# Higher Order Statistics

Matthias Hennig

School of Informatics, University of Edinburgh

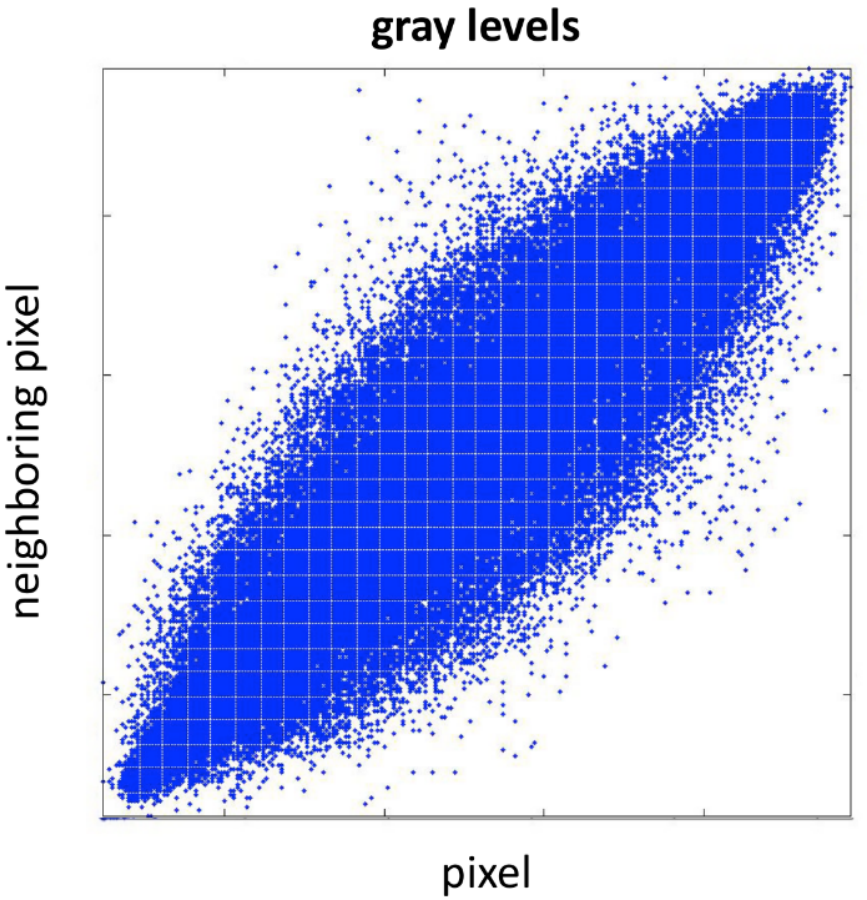
March 1, 2019

[0]Acknowledgements: Mark van Rossum and Chris Williams.

# Outline

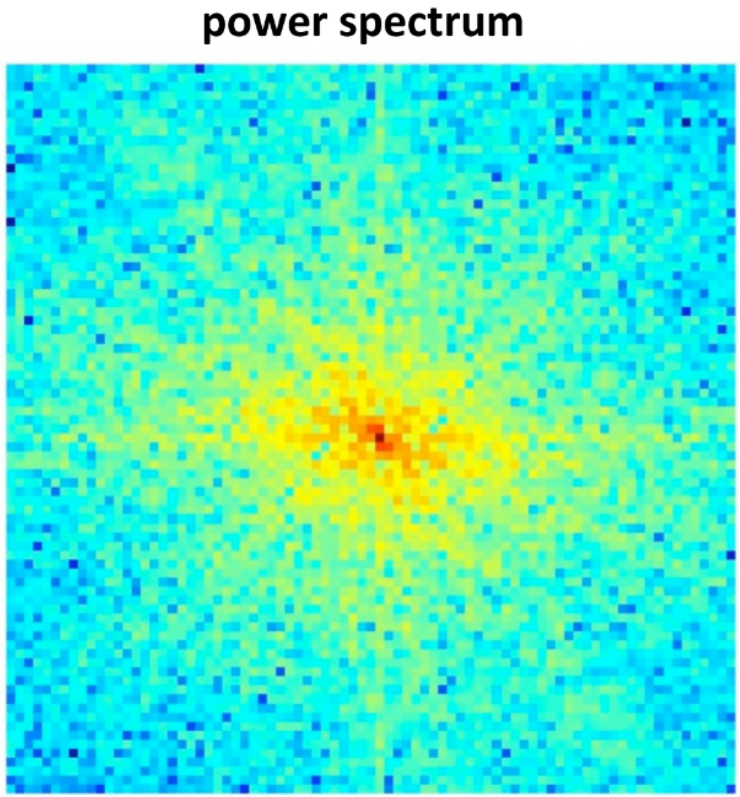
- First, second and higher-order statistics
- Generative models, recognition models
- Sparse Coding
- Independent Components Analysis

# Sensory information is highly redundant




[Figure: Matthias Bethge]

# and higher order correlations are relevant




[Figure: Matthias Bethge]

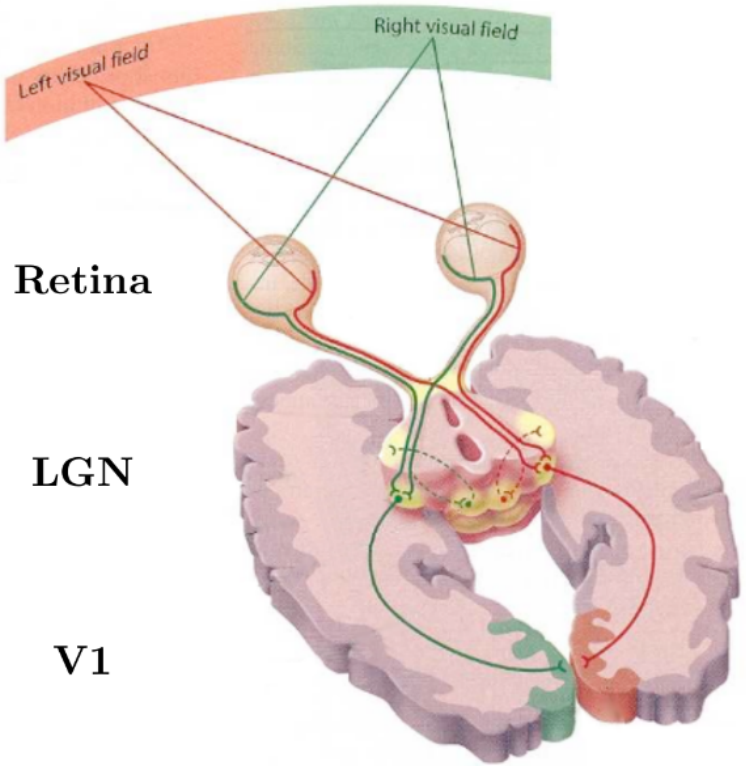
note Fourier transform of the autocorrelation function is equal to the power spectral density (Wiener-Khinchin theorem)

# Redundancy Reduction

(Barlow, 1961; Attneave 1954)

- Natural images are redundant in that there exist statistical dependencies amongst pixel values in space and time
- In order to make efficient use of resources, the visual system should *reduce* redundancy by removing statistical dependencies

# The visual system

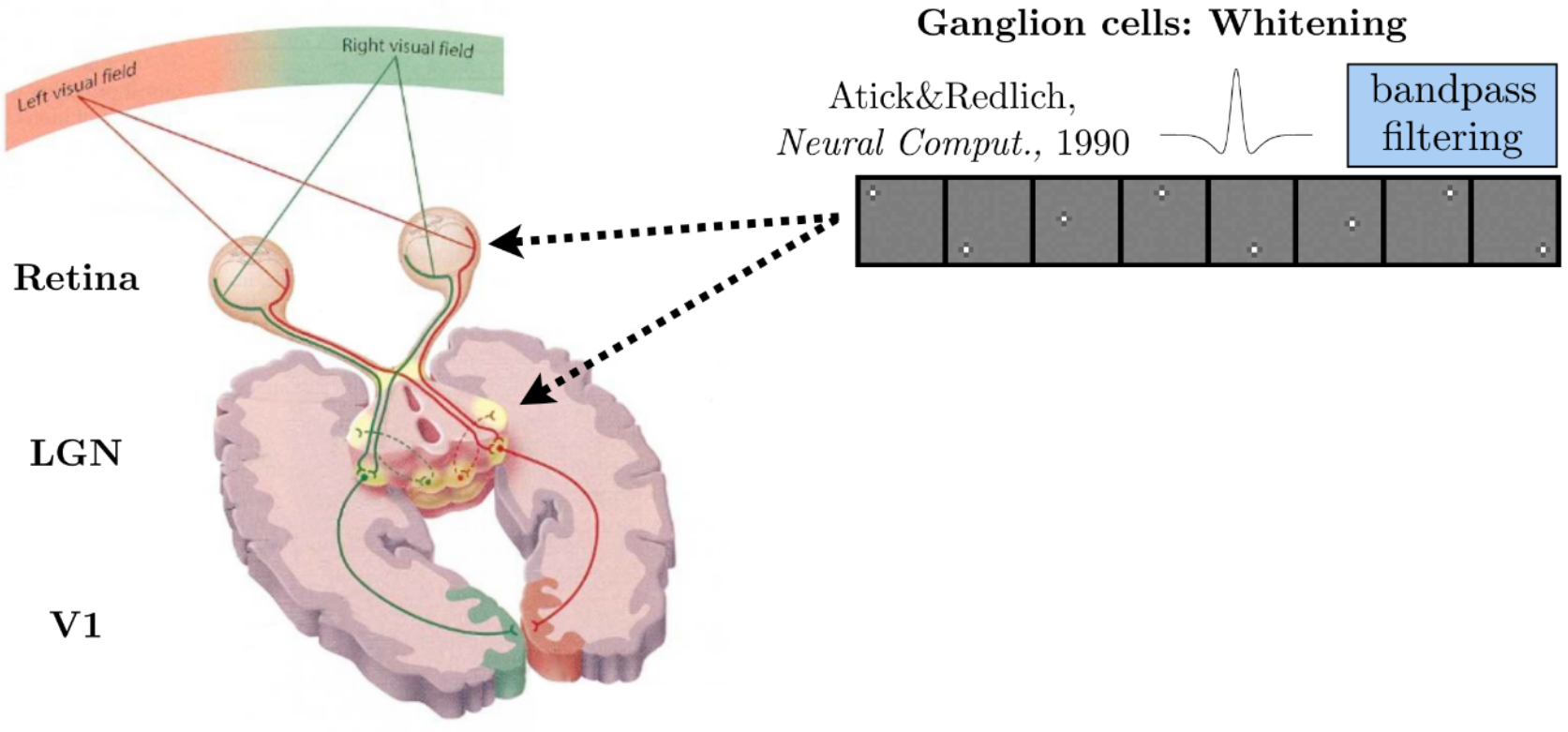



An attractive feature of [redundancy reduction] is that a code formed in response to redundancies in the input would constitute a distributed memory of this regularities---one that is used automatically and does not require a separate recall mechanism."

Horace Barlow, *BBS*, 2001

[Figure from Matthias Bethge]

# The visual system

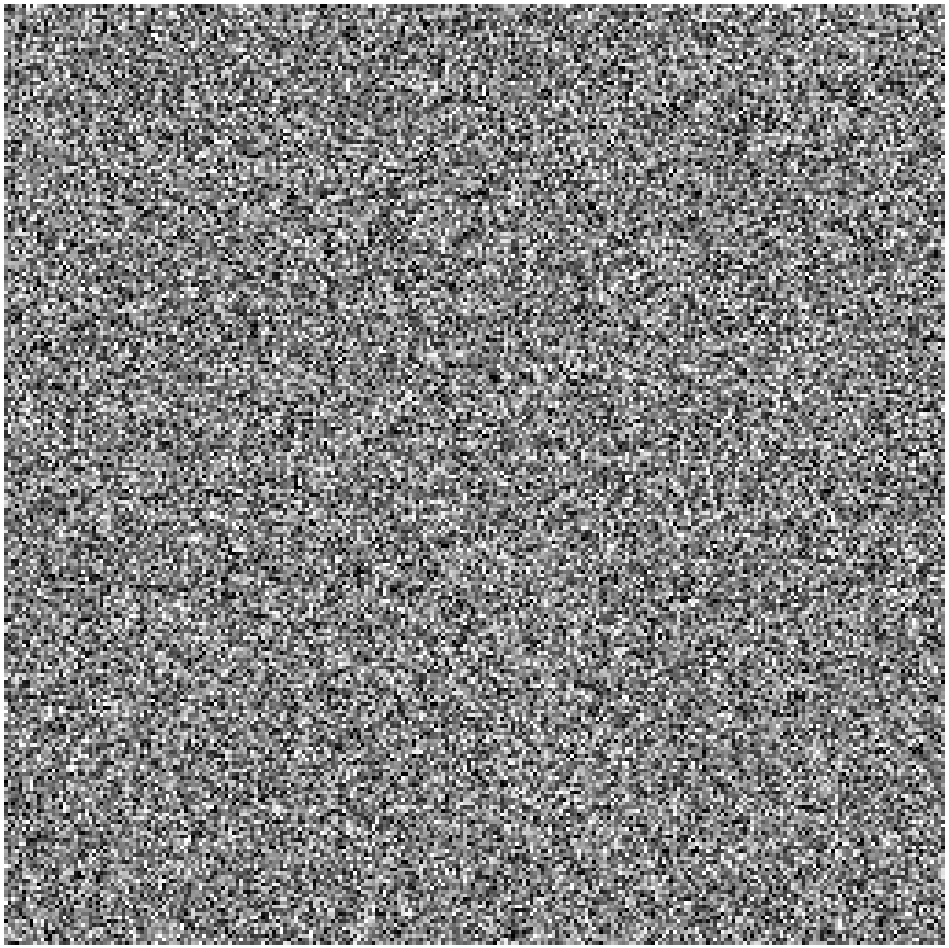



[Figure from Matthias Bethge]

# Natural Image Statistics and Efficient Coding

- First-order statistics
  - Intensity/contrast histograms $\Rightarrow$ e.g. histogram equalization
- Second-order statistics
  - Autocorrelation function ($1/f^2$ power spectrum)
  - Decorrelation/whitening
- Higher-order statistics
  - orientation, phase spectrum (systematically model higher orders)
  - Projection pursuit, sparse coding (find useful projections)

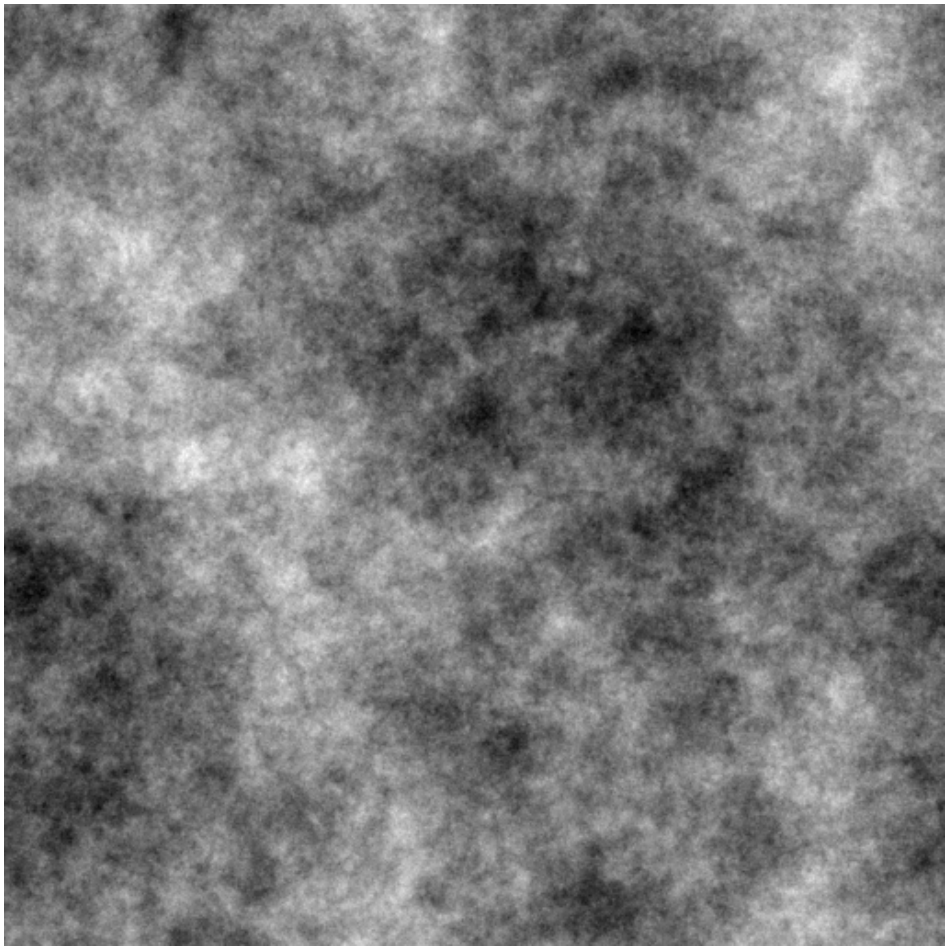
# Image synthesis: First-order statistics



[Figure: Olshausen, 2005]

Log-normal distribution of intensities.

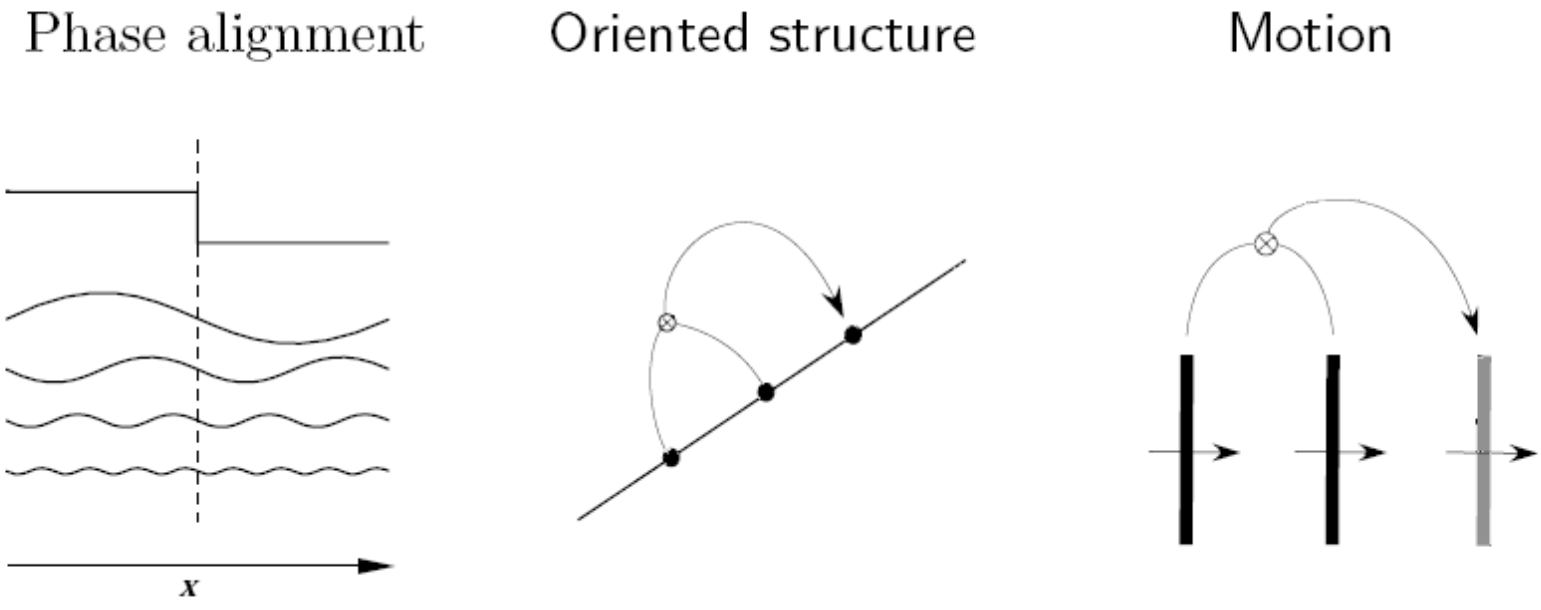
# Image synthesis: Second-order statistics



[Figure: Olshausen, 2005]

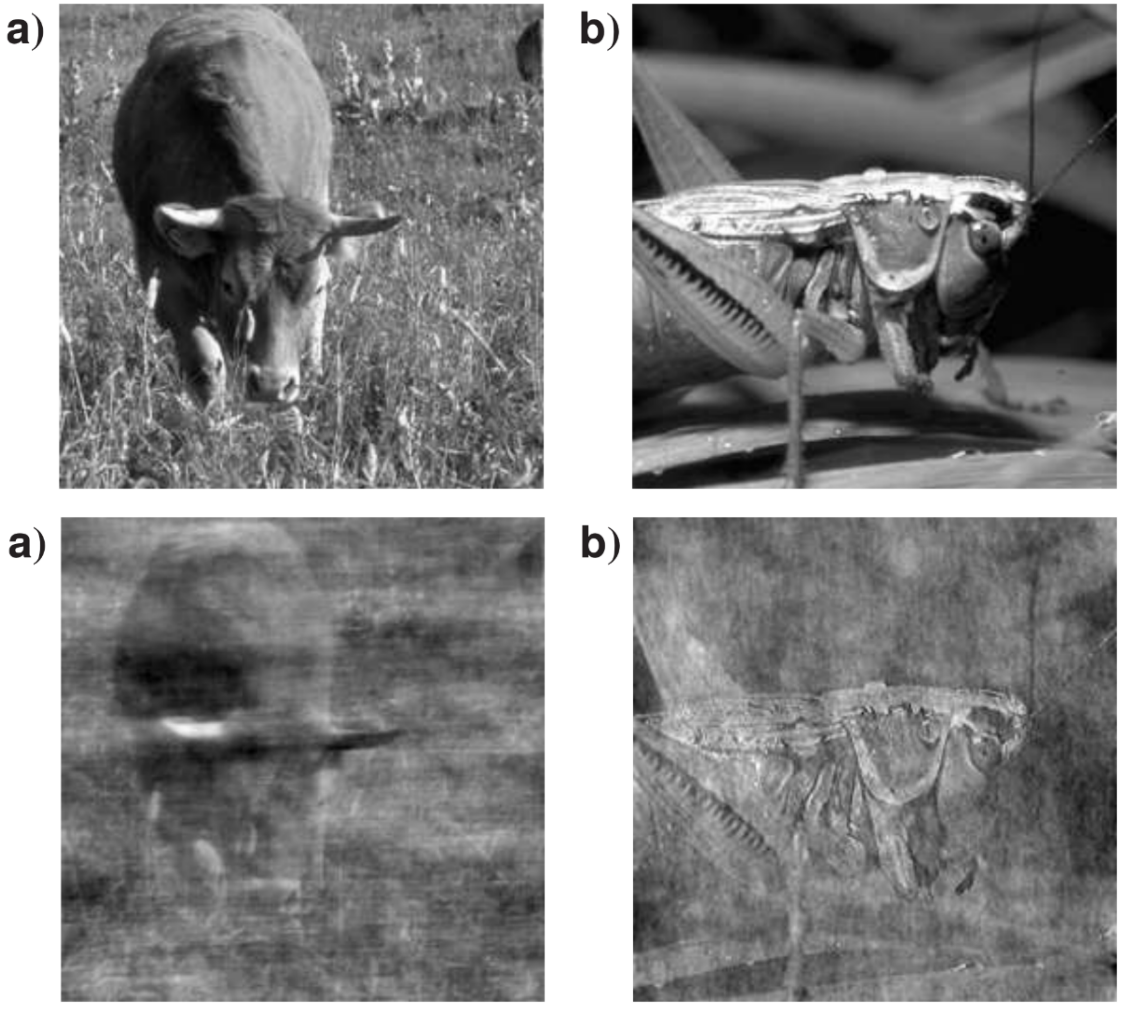
Describe as correlated Gaussian statistics, or equivalently, power spectrum.

# Higher-order statistics




[Figure: Olshausen, 2005]

# Importance of phase information




[Hyvärinen et al., 2009]

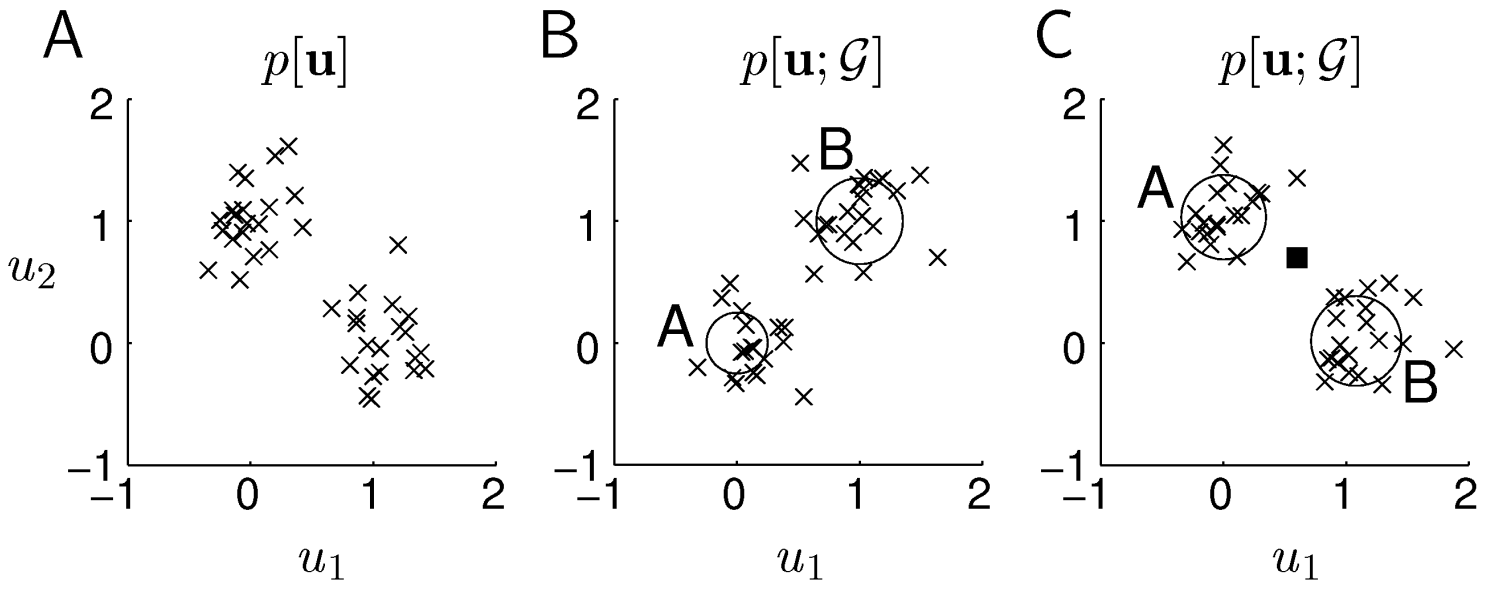
# Generative models, recognition models

(§10.1, Dayan and Abbott)

- How is sensory information encoded to support higher level tasks?
- Has to be based on the statistical structure of sensory information.
- Causal models: find the *causes* that give rise to observed stimuli.
- Generative models: reconstruct stimuli based on causes, model can fill in based on statistics.
- Allows the brain to generate appropriate actions (motor outputs) based on causes.
- A stronger constraint than optimal encoding alone (although it should still be optimal).

# Generative models, recognition models

(§10.1, Dayan and Abbott)




Left: observations. Middle: poor model; 2 latent causes (prior distribution) but wrong generating distribution given causes. Right: good model.

In image processing context one would want, e.g. A are cars, B are faces. They would explain the image, and could generate images with an appropriate generating distribution.

# Generative models, recognition models

- Hidden (latent) variables **h** (causes) that explain
- visible variables **u** (e.g. image)
- Generative model

$$p(\mathbf{u}|\mathcal{G}) = \sum_{\mathbf{h}} p(\mathbf{u}|\mathbf{h},\mathcal{G})p(\mathbf{h}|\mathcal{G})$$

- Recognition model

$$p(\mathbf{h}|\mathbf{u},\mathcal{G}) = \frac{p(\mathbf{u}|\mathbf{h},\mathcal{G})p(\mathbf{h}|\mathcal{G})}{p(\mathbf{u}|\mathcal{G})}$$

- Matching $p(\mathbf{u}|\mathcal{G})$ to the actual density $p(\mathbf{u})$. Maximize the log likelihood $L(\mathcal{G}) = \langle \log p(\mathbf{u}|\mathcal{G}) \rangle_{p(\mathbf{u})}$
- Train parameters $\mathcal{G}$ of the model using EM (expectation-maximization)

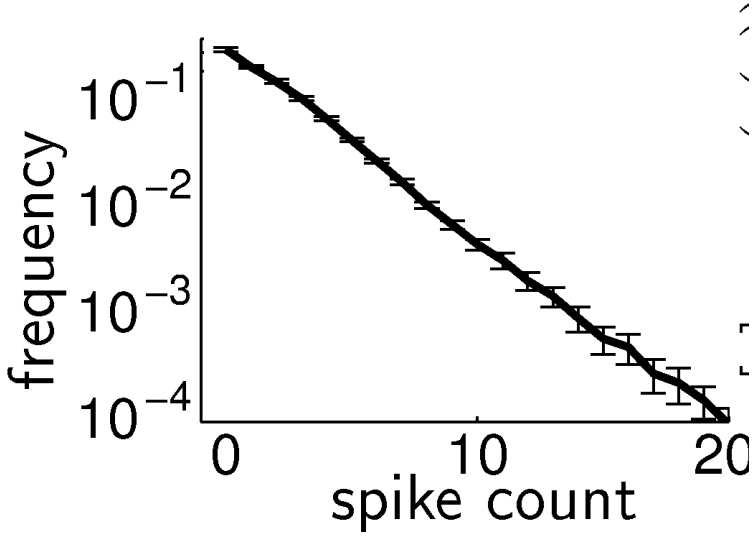
# Examples of generative models

(§10.1, Dayan and Abbott)

- Mixtures of Gaussians
- Factor analysis, PCA
- Sparse Coding
- Independent Components Analysis

# Sparse Coding

- Area V1 is highly overcomplete. V1 : LGN $\approx$ 25:1 (in cat)
- Firing rate distribution is typically exponential (i.e. sparse)
- Experimental evidence for sparse coding in insects, zebra finch, mouse, rabbit, rat, macaque monkey, human
  [Olshausen and Field, 2004]




Activity of a macaque IT cell in response to video images [Figure: Dayan and Abbott, 2001]

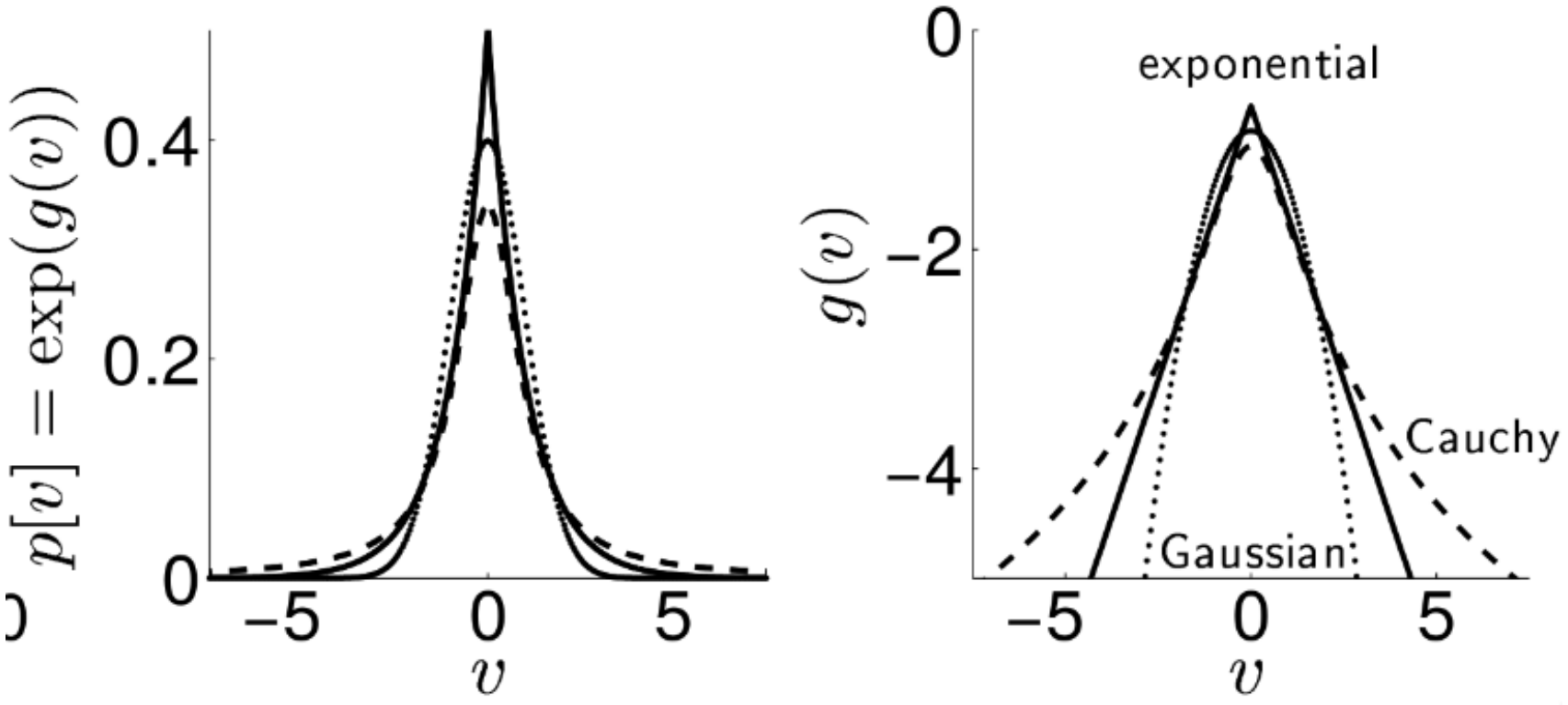
# Sparse Coding

- Distributions that are close to zero most of the time but occasionally far from 0 are called *sparse*
- Sparse distributions are more likely than Gaussians to generate values near to zero, and also far from zero (heavy tailed)

$$\text{kurtosis} = \frac{\int p(x)(x-\overline{x})^4 dx}{\left(\left[\int p(x)(x-\overline{x})^2 dx\right]^2} - 3$$

- Gaussian has kurtosis 0, positive $k$ implies sparse distributions (super-Gaussian, leptokurtotic)
- Kurtosis is sensitive to outliers (i.e. it is not robust). See HHH §6.2 for other measures of sparsity

# Skewed distributions




$p(h) = \exp(g(h))$
exponential: $g(h) = -|h|$
Cauchy: $g(h) = -\log(1 + h^2)$
Gaussian: $g(h) = -h^2/2$

[Figure: Dayan and Abbott, 2001]

# The sparse coding model

Single component model for image: $\mathbf{u} = \mathbf{g}h$.
Find $\mathbf{g}$ so that sparseness maximal, while $\langle h \rangle = 0$, $\langle h^2 \rangle = 1$. Multiple components:

$$\mathbf{u} = G\mathbf{h} + \mathbf{n}$$

Minimize [Olshausen and Field, 1996]
$E =$ [reconstruction error] $- \lambda$[sparseness]

- Factorial: $p(\mathbf{h}) = \prod_i p(h_i)$
- Sparse: $p(h_i) \propto \exp(g(h_i))$ (non-Gaussian)
  - Laplacian: $g(h) = -\alpha|h|$
  - Cauchy: $g(h) = -\log(\beta^2 + h^2)$
- $\mathbf{n}$ is a noise term
- Goal: find set of basis functions $G$ such that the coefficients $\mathbf{h}$ are as sparse and statistically independent as possible
- See D and A pp 378-383, and HHH §13.1.1-13.1.4

# Recognition step

- Suppose $G$ is given. For given image, what is $\mathbf{h}$?
- For $g(h)$ is Cauchy distribution, $p(\mathbf{h}|\mathbf{u}, \mathcal{G})$ is difficult to compute exactly
- The overcomplete model is not invertible

$$p(\mathbf{h}|\mathbf{u}) = \frac{p(\mathbf{u}|\mathbf{h})p(\mathbf{h})}{p(\mathbf{u})}$$

- Olshausen and Field (1996) used MAP approximation. As $p(\mathbf{u})$ does not depend on $\mathbf{h}$, we can find $\mathbf{h}$ by maximising:

$$\log p(\mathbf{h}|\mathbf{u}) = \log(p(\mathbf{u}|\mathbf{h})) + \log(p(\mathbf{h}))$$

# Recognition step

- We assume a sparse and independent prior $p(\mathbf{h})$, so

$$\log p(\mathbf{h}) = \sum_{a=1}^{N_h} g(h_a)$$

- Assuming Gaussian noise $\mathbf{n} \sim N(\mathbf{0}, \sigma^2 I)$, $p(\mathbf{u}|\mathbf{h})$ is drawn from a Gaussian distribution at $\mathbf{u} - G\mathbf{h}$ and variance $\sigma^2$:

$$\log p(\mathbf{h}|\mathbf{u}, \mathcal{G}) = -\frac{1}{2\sigma^2}|\mathbf{u} - G\mathbf{h}|^2 + \sum_{a=1}^{N_h} g(h_a) + \text{const}$$

# Recognition step

- At maximum (differentiate w.r.t. to $\mathbf{h}$)

$$\sum_{b=1}^{N_h} \frac{1}{\sigma^2}[\mathbf{u} - G\hat{\mathbf{h}}]_b G_{ba} + g'(\hat{h}_a) = 0$$

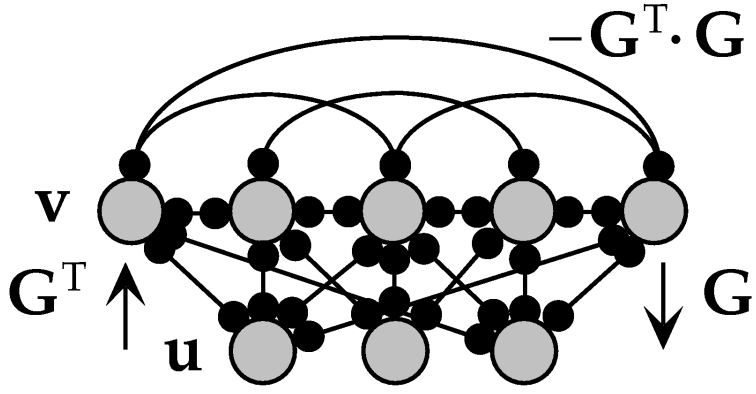
$$\text{or} \qquad \frac{1}{\sigma^2} G^T[\mathbf{u} - G\hat{\mathbf{h}}] + \mathbf{g}'(\hat{\mathbf{h}}) = \mathbf{0}$$

To solve this equation, follow dynamics

$$\tau_h \frac{dh_a}{dt} = \frac{1}{\sigma^2} \sum_{b=1}^{N_h} [\mathbf{u} - G\mathbf{h}]_b G_{ba} + g'(h_a)$$

Neural network interpretation (notation, $\mathbf{v} = \mathbf{h}$)

[Figure: Dayan and Abbott, 2001]




- Dynamics does gradient ascent on log posterior.
- A combination of feed forward excitation, lateral inhibition and relaxation of neural firing rates.
- Process is guaranteed only to find a local (not global) maximum

# Learning of the model

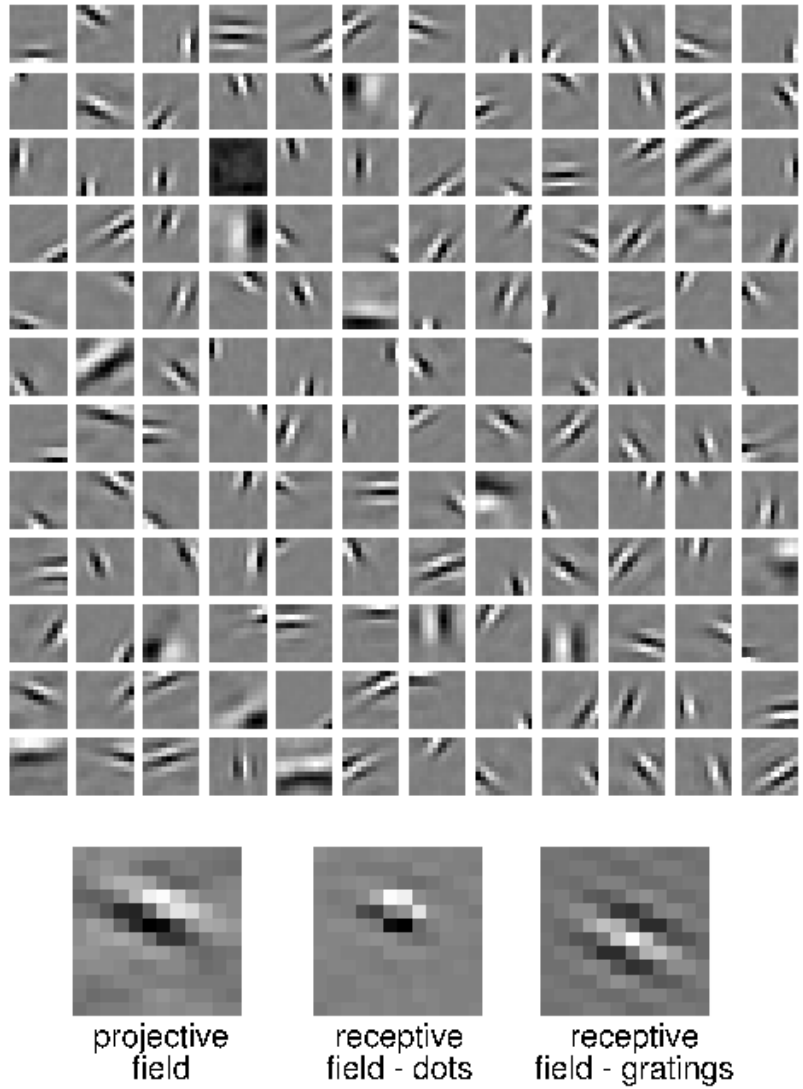
- Now we have $\mathbf{h}$, we can compare
- Log likelihood $L(\mathcal{G}) = \langle \log p(\mathbf{u}|\mathcal{G}) \rangle$. Learning rule:

$$\Delta G \propto \frac{\partial L}{\partial G}$$

- Basically linear regression (mean-square error cost)

$$\Delta G = \epsilon(\mathbf{u} - G\hat{\mathbf{h}})\hat{\mathbf{h}}^T$$

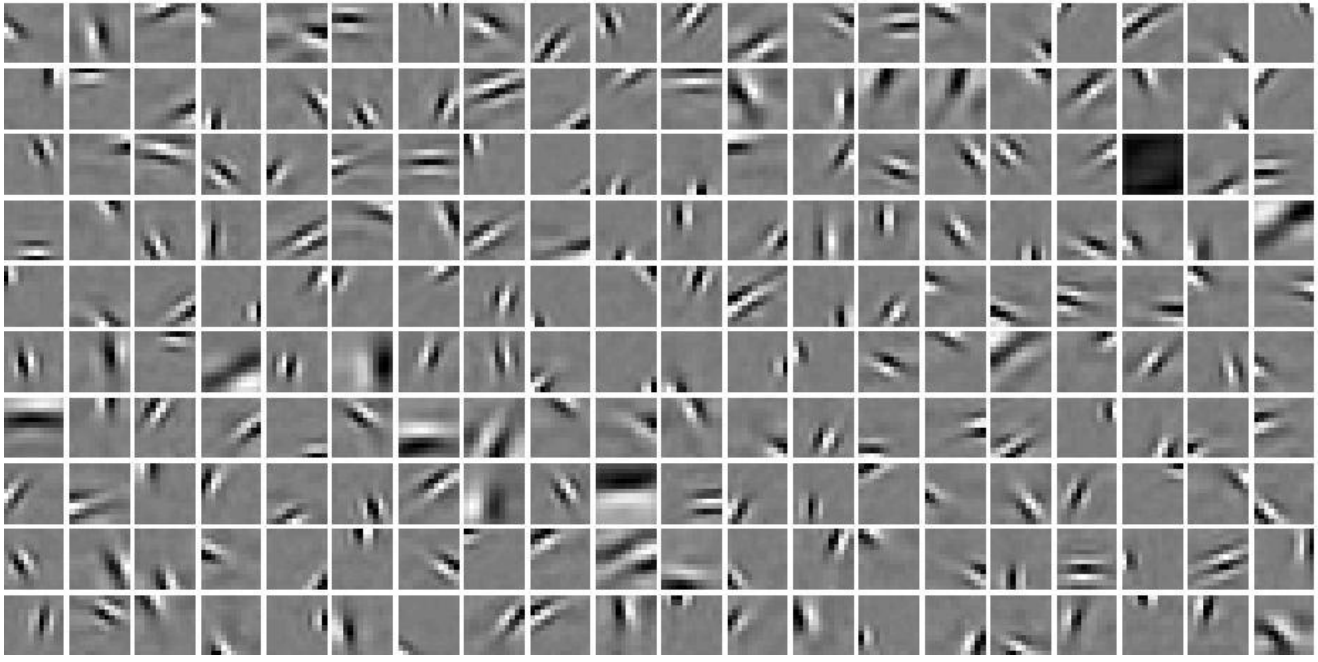
- Small values of $h$ can be balanced by scaling up $G$. Hence impose constraint on $\sum_b G_{ba}^2$ for each cause $a$ to encourage the variances of each $h_a$ to be approximately equal
- It is common to whiten the inputs before learning (so that $\langle \mathbf{u} \rangle = \mathbf{0}$ and $\langle \mathbf{u}\mathbf{u}^T \rangle = I$), to force the network to find structure beyond second order




[Figure: Dayan and Abbott (2001), after Olshausen and Field (1997)]

# Projective Fields and Receptive Fields

- Projective field for $h_a$ is $G_{ba}$ for all $b$ values
- Note resemblance to simple cells in V1
- Receptive fields: includes network interaction.
- Outputs of network are sparser than feedforward input, or pixel values
- Comparison with physiology: spatial-frequency bandwidth, orientation bandwidth

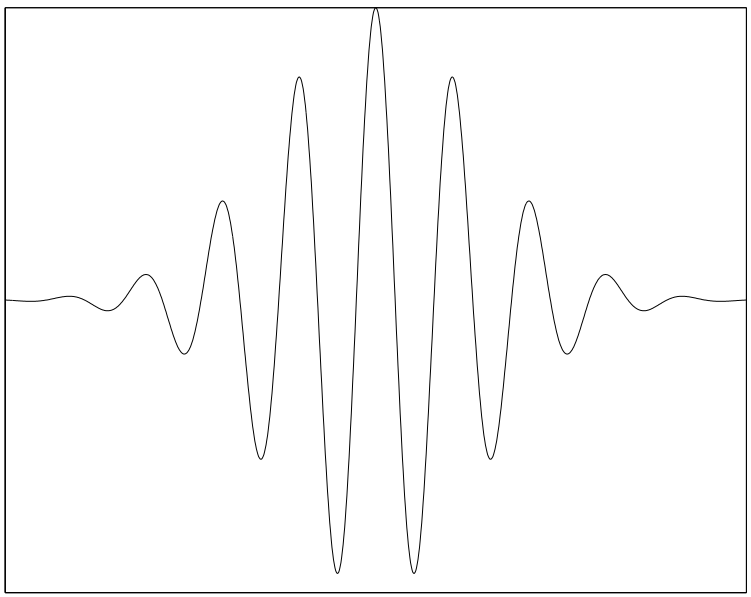


Overcomplete: 200 basis functions from 12 $\times$ 12 patches [Figure: Olshausen, 2005]

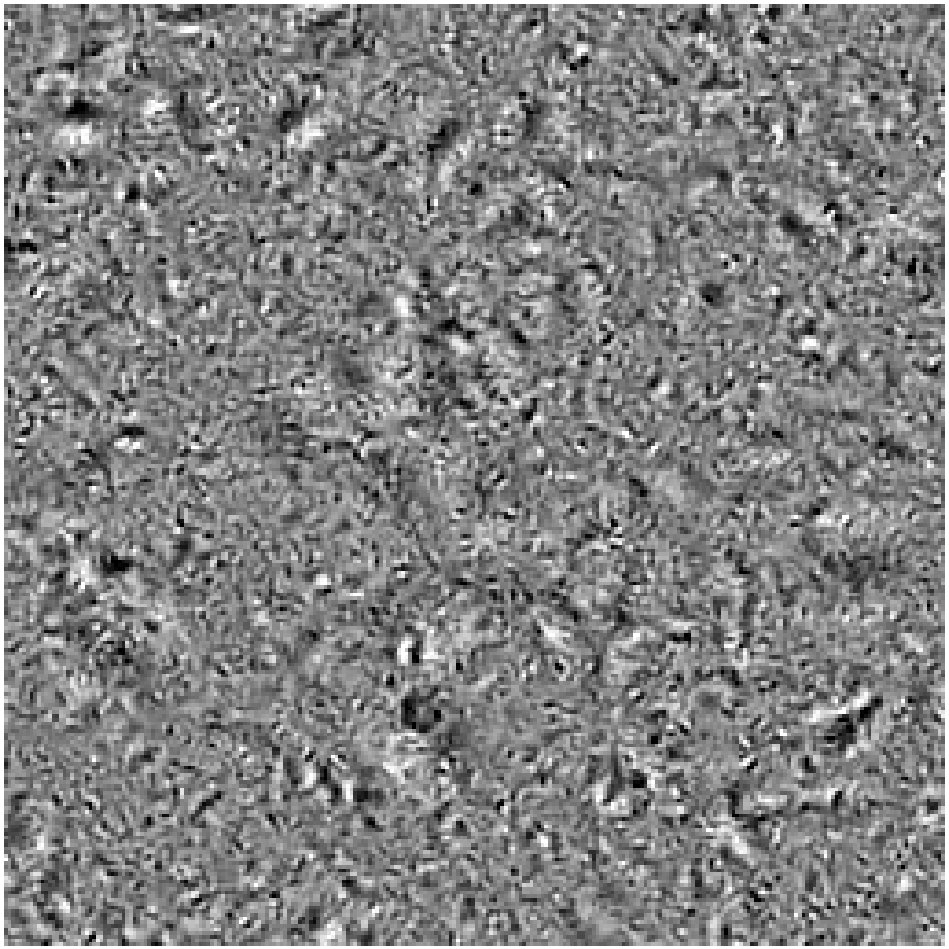
# Gabor functions

- Can be used to model the receptive fields.
- A sinusoid modulated by a Gaussian envelope

$$\frac{1}{2\pi\sigma_x\sigma_y}\exp\left(-\frac{x^2}{2\sigma_x^2}-\frac{y^2}{2\sigma_y^2}\right)\cos(kx-\phi)$$



# Image synthesis: sparse coding



[Figure: Olshausen, 2005]

# Spatio-temporal sparse coding

(Olshausen 2002)

$$\mathbf{u}(t) = \sum_{m=1}^{M} \sum_{n=1}^{n_m} h_i^m \mathbf{g}_m(t - \tau_i^m) + \mathbf{n}(t)$$

- $G$ is now 3-dimensional, having time slices as well
- Goal: find a set of space-time basis functions for representing natural images such that the time-varying coefficients $\{h_i^m\}$ are as sparse and statistically independent as possible over both space and time.
- 200 bases, $12 \times 12 \times 7$:
  http://redwood.berkeley.edu/bruno/bfmovie/bfmovie.html

# Sparse coding: limitations

- Sparseness-enforcing non-linearity choice is arbitrary
- Learning based on enforcing uncorrelated $h$ is ad hoc
- Unclear if $p(\mathbf{h})$ is a proper prior distribution

Solution: a generative model which describes how the image was generated from a transformation of the latent variables.

# ICA: Independent Components Analysis [Bell and Sejnowski, 1995]

- Linear network with output non-linearity $\mathbf{h} = W\mathbf{u}$, $y_j = f(h_j)$.
- $h_j$ are statistically independent random variables.
- $h_j$ are from a non-Gaussian distribution (as in sparse coding).
- Find weight matrix maximizing information between $\mathbf{u}$ and $\mathbf{y}$
- No noise (cf. Linsker): $I(\mathbf{u}, \mathbf{y}) = H(y) - H(y|u) = H(y)$

$$H(y) = \langle \log p(\mathbf{y}) \rangle_y = \langle \log p(\mathbf{u}) / \det J \rangle_u$$

$$\text{with } J_{ji} = \frac{\partial y_j}{\partial u_i} = \frac{\partial h_j}{\partial u_i} \frac{\partial y_j}{\partial h_j} = w_{ij} \prod_j f'(h_j)$$

(for a transformation, the PDF is multiplied by the absolute value of the determinant of the transformation matrix to ensure nominalisation)

# ICA: Independent Components Analysis [Bell and Sejnowski, 1995]

- Mutual information:

$$H(y) = \log\det W + \langle \sum_j logf'(h_j) \rangle + \text{const}$$

- Maximize entropy by producing a uniform distribution (histogram equalization):

$$p(h_i) = f'(h_i)$$

- Choose $f$ so that it encourages sparse $p(h)$, e.g. $1/(1+\mathrm{e}^{-h})$.
- For $f(h) = 1/(1+\mathrm{e}^{-h})$:

$$\frac{dH(y)}{dW} = (W^T)^{-1} + (\mathbf{1} - \mathbf{2y})\mathbf{x^T}$$

# ICA: how does it differ from PCA?

- The (symmetric) covariance matrix only constrains $n(n-1)/2$ components.
- Hence in a larger model (e.g. $n^2$) the coefficients are not fully constrained.
- The more random variables are added, the more Gaussian. So we look for the most non-Gaussian projection.
- Often, but not always, this is most sparse projection.
- Can use ICA to de-mix (e.g. blind source separation of sounds)

left: whitened by PCA; middle: 2 mixed independent components; right: 2 independent compoennts

# ICA as generative model

- Simplify sparse coding network, let $G$ be square
- $\mathbf{u} = G\mathbf{h}$, $W = G^{-1}$

$$p(\mathbf{u}) = |\det W| \prod_{a=1}^{N_h} p_h([W\mathbf{u}]_a)$$

- Log likelihood

$$L(W) = \left\langle \sum_a g([W\mathbf{u}]_a) + \log|\det W| \right\rangle + \text{const}$$

- See Dayan and Abbott pp 384-386 [also HHH ch 7]

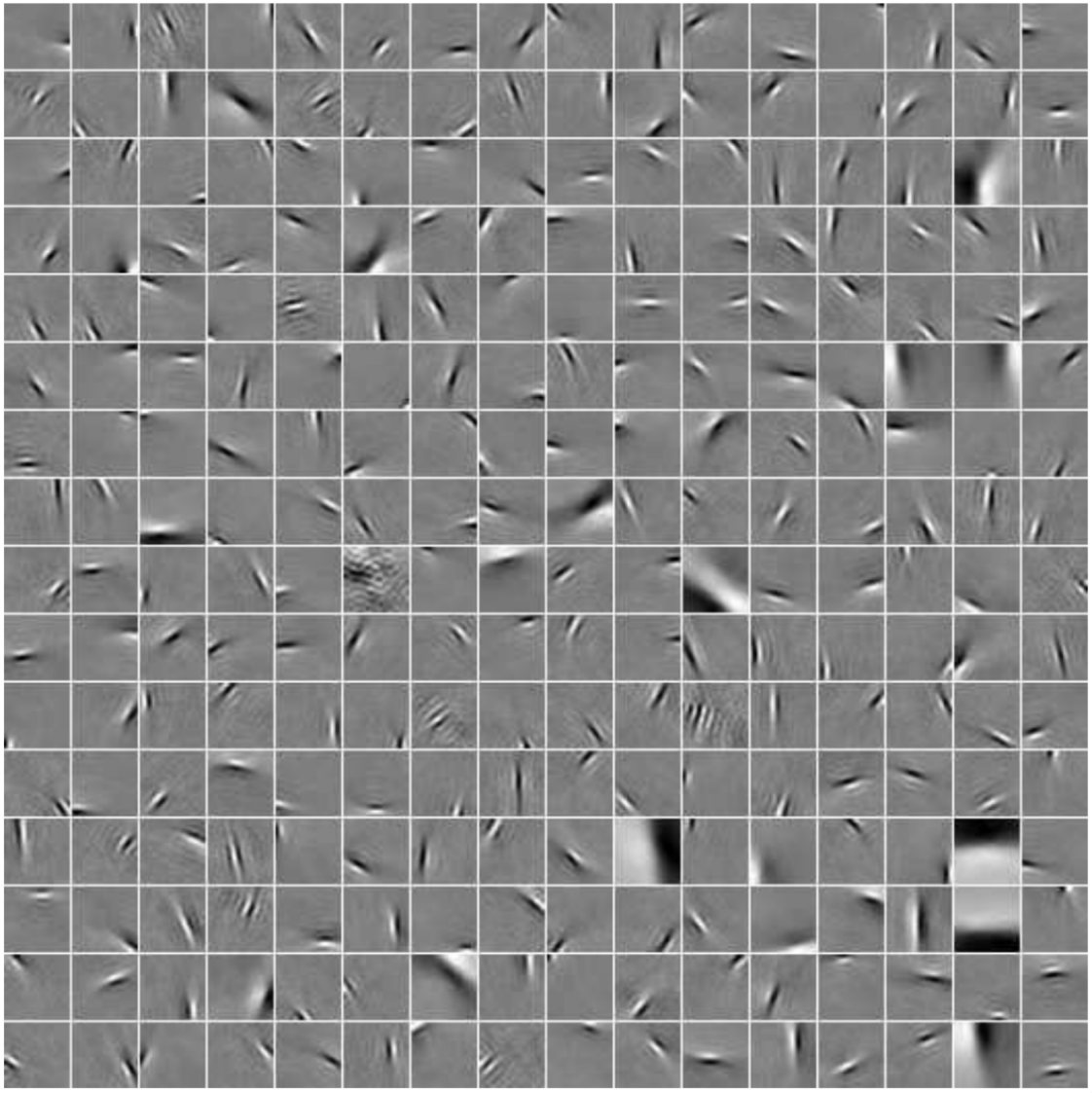
- Stochastic gradient ascent gives update rule

$$\Delta W_{ab} = \epsilon([W^{-1}]_{ba} + g'(h_a)u_b)$$

  using $\partial \log \det W / \partial W_{ab} = [W^{-1}]_{ba}$
- Natural gradient update: multiply by $W^T W$ (which is positive definite) to get

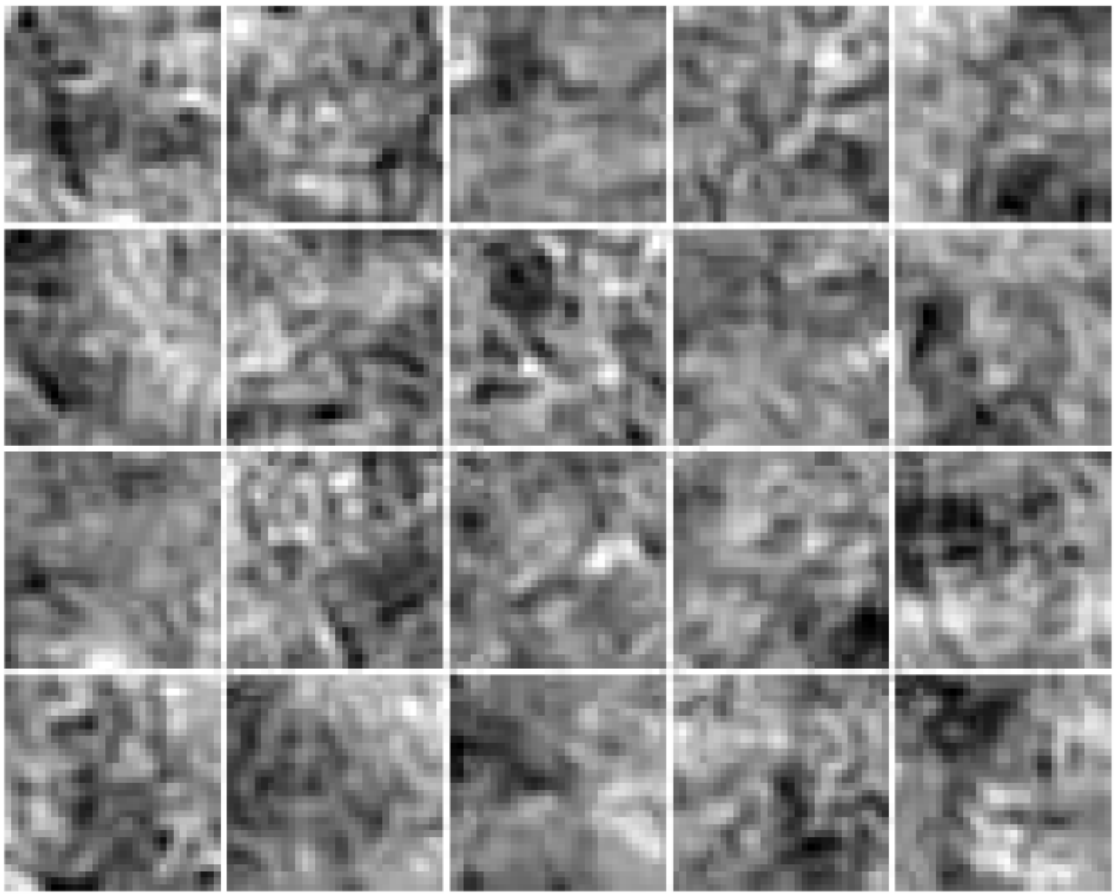
$$\Delta W_{ab} = \epsilon(W_{ab} + g'(h_a)[\mathbf{h}^T W]_b)$$

# ICA features



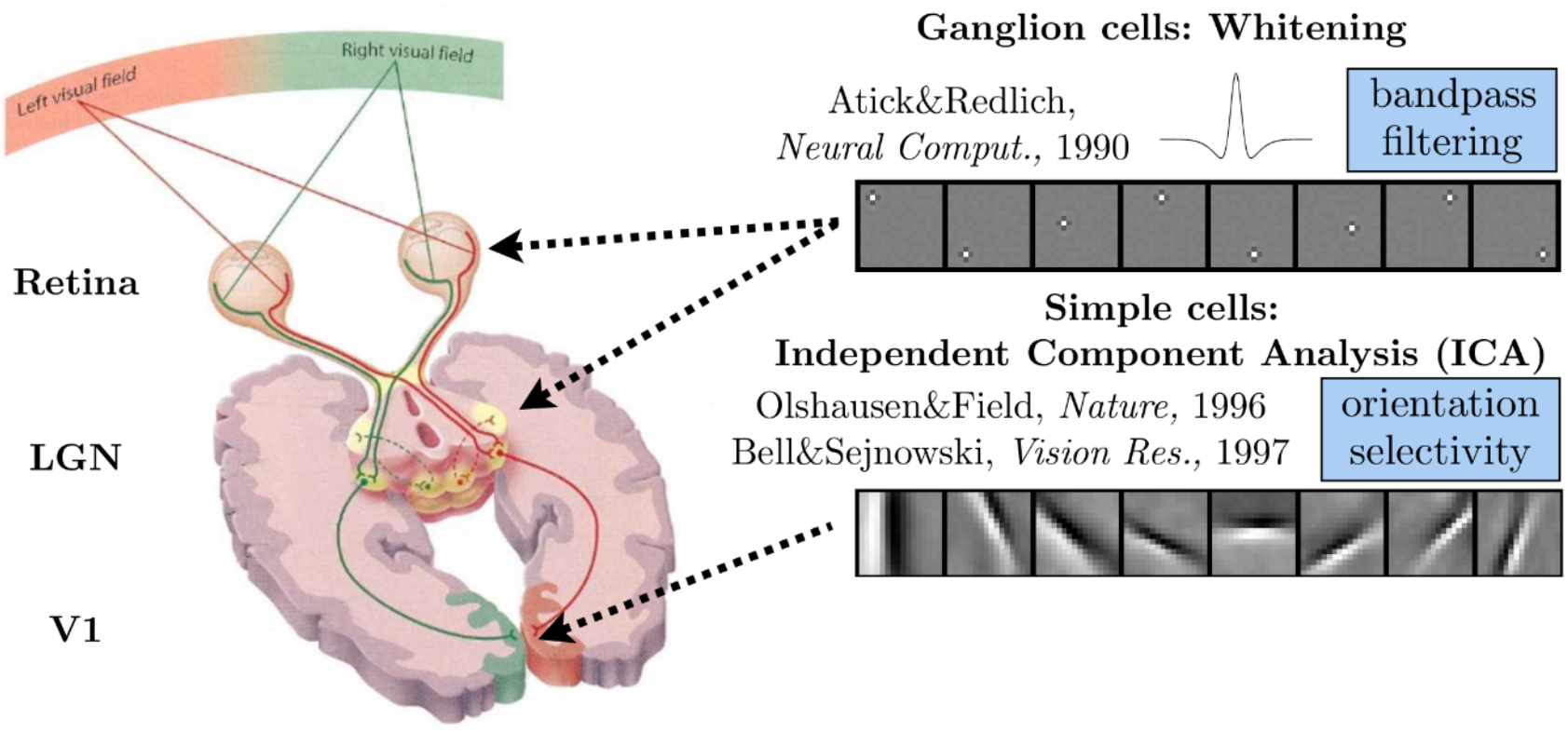
[Hyvärinen et al., 2009]

# ICA synthesised images



[Hyvärinen et al., 2009]

# The visual system




[Figure from Matthias Bethge]

# Are Gabor patches what we want?

Dayan and Abbott (2001) p. 382 say:

> *In a generative model, projective fields are associated with the causes underlying the visual images presented during training. The fact that the causes extracted by the sparse coding model resemble Gabor patches within the visual field is somewhat strange from this perspective. It is difficult to conceive of images arising from such low-level causes, instead of causes couched in terms of objects within images, for example. From the perspective of good representation, causes more like objects and less like Gabor patches would be more useful. To put this another way, although the prior distribution over causes biased them toward mutual independence, the causes produced by the recognition model in response to natural images are not actually independent...*

*This is due to the structure in images arising from more complex objects than bars and gratings. It is unlikely that this higher-order structure can be extracted by a model with only one set of causes. It is more natural to think of causes in a hierarchical manner, with causes at a higher level accounting for structure in the causes at a lower level. The multiple representations in areas along the visual pathway suggest such a hierarchical scheme, but the corresponding models are still in the rudimentary stages of development.*

# Summary

- Both ICA and Sparse Coding lead to similar RFs, and sparse output for natural images.
- Both give good description of V1 simple cell RFs, although not perfectly [van Hateren and van der Schaaf, 1998] )
- (and so do many other algorithms)
- Different objectives:
  - ICA - maximize information
  - Sparse Coding - sparse reconstruction
- What about deeper layers? See [Hyvärinen et al., 2009] for discussion of these points.

# References I



Bell, A. J. and Sejnowski, T. J. (1995).
An information-maximisation approach to blind separation and blind deconvolution.
*Neural Comp.*, 6:1004–1034.


Hyvärinen, A., Hurri, J., and Hoyer, P. (2009).
*Natural Image Statistics*.
Spinger.


Olshausen, B. A. and Field, D. J. (1996).
Emergence of simple cell receptive field properties by learning a sparse code for natural images.
*Nature*, 381:607–609.


Olshausen, B. A. and Field, D. J. (2004).
Sparse coding of sensory inputs.
*Curr Opin Neurobiol*, 14(4):481–487.