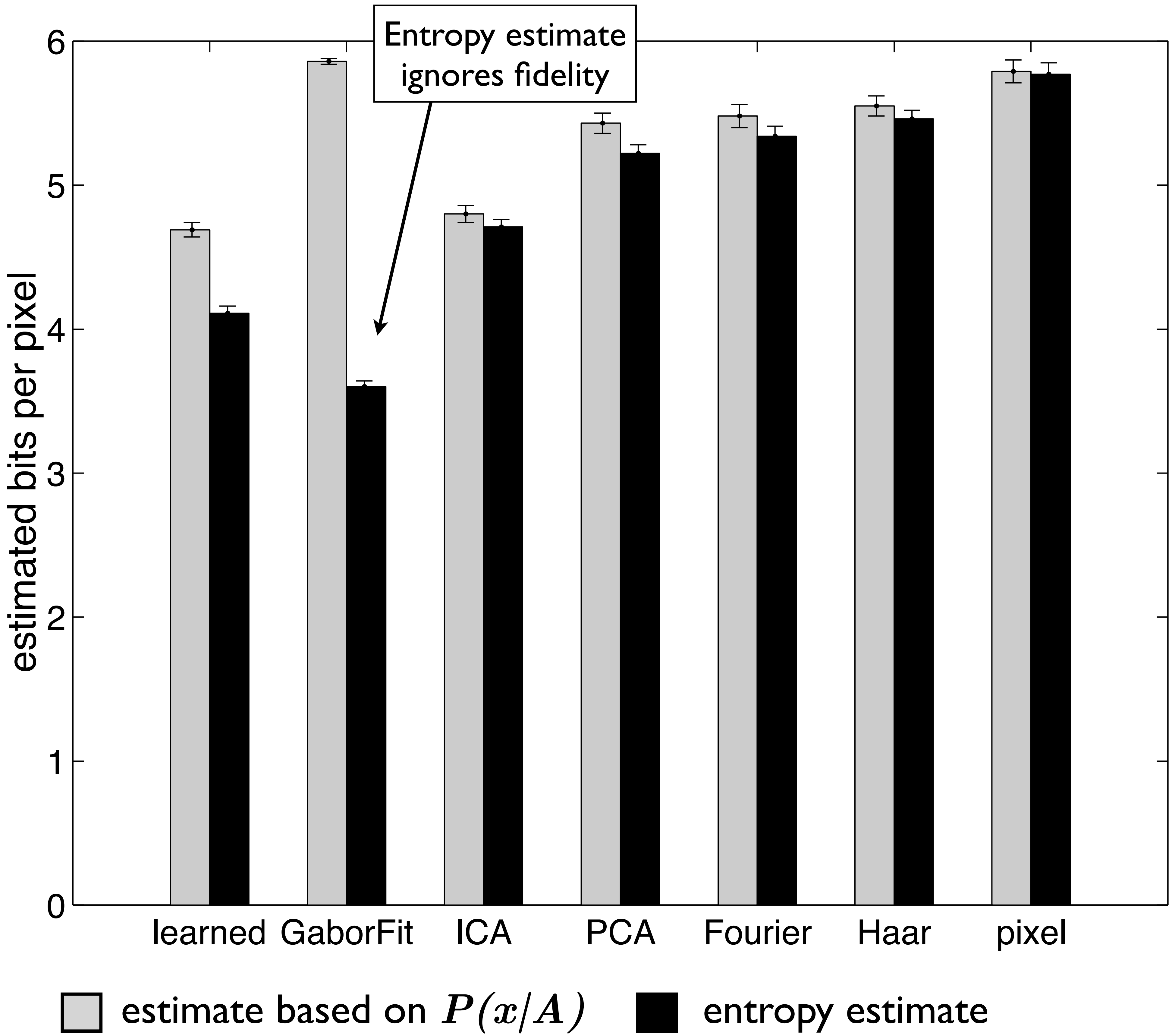
Fit of 2D Gabor wavelet is indistinguishable from noise.

figure from Daugman, 1990
data from Jones and Palmer, 1987

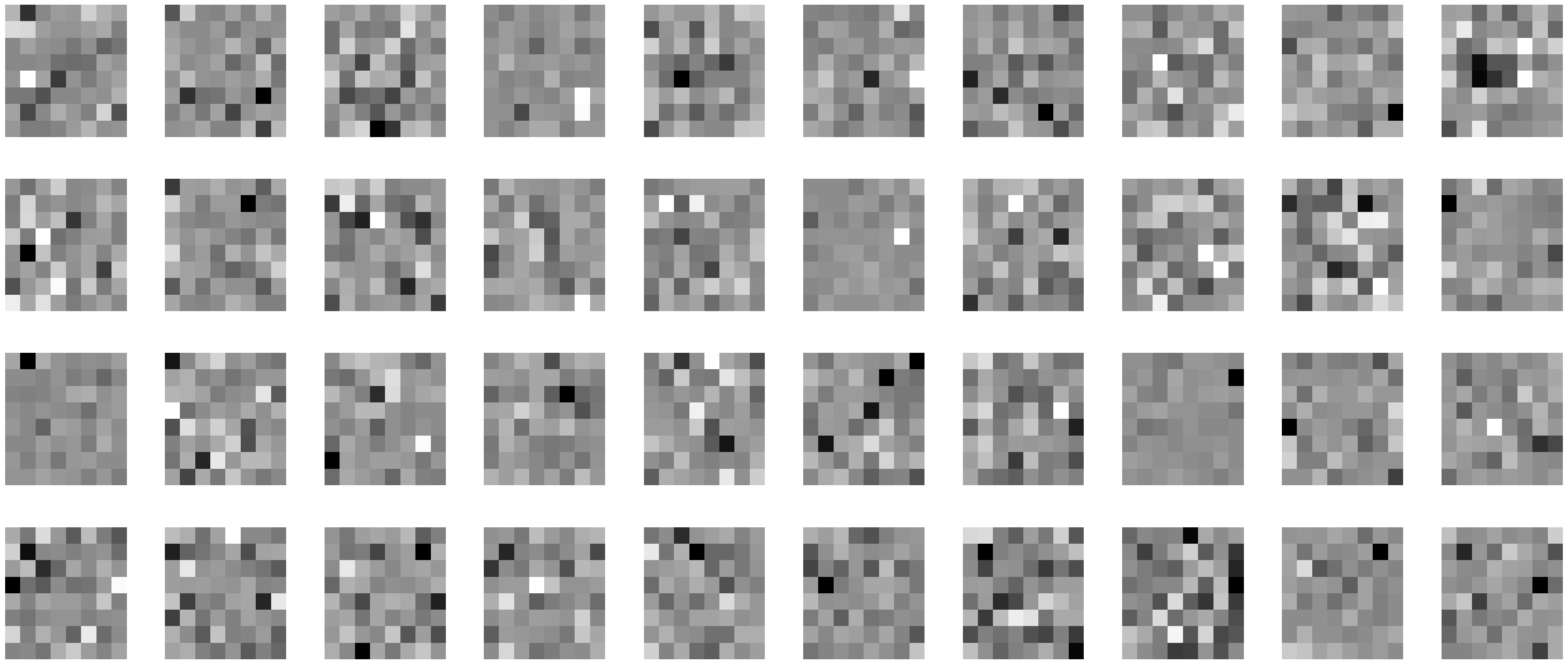
Comparing coding efficiency on natural images



Comparing efficiency estimates using entropy and probability

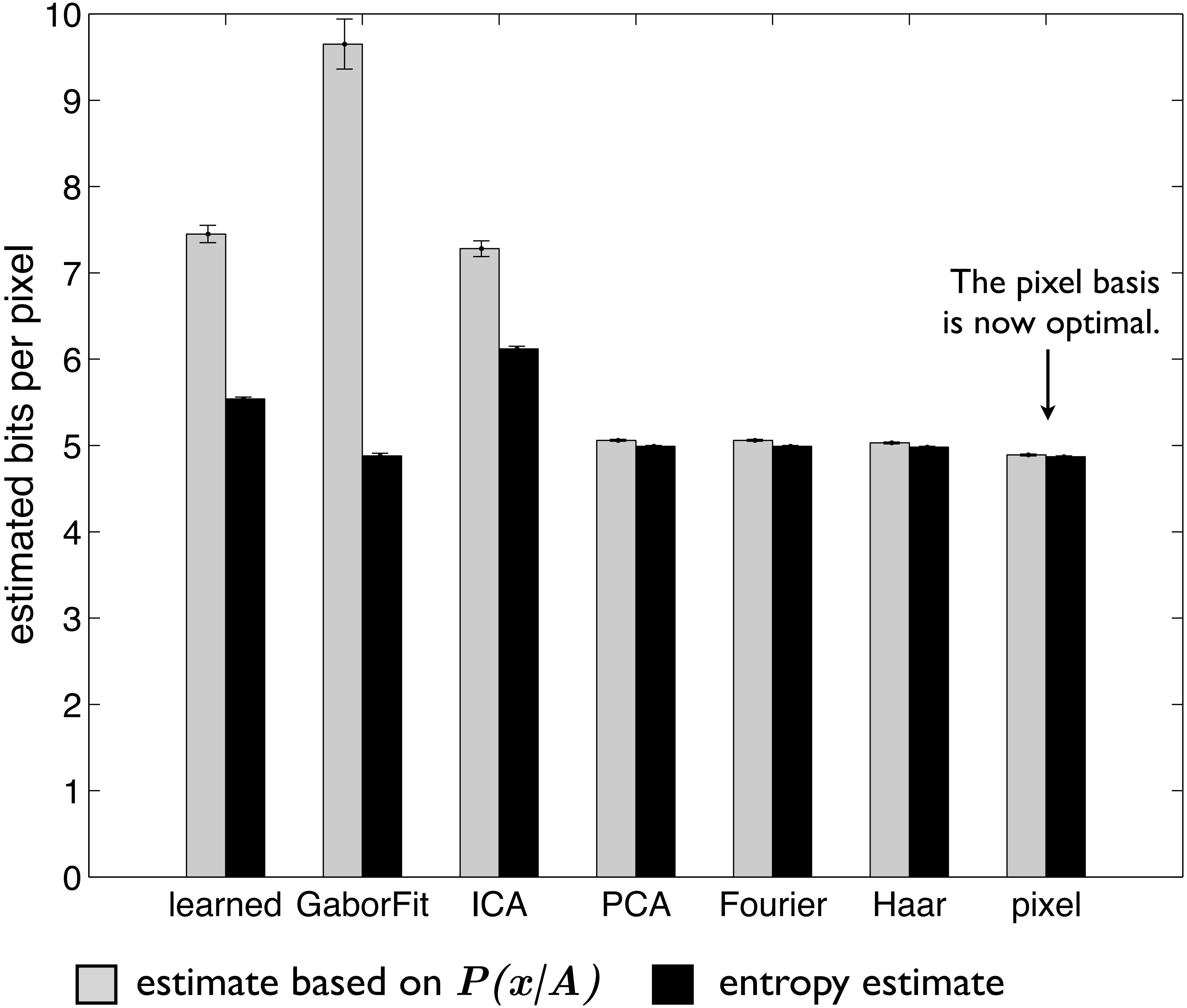


Optimality depends on data

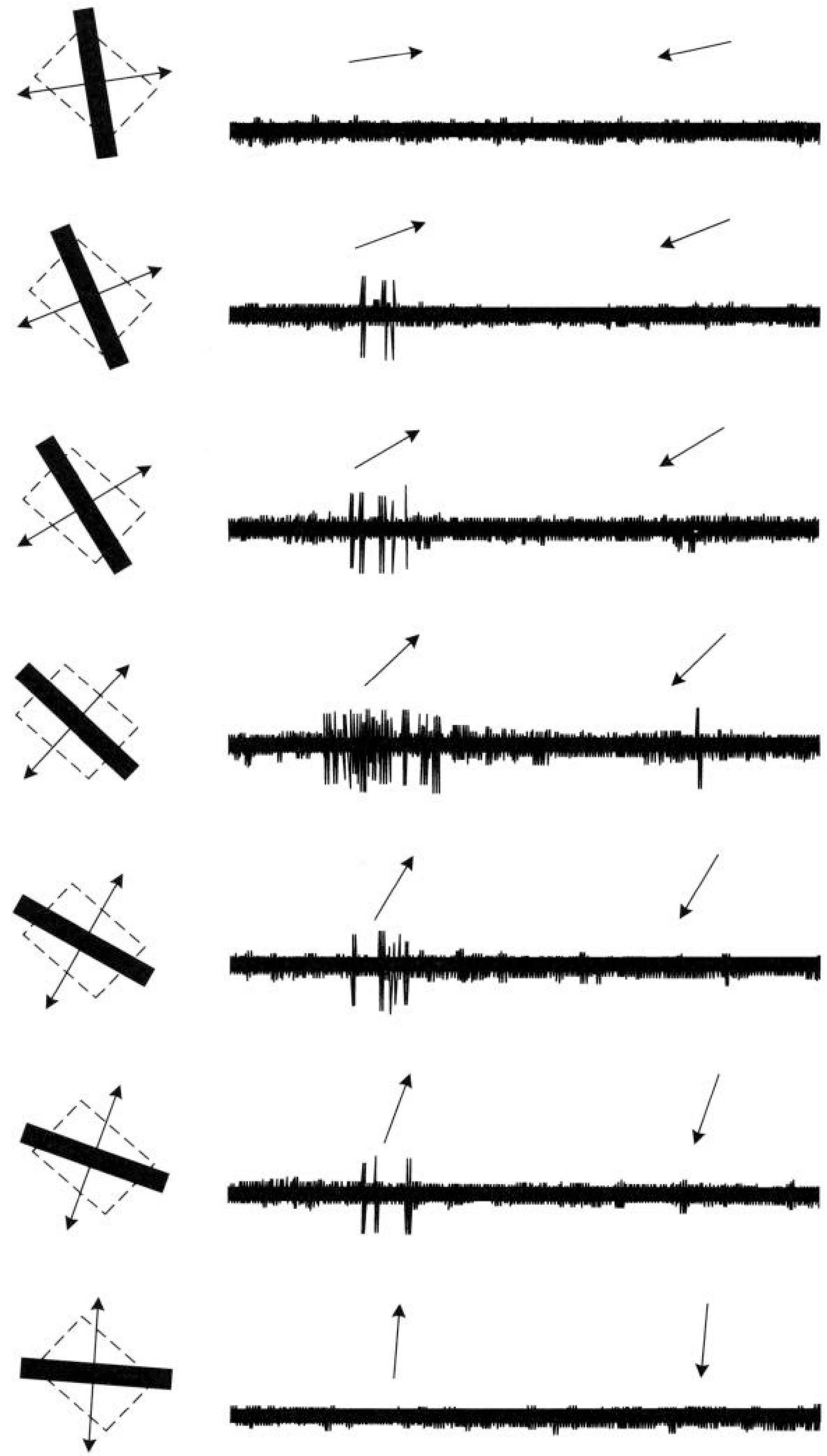


Which code will be best for random, sparse pixels?

Now the coding efficiencies are reversed

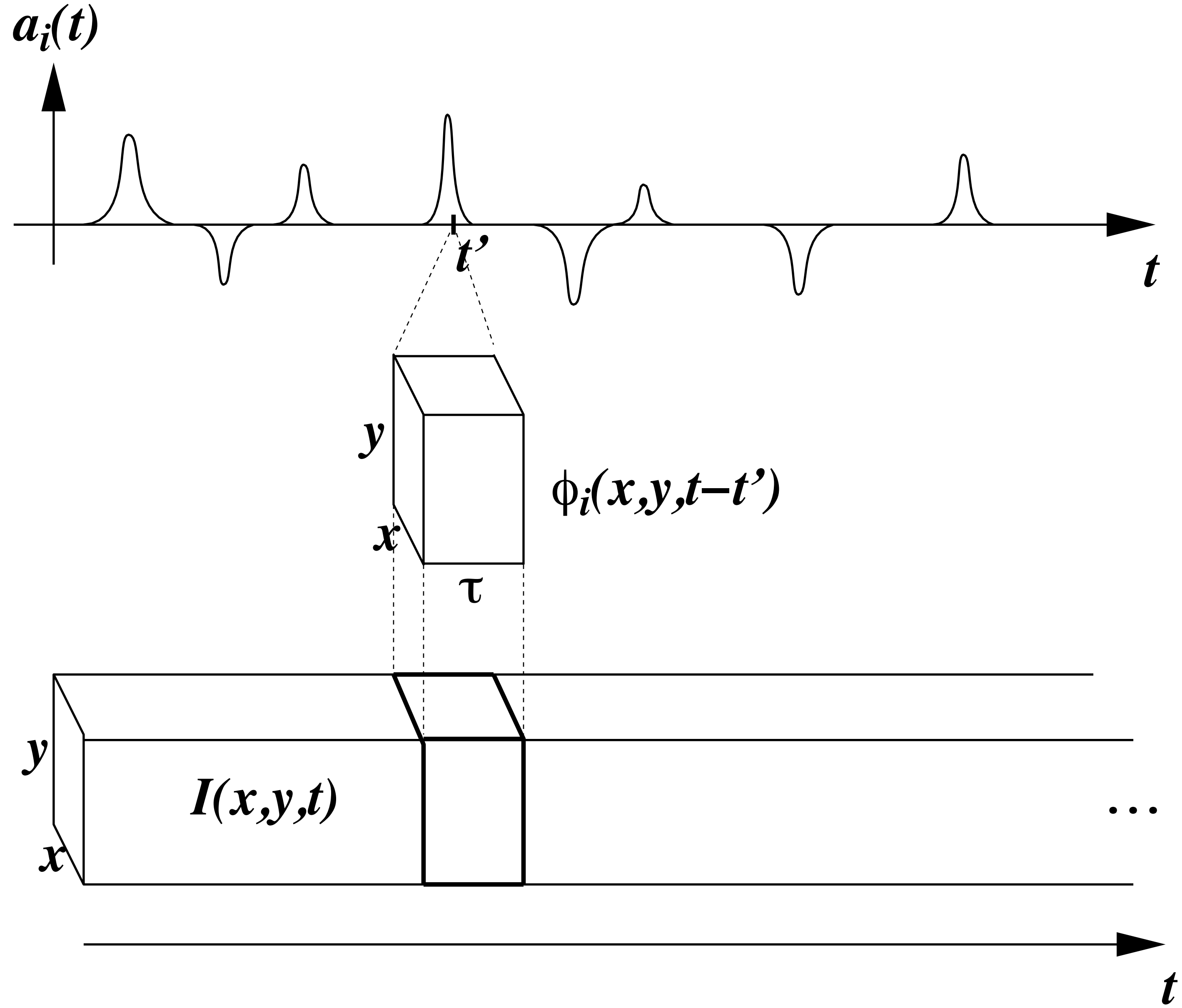


Responses in primary visual cortex to visual motion



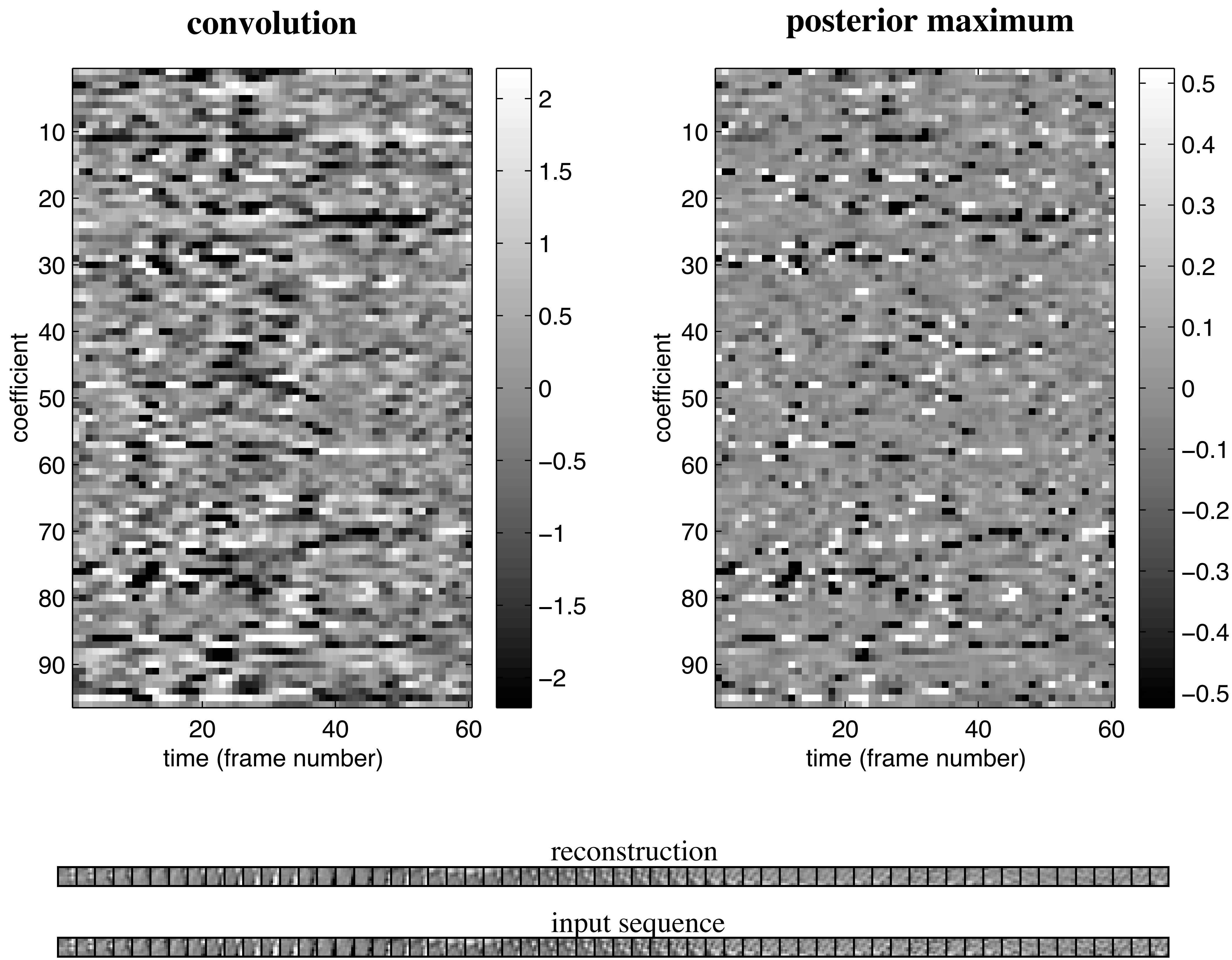
from Wandell, 1995

Sparse coding of time-varying images (Olshausen, 2002)



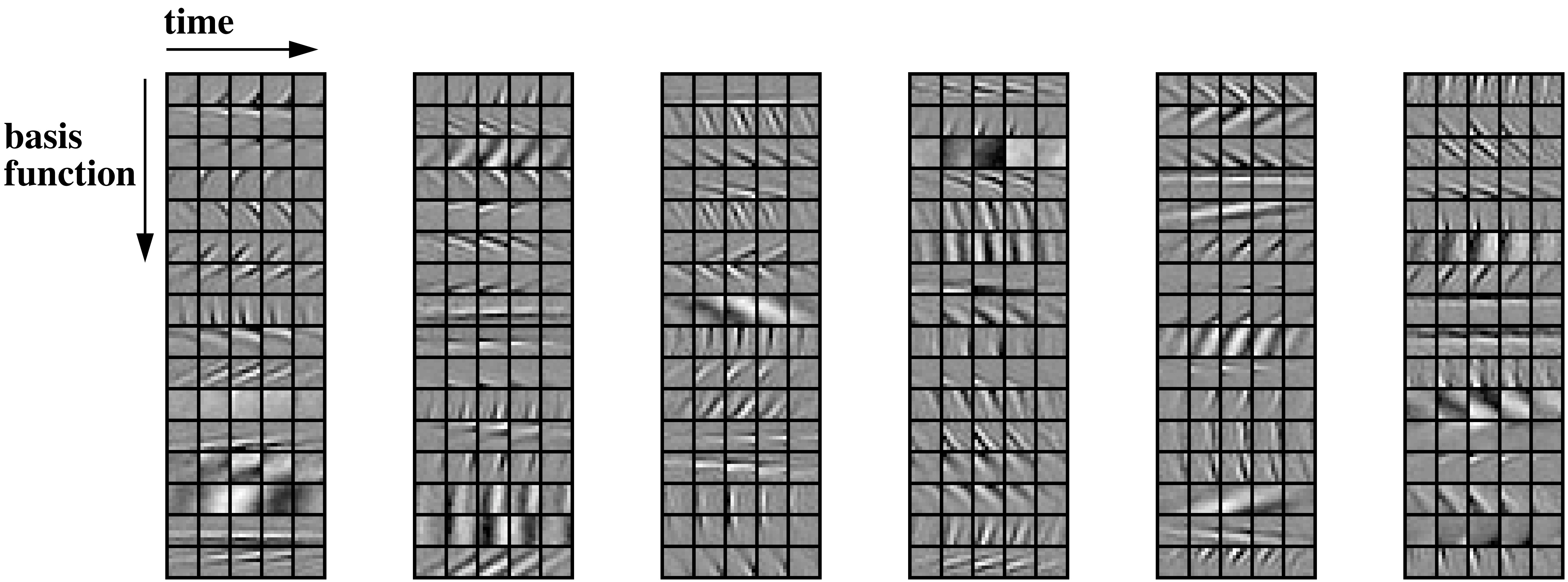
$$
\begin{aligned}
I(x,y,t) &= \sum_i \sum_{t'} a_i(t')\phi_i(x,y,t-t') + \epsilon(x,y,t) \\
&= \sum_i a_i(t) * \phi_i(x,y,t) + \epsilon(x,y,t)
\end{aligned}
$$

Sparse decomposition of image sequences



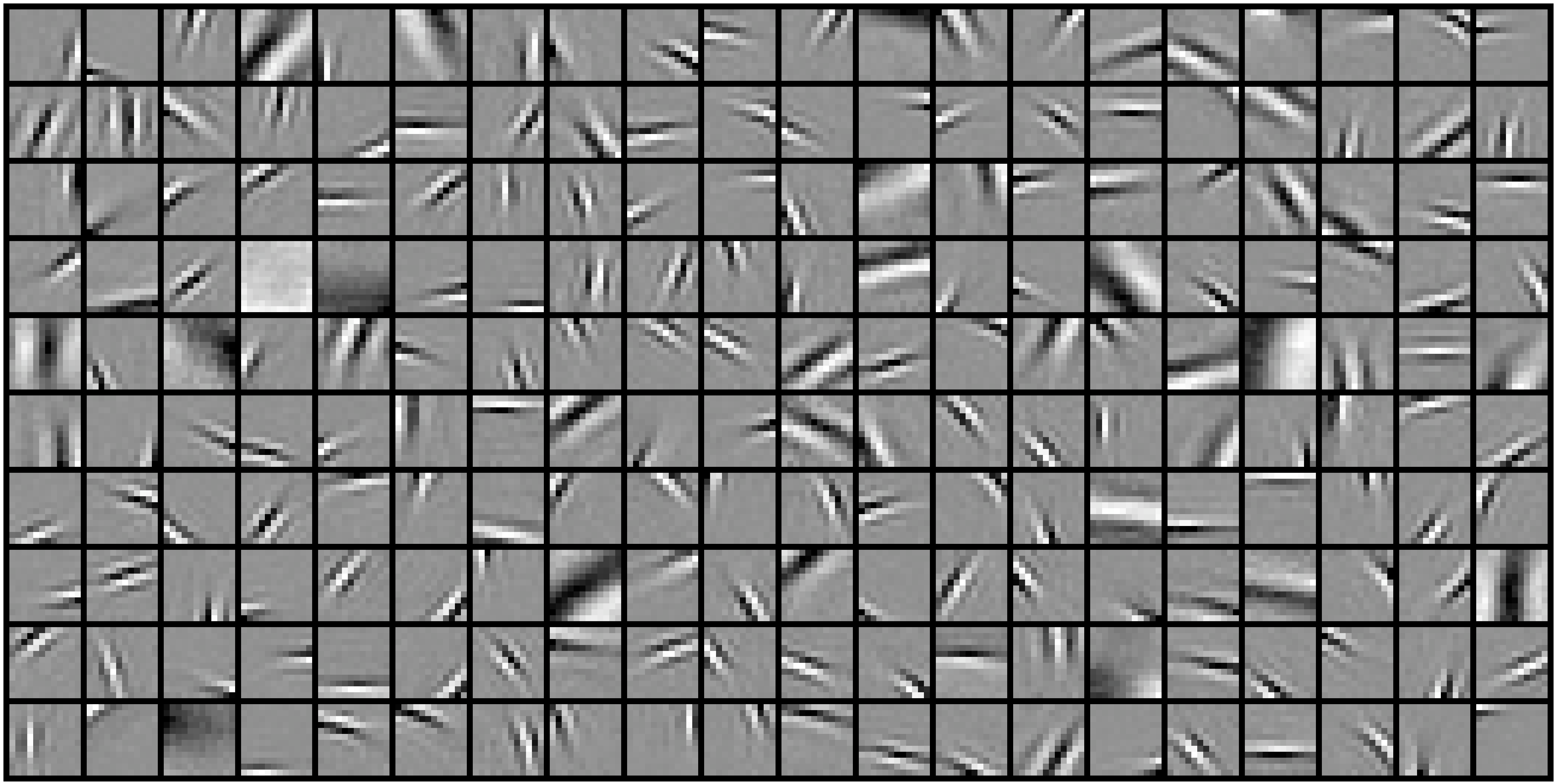
from Olshausen, 2002

Learned spatio-temporal basis functions



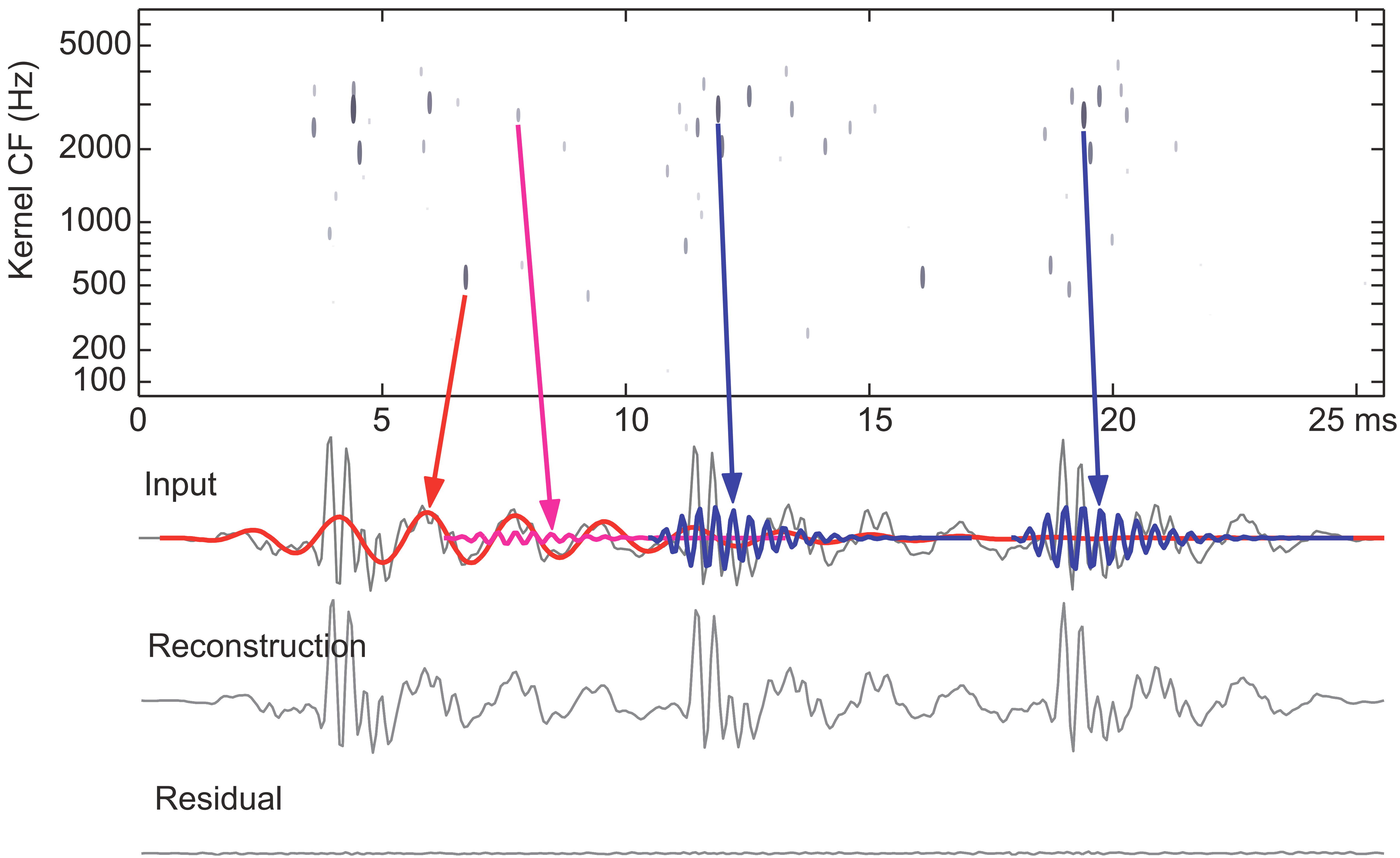
from Olshausen, 2002

Animated spatial-temporal basis functions

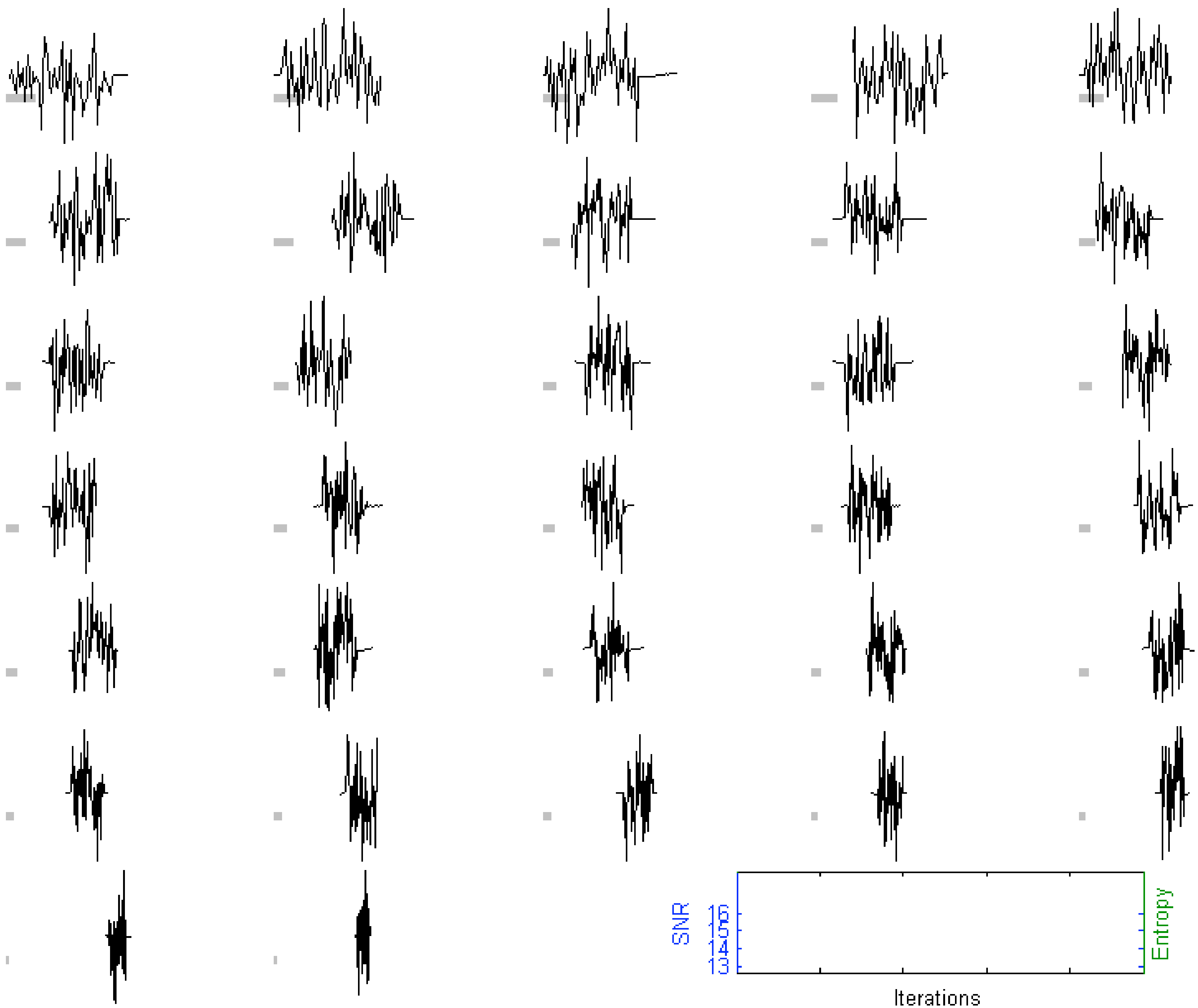


from Olshausen, 2002

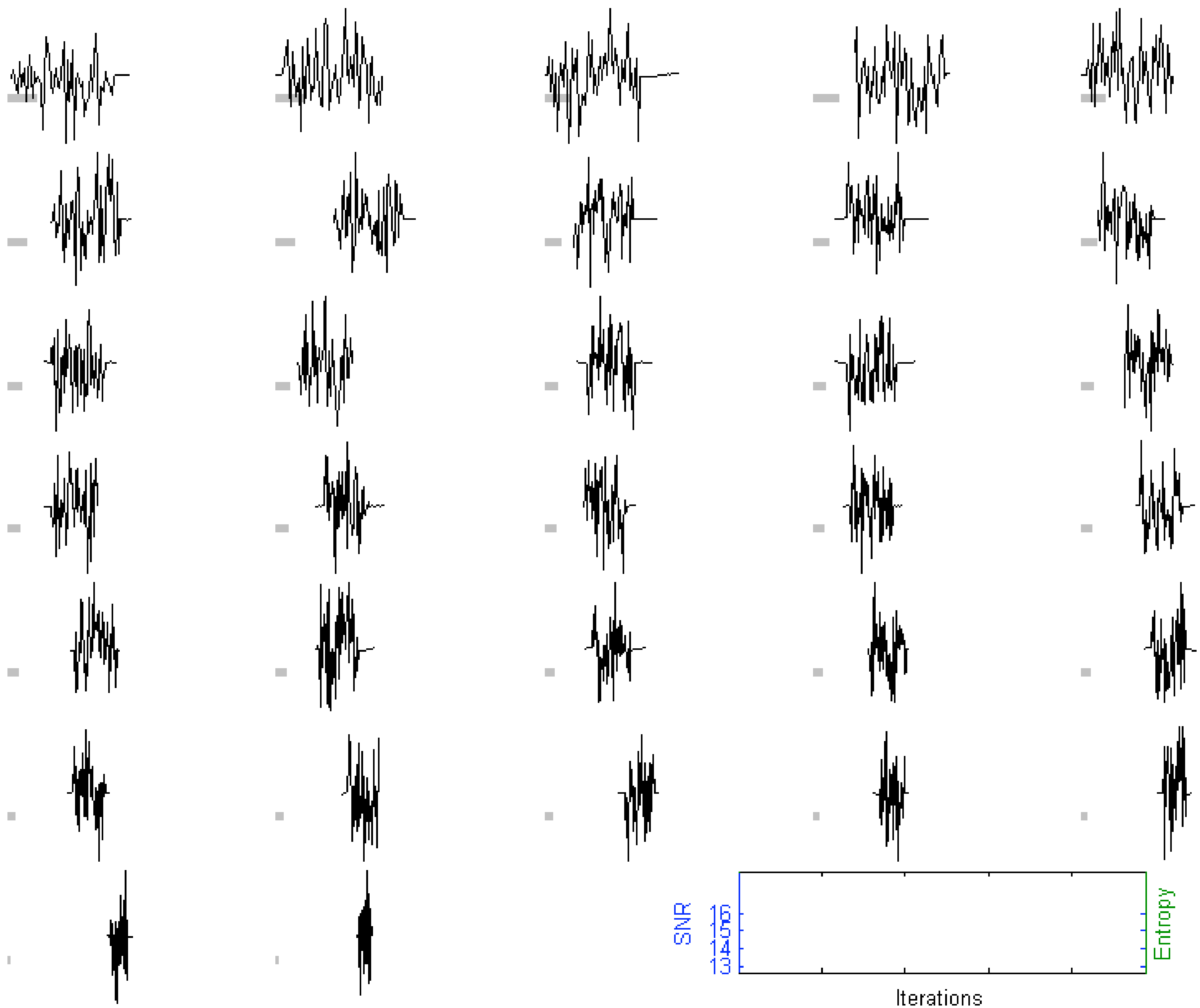
Coding audio signals with spikes



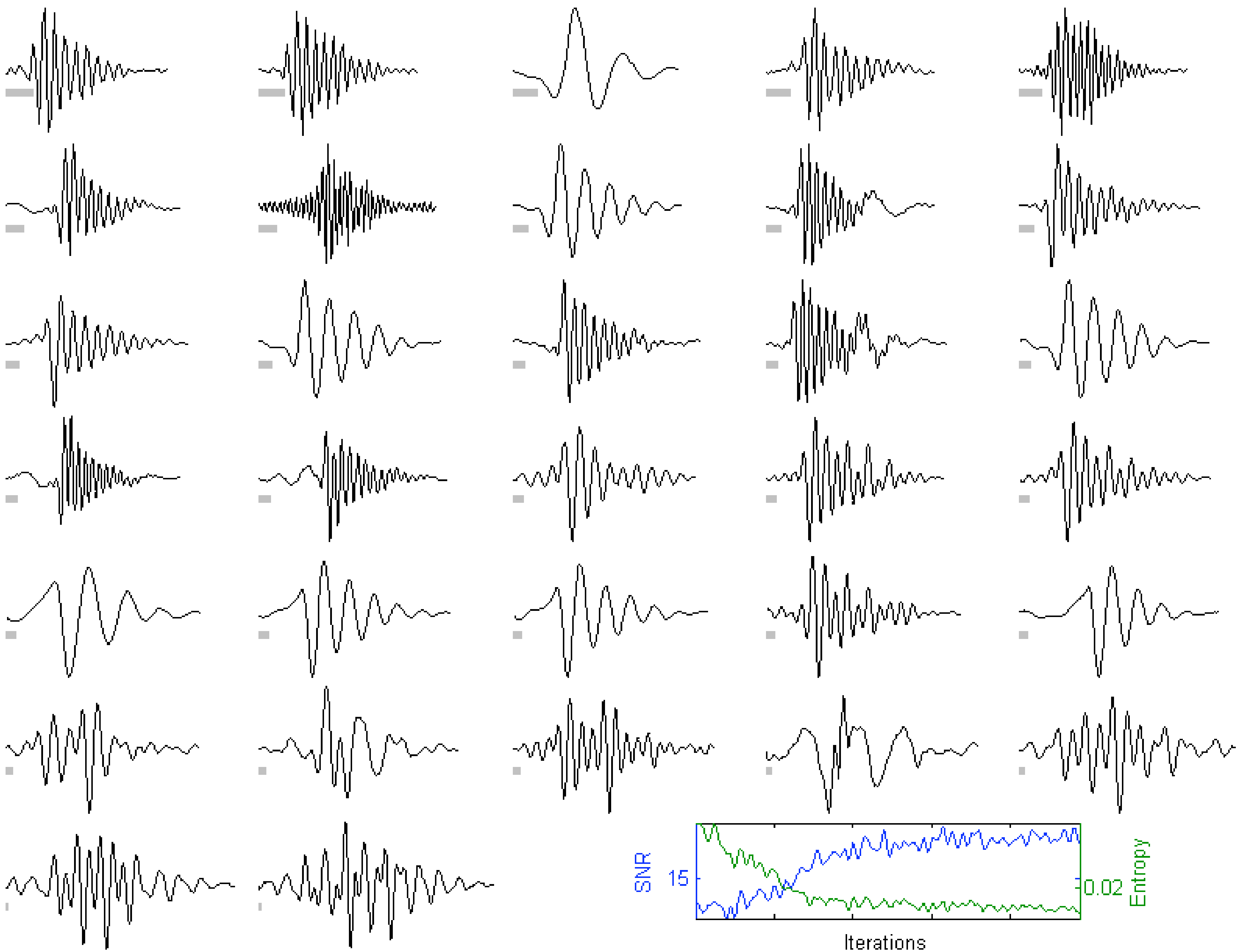
Kernel functions are initialized to random vectors



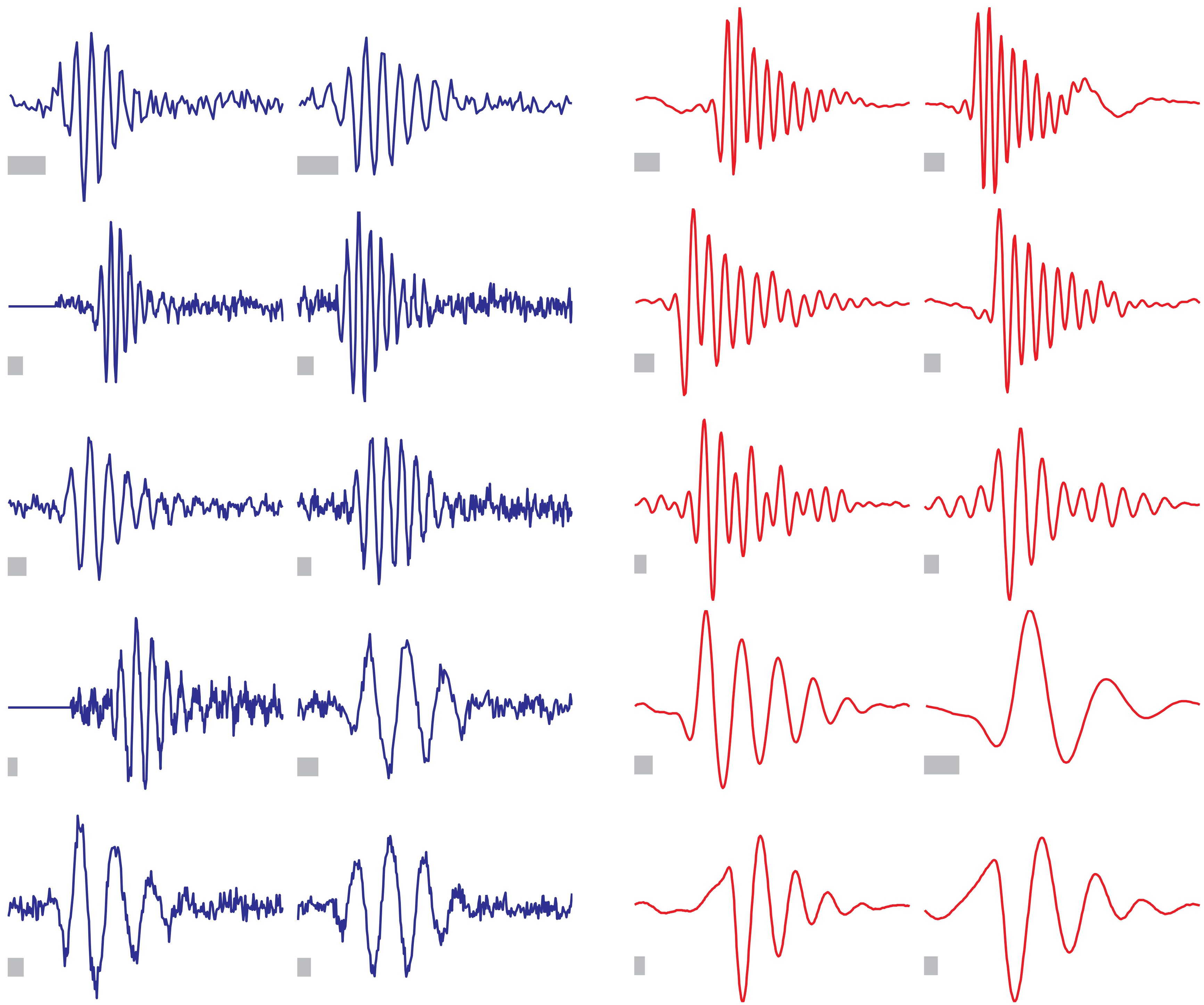
Adapting the optimal kernel shapes



Kernel functions optimized for coding speech



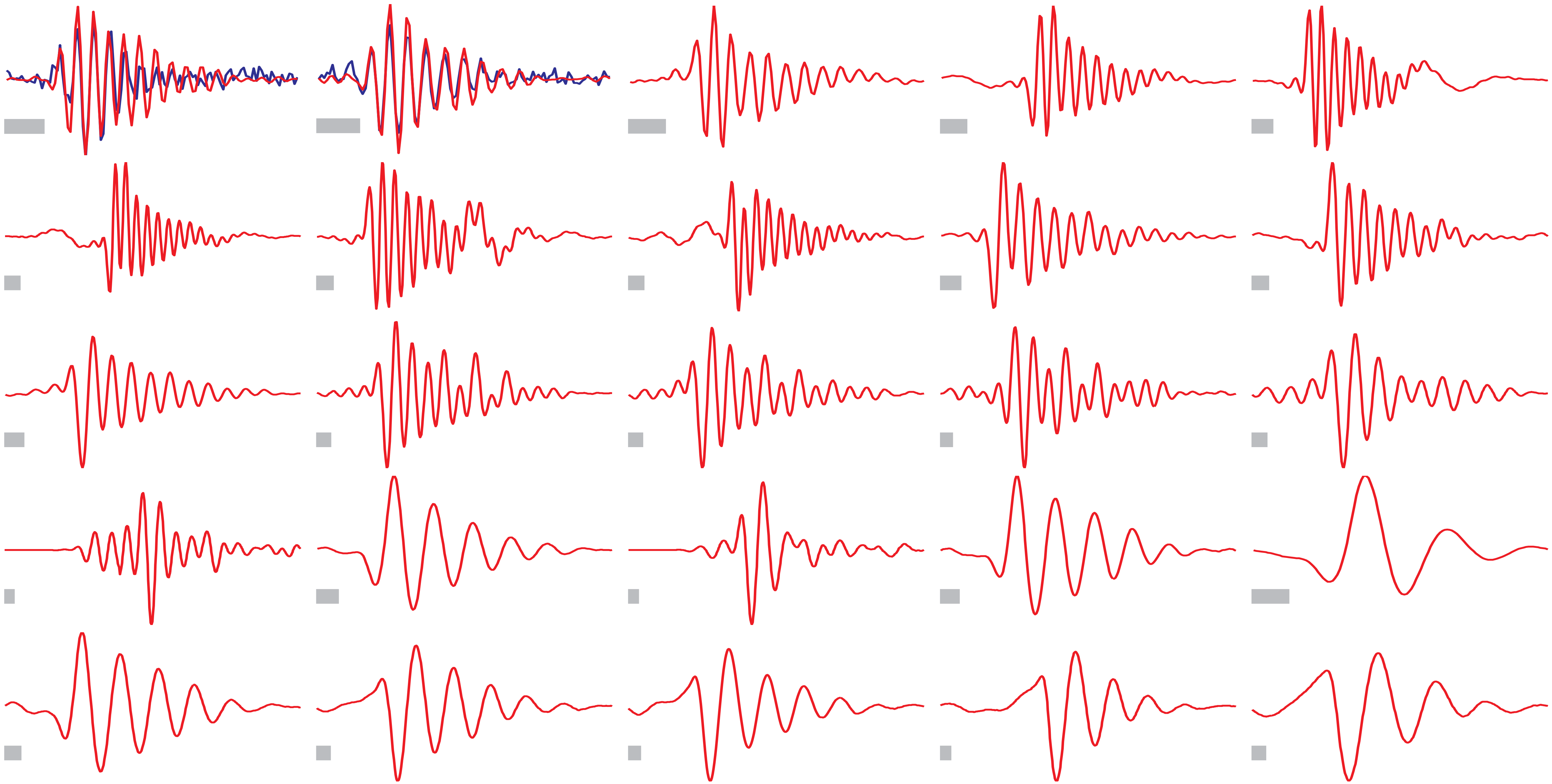
Learned kernels share features of auditory nerve filters



Auditory nerve filters
from Carney, McDuffy, and Shekhter, 1999

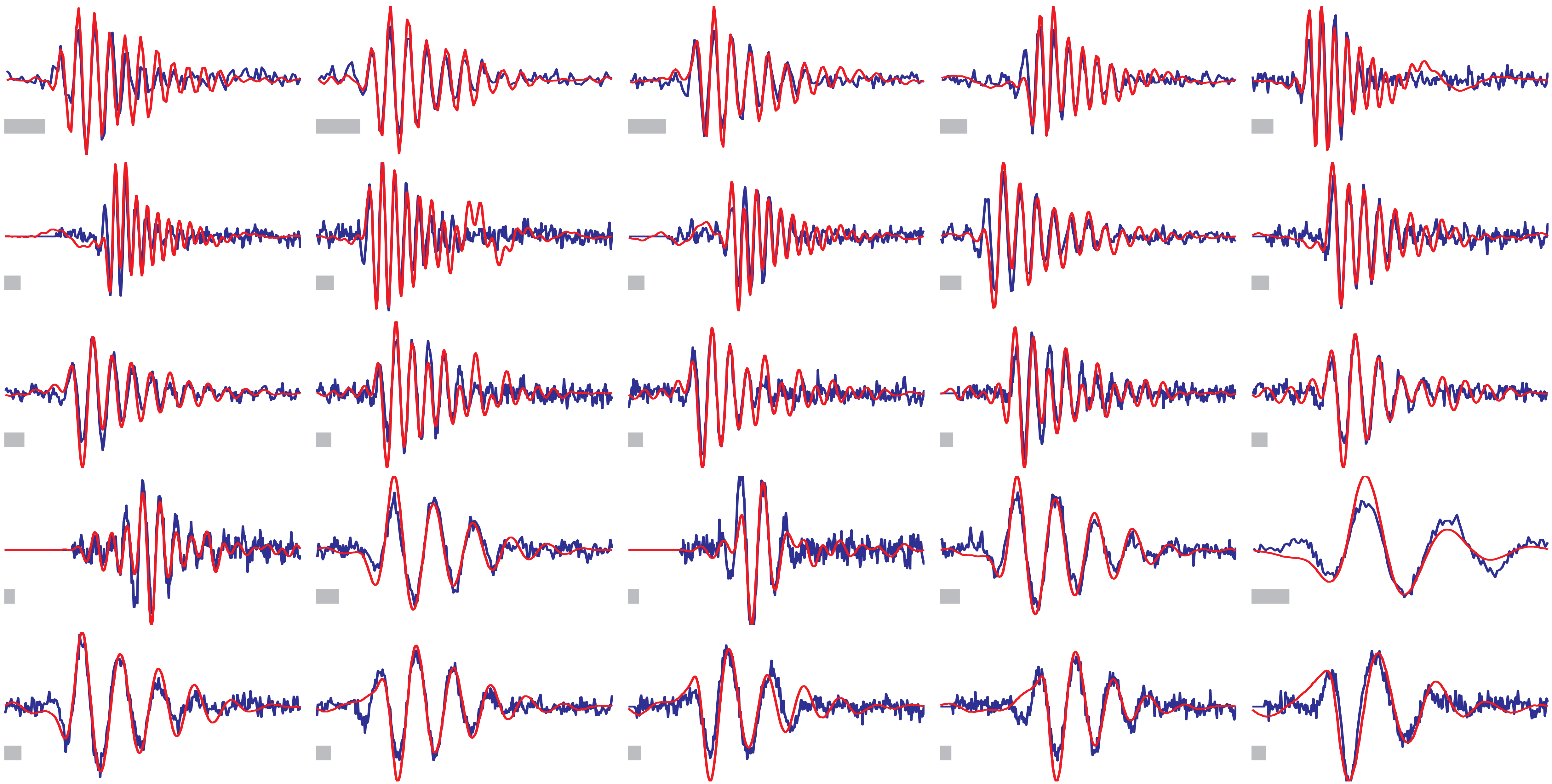
Optimized kernels
scale bar = 1 msec

Learned kernels closely match individual auditory nerve filters



for each kernel find closet matching auditory nerve filter
in Laurel Carney's database of ~100 filters.

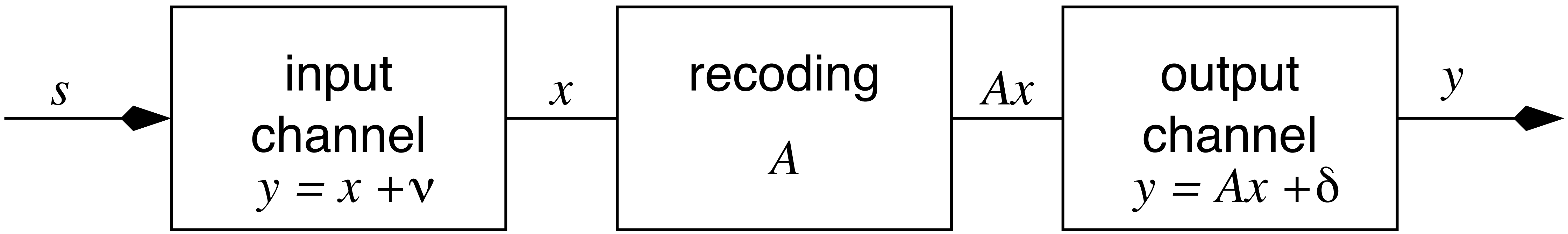
Learned kernels overlaid on selected auditory nerve filters



For almost all learned kernels there is a closely matching auditory nerve filter.

from Smith and Lewicki (2006)

Redundancy reduction for noisy channels (Atick, 1992)



Mutual information

$$I(x,s) = \sum_{s,x} P(x,s) \log_2 \left[\frac{P(x,s)}{P(s)P(x)} \right]$$

$I(x,s) = 0$ iff $P(x,s) = P(x)P(s)$, i.e. x and s are independent.

A second order statistical model: Gaussian

- To calculate I(x,s), we need P(s), P(x), and P(x,s)
- Assume it is sufficient to measure 2nd order correlations

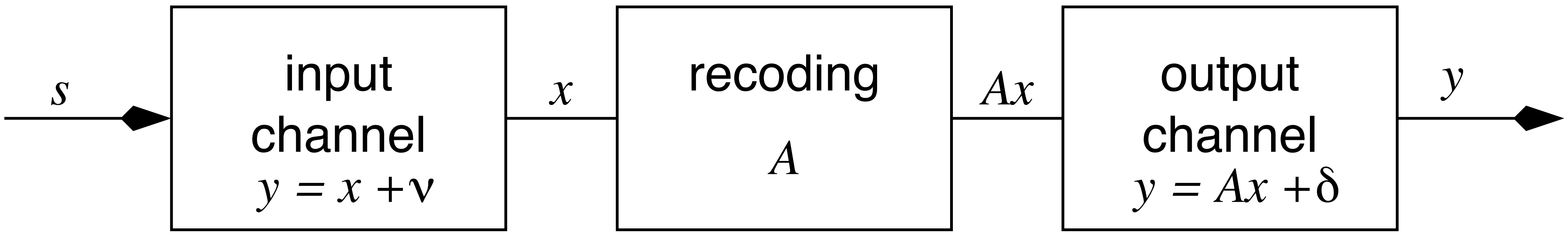
  (this is equivalent to measuring the avg. spatial frequency):

$$
\begin{aligned}
\langle s[n]s[m] \rangle &= R_0[n,m] \\
\langle x[n]x[m] \rangle &= R_0[n,m] + N^2\delta_{n,m} \equiv R[n,m] \\
\langle x[n]s[m] \rangle &= \langle s[n]s[m] \rangle
\end{aligned}
$$

for $u = s, x, y$ and $R_{uu}[n,m] \equiv \langle u[n]u[m] \rangle$.

$$
\begin{aligned}
P(u) &= \left[(2\pi)^d \det(R_{uu})\right]^{-1/2} \exp\left[-\frac{1}{2}\sum_{n,m}(u[n]-\bar{u})R_{uu}^{-1}[n,m](u[m]-\bar{u})\right] \\
&= \mathcal{N}(\bar{u}, R_{uu})
\end{aligned}
$$

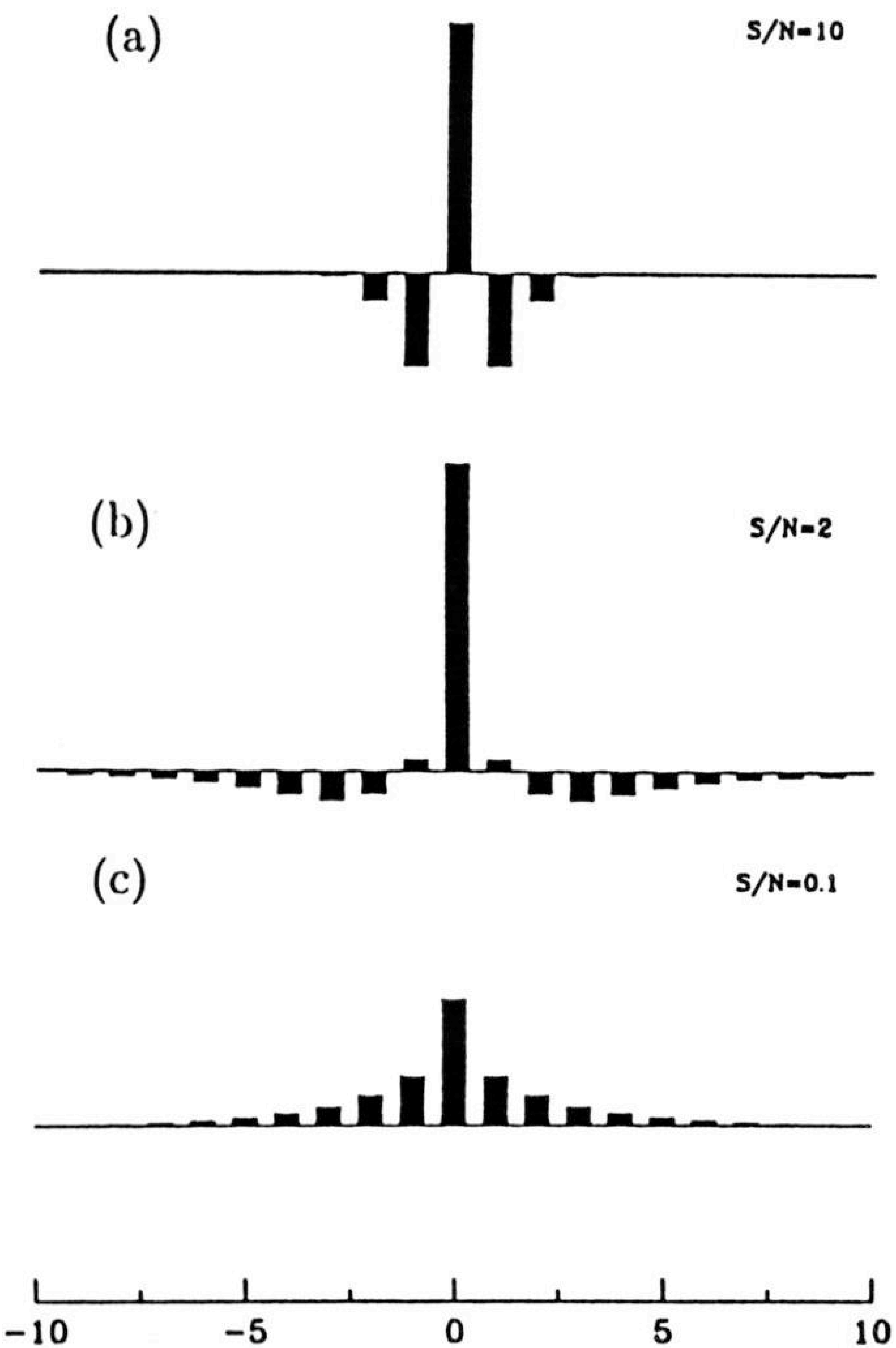
Mutual information between stages of the model



$$
\begin{aligned}
I(x,s) &= \frac{1}{2}\left[\frac{\det(R_0+N^2)}{\det N^2}\right] \\
I(y,s) &= \frac{1}{2}\left[\frac{\det(A(R_0+N^2)A^T+N_\delta^2)}{\det(AN^2A^T+N_\delta^2)}\right]
\end{aligned}
$$

- Mutual information depends on noise and correlations
- Increasing noise ⇒ decreasing mutual information
- Can derive the optimal A for different signal to noise ratios.

1D profile of optimal filters

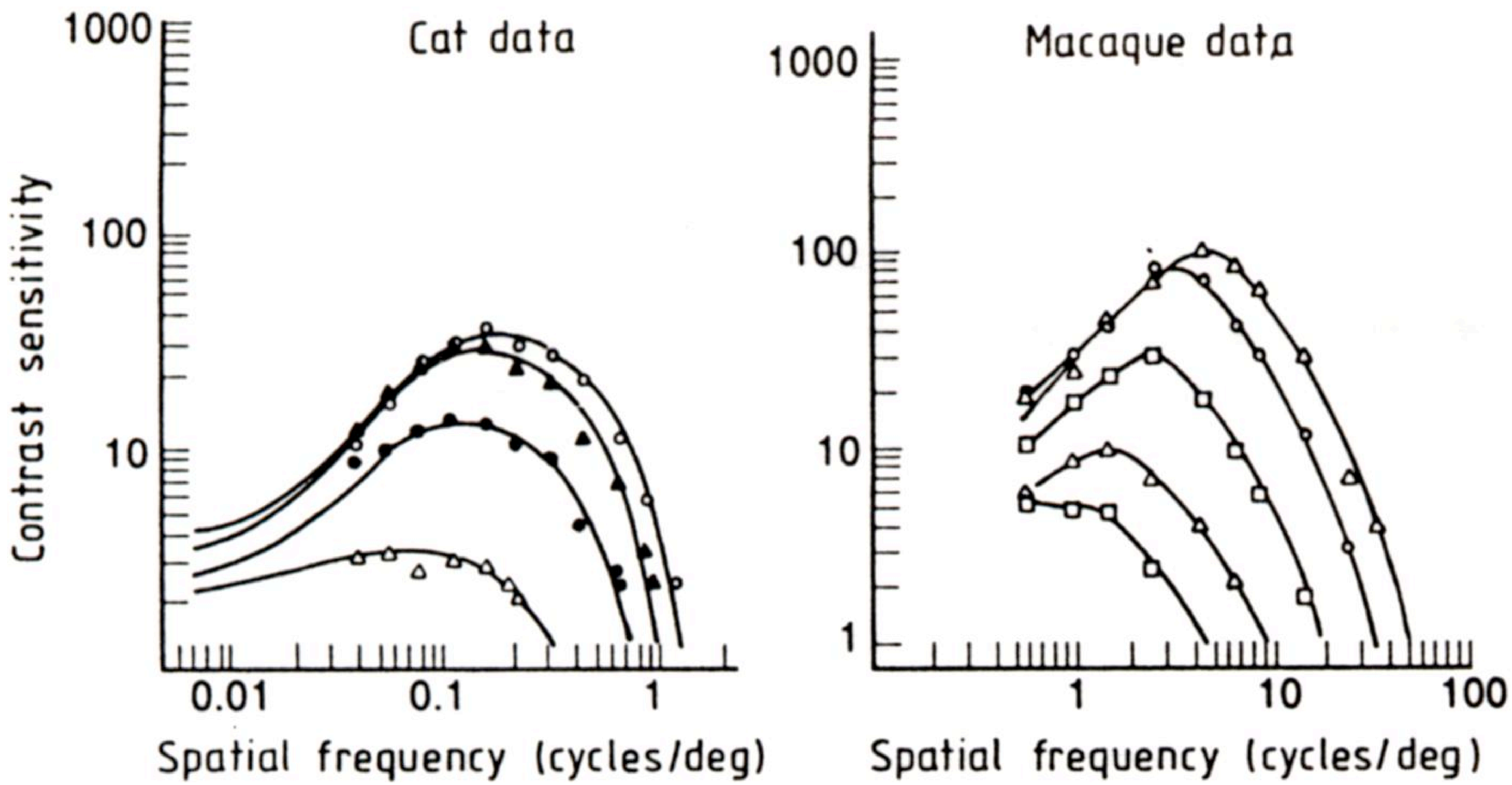


- high SNR
  - reduce redundancy
  - center-surround structure

- low SNR
  - average
  - low-pass filter

- matches behavior of retinal ganglion cells

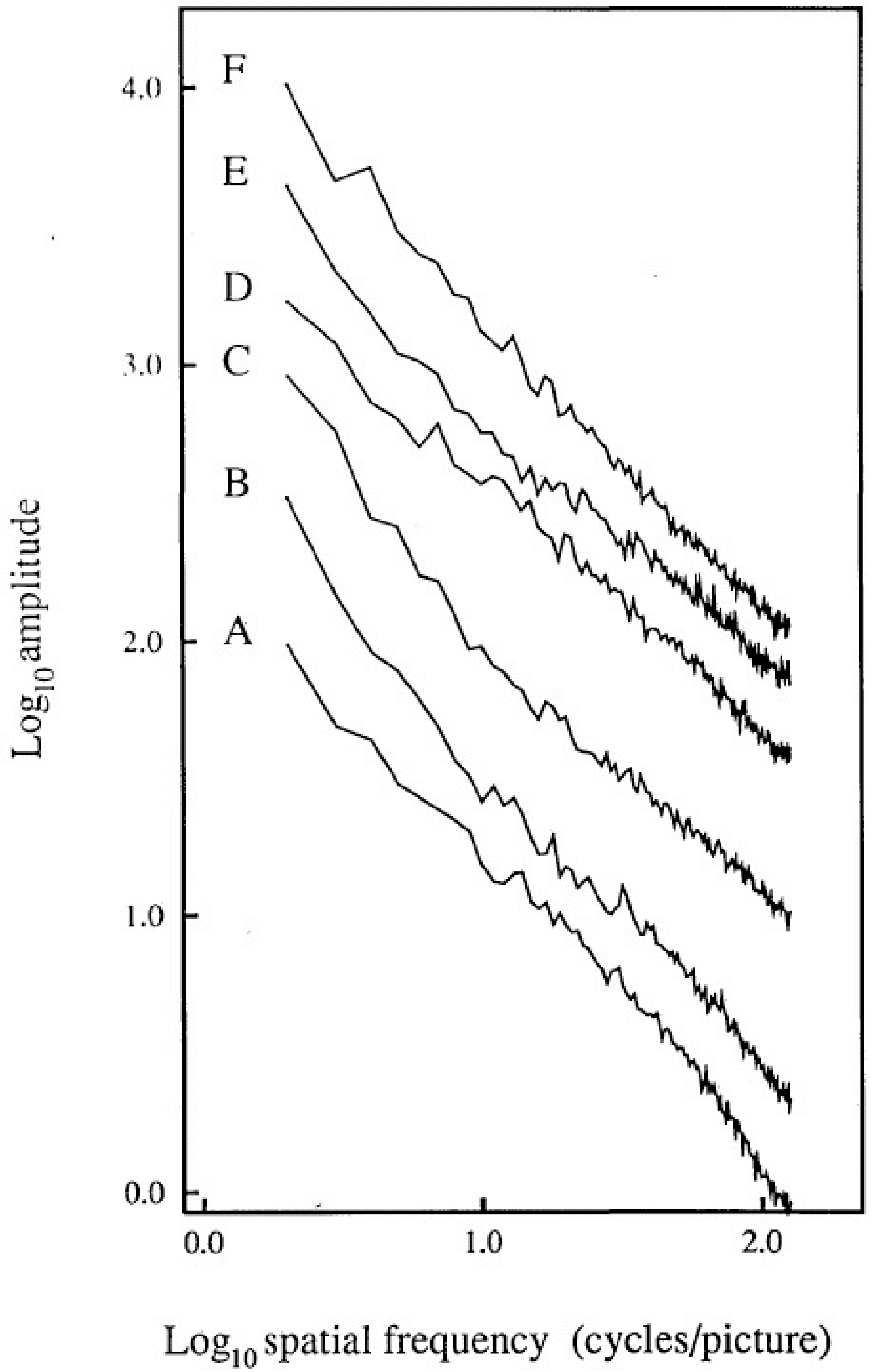
An observation: Contrast sensitivity of ganglion cells



Luminance level decreases one log unit each time we go to lower curve.

What is happening at low luminance levels?

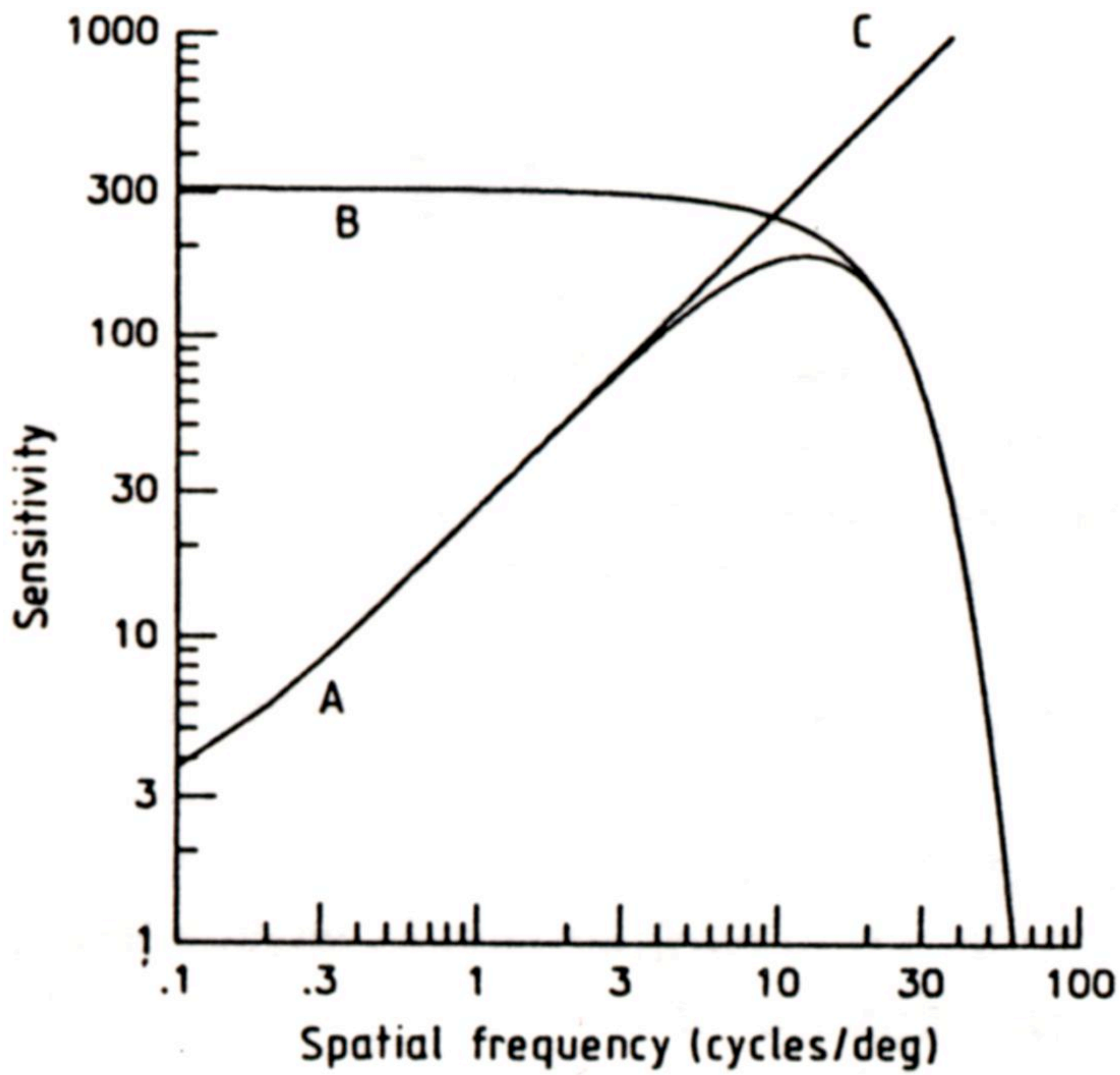
Natural images have a $1/f^2$ power spectrum



- Field 1987
  - amplitude spectra for 6 images (shifted for clarity)
  - power spectra fall off as $1/f^2$
  - Fourier coefficients fall off as $1/f$

- *How does this reflect the correlated structure of natural images?*
- *What would an uncorrelated structure look like?*
- *What transformation would yield a more efficient code?*

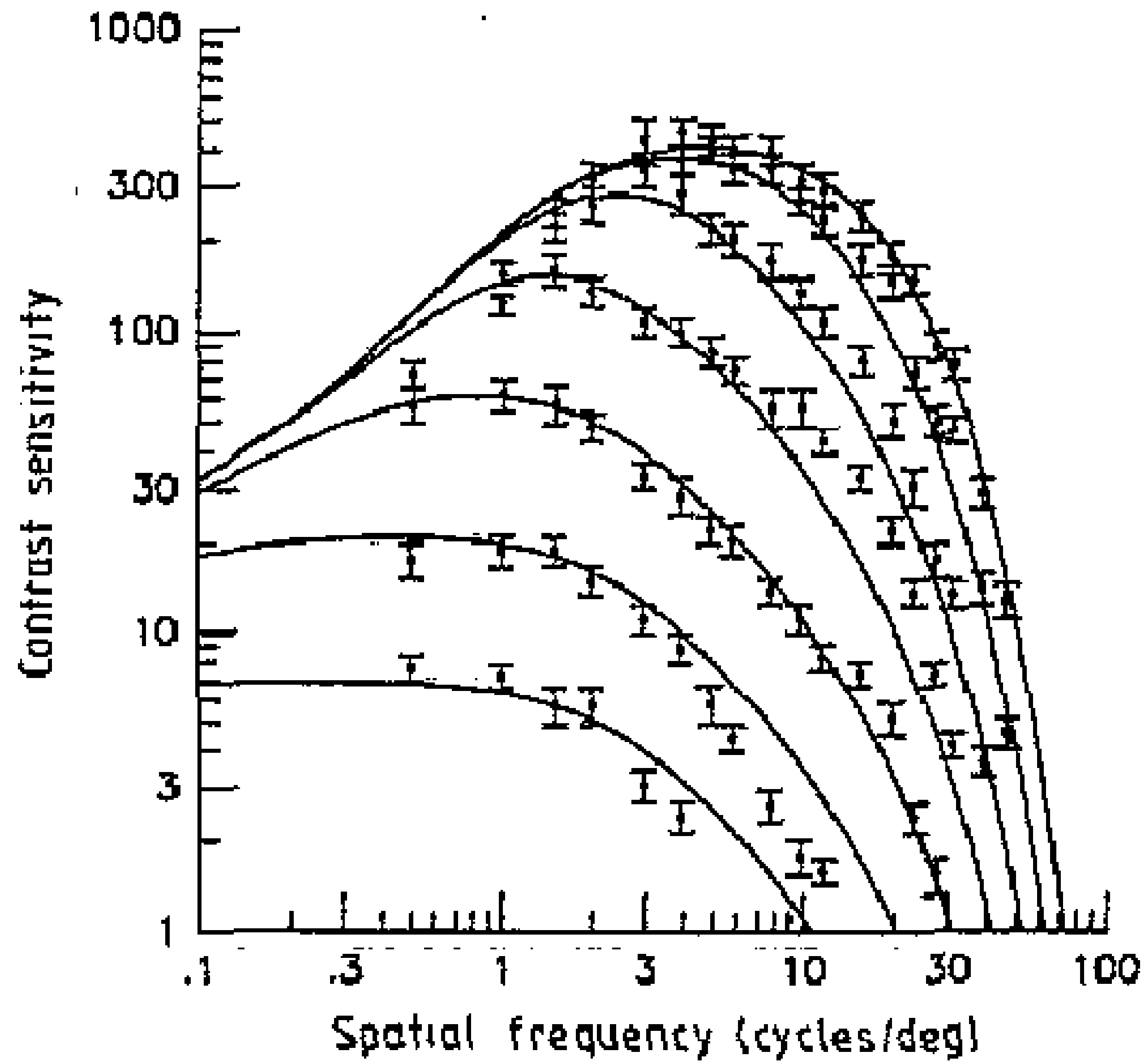
Components of predicted filters



from Atick, 1992

The predicted form of the optimal filter (A), is a combination of a low-pass filter (B) plus a whitening filter (C).

Predicted contrast sensitivity functions match neural data



from Atick, 1992

Part 3