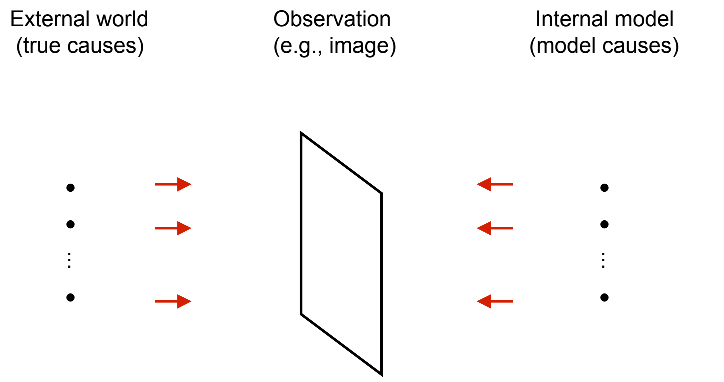
Lecture 15
Representational learning

jochen.braun@nat.uni-magdeburg.de

June 21, 2006

Credits:

Dayan & Abbott, Chapter 10
Olshausen & Field (2000) American Scientist 88: 238.
Olshausen & Field (2004) Current Biology 14: 481-487.
Vinje & Gallant (2002) J Neuroscie 22: 2904-2915.



External world (true causes)

Observation (e.g., image)

Internal model (model causes)

Our goal is to understand complex observations in terms of few causes

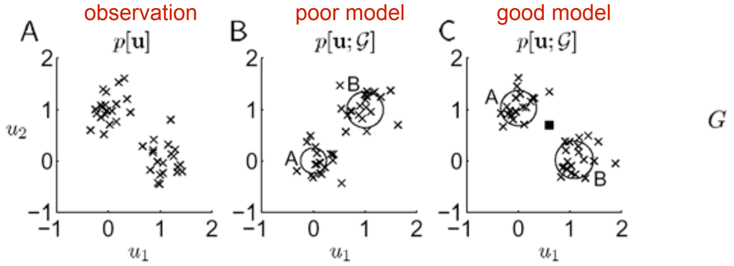
1 Introduction to causal models

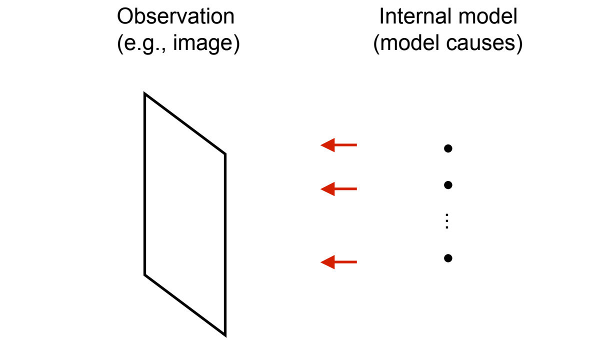
The structure of some data distributions suggests discrete underlying causes. In the example from the book, there are clearly two clusters of data points $\boldsymbol{u}$. A more compact and causal description of the same data would involve a causal variable v, which in this case would take one of two values, A or B. The value of v cannot always be unambiguously determined.

In *probabilistic recognition*, we want to know the probability that a given data point $\boldsymbol{u}$ was caused by cause A or by cause B. In *deterministic recognition*, we simply want to know the most likely cause, A or B.

We consider models that infer causes without supervision or additional information. The success of such models is judged by their ability to reproduce (and thus 'explain') the input data. We iteratively adjust the parameters of a generative model until we obtain a good match between observation and reproduction.

Data vector	$\boldsymbol{u} = (u_1, u_2)$
Scalar cause	$v \in \{A, B\}$





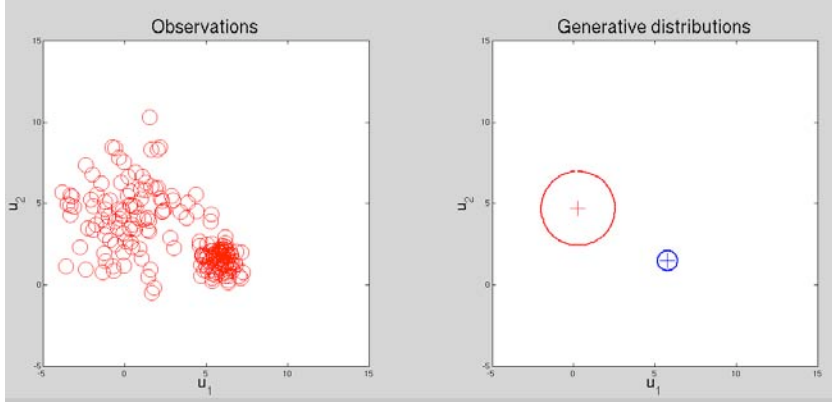
Observation (e.g., image)

Internal model (model causes)

Generative model	observation	←	cause
Recognition model	observation	→	cause
Iteration	observation	↔	cause

To iteratively improve recognition, we need both generation and recognition

Generative model



$$G = \{\gamma_A, \boldsymbol{g}_A, \boldsymbol{g}_B, \Sigma_A, \Sigma_B\}$$

1.1 Generative model

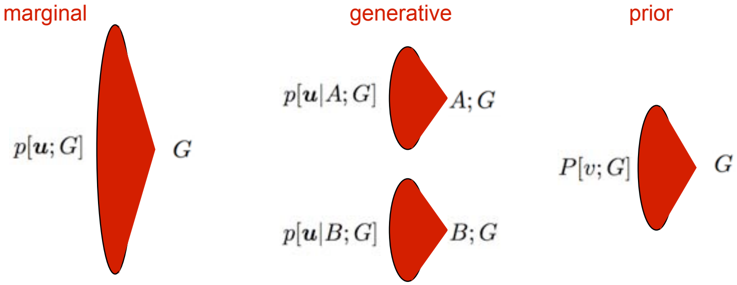
Prior distribution of causes $\quad P[v;G] = \gamma_v \quad$ with $\quad v = A$ or $B \quad$ and $\quad \sum \gamma_v = 1$

Generative distribution (given cause) $\quad p[\boldsymbol{u}|v;G] = \frac{1}{2\pi\,\Sigma_{A,B}} \exp\left[-\frac{(u_1 - g_1^{A,B})^2 + (u_2 - g_2^{A,B})^2}{2\Sigma_{A,B}}\right]$

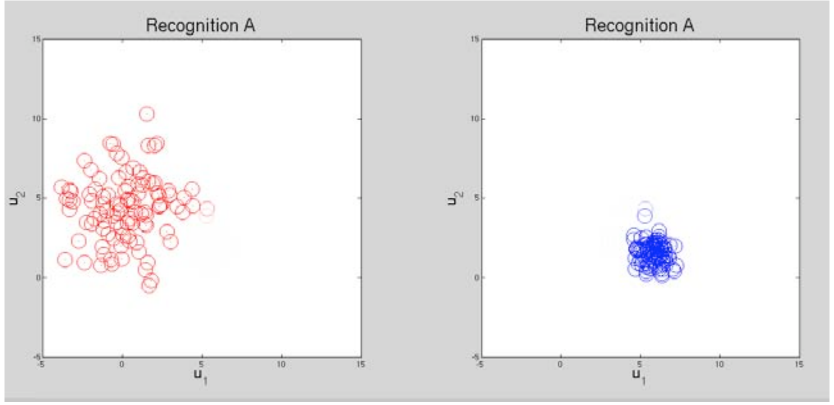
Parameter set $\quad G = \{\gamma_A, \boldsymbol{g}_A, \boldsymbol{g}_B, \Sigma_A, \Sigma_B\}$

Marginal distribution $\quad p[\boldsymbol{u};G] = \sum_v p[\boldsymbol{u}|v;G]\,P[v;G]$

The structure of our generative model reflects heuristic information (prejudices, assumptions, analogies, ...).



Recognition model



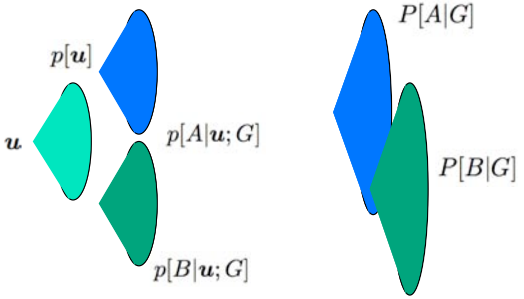
$$P[A|\boldsymbol{u};G] \qquad P[B|\boldsymbol{u};G]$$

1.2 Recognition model

Once the generative model is optimized, we can use it to classify new observations in terms of probable causes, in other words, we can use it for "recognition" or "classification" problems. To this end, we use Bayes' theorem to compute the most likely cause, given a particular observation:

$$P[v|\boldsymbol{u};G] = \frac{p[\boldsymbol{u}|v;G]\,P[v;G]}{p[\boldsymbol{u};G]}$$

Not all situations are *invertible* in this way, that is, we cannot always compute the conditional distribution of causes from the conditional distribution of events.



2 Expectation maximization

We introduce "expectation maximization" as a method for adjusting a generative model G. Our generative model includes two causes, each producing Gaussian-distributed observations. In this case, G comprises the means $\boldsymbol{g}_{A,B}$ and variances $\Sigma_{A,B}$ of the conditional distributions $p[u|v;G]$ and the prior probabilities $p[A|G] = \gamma_A$ and $p[B|G] = \gamma_B$.

If we knew which observation stems from which cause, it would be a simple matter to compute the parameters in G. As we do not have this information yet, we instead use the classification distribution $P[v|\boldsymbol{u};G]$ as a current best guess.

The EM algorithm consists of two alternating steps, the E (expectation) step of inferring "probable causes" from the classification distribution $P[v|\boldsymbol{u};G]$, and the M (maximization) step of computing parameters G from weighted averages over observations $\boldsymbol{u}$. It is far from obvious that this process (which seems suspiciously circular) will converge to an optimal generative model G!

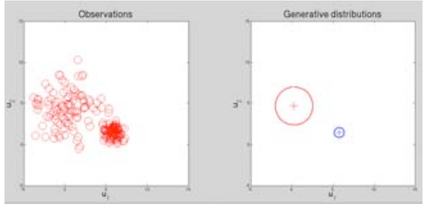
$$\begin{tikzcd}
G = \{\gamma_A, \boldsymbol{g}_A, \boldsymbol{g}_B, \Sigma_A, \Sigma_B\} \arrow[dr] & & \\
& P[A|\boldsymbol{u};G] \arrow[dl] & P[B|\boldsymbol{u};G] \\
G = \{\gamma_A, \boldsymbol{g}_A, \boldsymbol{g}_B, \Sigma_A, \Sigma_B\} \arrow[dr] & & \\
& P[A|\boldsymbol{u};G] \arrow[dl] & P[B|\boldsymbol{u};G] \\
G = \{\gamma_A, \boldsymbol{g}_A, \boldsymbol{g}_B, \Sigma_A, \Sigma_B\} \arrow[dr] & & \\
& P[A|\boldsymbol{u};G] \arrow[dl] & P[B|\boldsymbol{u};G] \\
{} & &
\end{tikzcd}$$

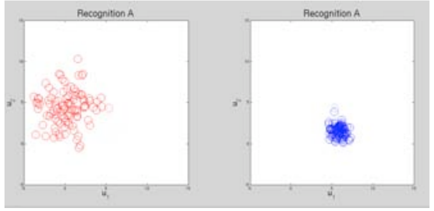
Expectation step

$$P[v|\boldsymbol{u};G] = \frac{p[\boldsymbol{u}|v;G]\,P[v;G]}{p[\boldsymbol{u};G]}$$

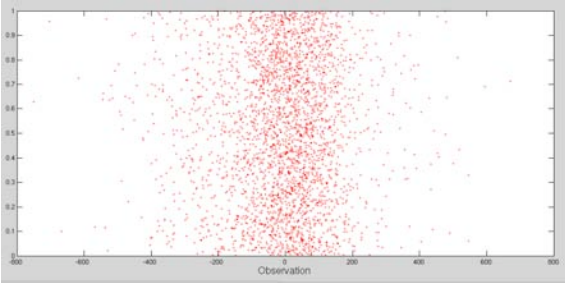
Maximization step

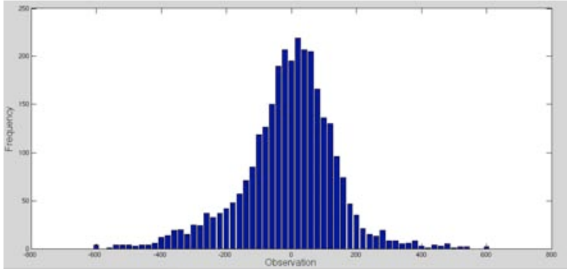
$$\gamma_v = \langle P[v|\boldsymbol{u};G]\rangle_v \qquad \boldsymbol{g}_v = \frac{\langle \boldsymbol{u}\; P[v|\boldsymbol{u};G]\rangle_v}{\gamma_v} \qquad \Sigma_v = \frac{\langle (\boldsymbol{u}-\boldsymbol{g}_v)^2\; P[v|\boldsymbol{u};G]\rangle_v}{\gamma_v}$$





Example: 1D observations with two Gaussian causes





$$G = \{\gamma_A, \boldsymbol{g}_A, \boldsymbol{g}_B, \Sigma_A, \Sigma_B\} \longrightarrow P[A|\boldsymbol{u};G] \qquad P[B|\boldsymbol{u};G]$$

2.1 Expectation step

Probability of observing u, given cause A or cause B:

$$p(u|A) = \frac{1}{\sqrt{2\pi\,\Sigma_A^2}}\,\exp\left(-\frac{(u-g_A)^2}{2\,\Sigma_A^2}\right) \qquad 1 = \int p_A(u)\,du$$

$$p(u|B) = \frac{1}{\sqrt{2\pi\,\Sigma_B^2}}\,\exp\left(-\frac{(u-g_B)^2}{2\,\Sigma_B^2}\right) \qquad 1 = \int p_B(u)\,du$$

Joint probability of observing u *and* it being due to cause A or cause B:

$$p(u;A) = \gamma_A\,p(u|A) \qquad \gamma_A = \int\!\!\int p(u;A)\,du$$

$$p(u;B) = (1-\gamma_A)\,p(u|B) \qquad 1-\gamma_A = \int\!\!\int p(u;B)\,du$$

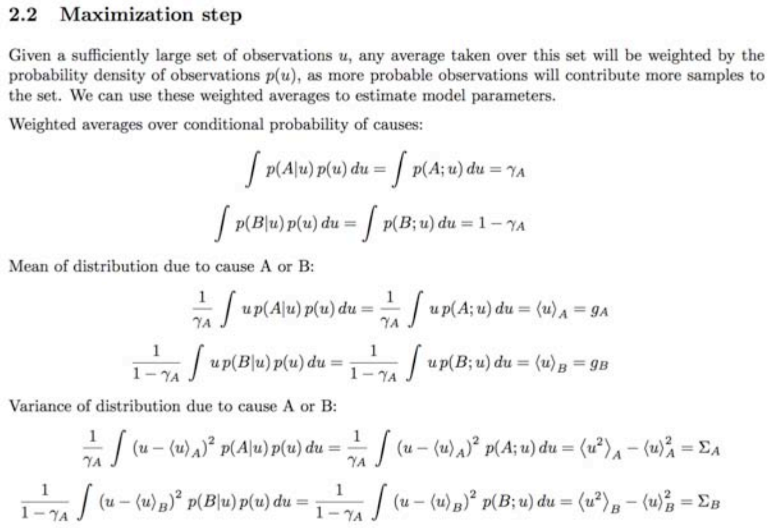
Total probability of observing u (due to either cause).

$$p(u) = p(u;A) + p(u;B) \qquad 1 = \int\!\!\int p(u)\,du$$

Conditional probability of cause A or B, given an observation u:

$$p(A|u) = \frac{p(u;A)}{p(u)} \qquad p(B|u) = \frac{p(u;B)}{p(u)}$$

$$P[A|\boldsymbol{u};G] \qquad P[B|\boldsymbol{u};G] \qquad \longrightarrow \qquad G = \{\gamma_A, \boldsymbol{g}_A, \boldsymbol{g}_B, \Sigma_A, \Sigma_B\}$$



2.2 Maximization step

Given a sufficiently large set of observations u, any average taken over this set will be weighted by the probability density of observations $p(u)$, as more probable observations will contribute more samples to the set. We can use these weighted averages to estimate model parameters.

Weighted averages over conditional probability of causes:

$$\int p(A|u)\,p(u)\,du = \int p(A;u)\,du = \gamma_A$$

$$\int p(B|u)\,p(u)\,du = \int p(B;u)\,du = 1 - \gamma_A$$

Mean of distribution due to cause A or B:

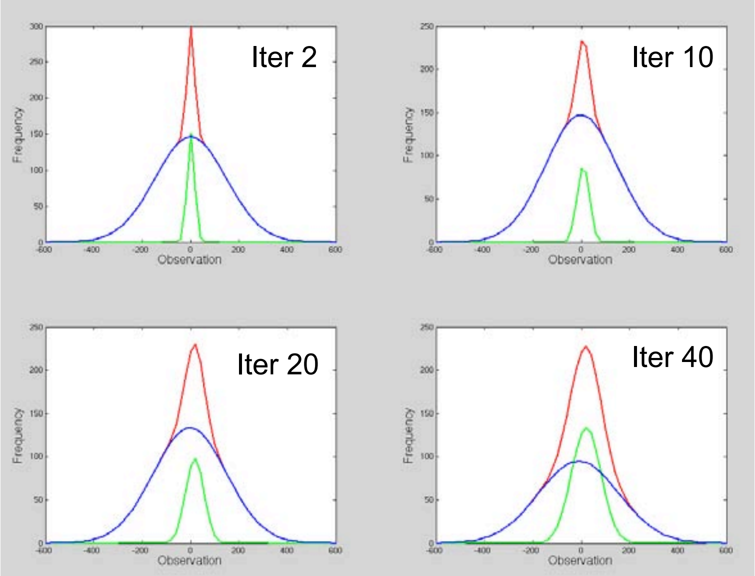
$$\frac{1}{\gamma_A}\int u\,p(A|u)\,p(u)\,du = \frac{1}{\gamma_A}\int u\,p(A;u)\,du = \langle u\rangle_A = g_A$$

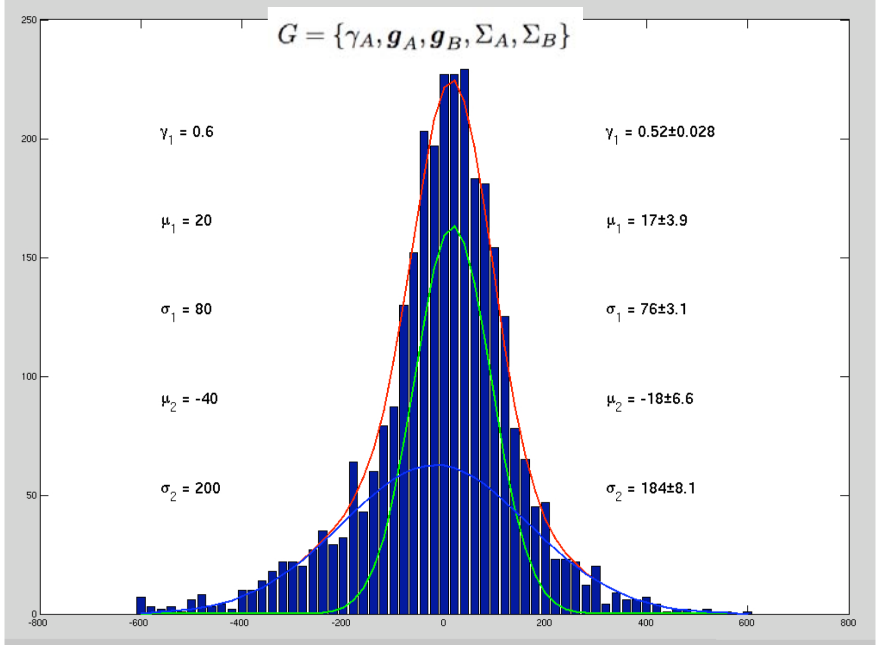
$$\frac{1}{1-\gamma_A}\int u\,p(B|u)\,p(u)\,du = \frac{1}{1-\gamma_A}\int u\,p(B;u)\,du = \langle u\rangle_B = g_B$$

Variance of distribution due to cause A or B:

$$\frac{1}{\gamma_A}\int (u-\langle u\rangle_A)^2\,p(A|u)\,p(u)\,du = \frac{1}{\gamma_A}\int (u-\langle u\rangle_A)^2\,p(A;u)\,du = \langle u^2\rangle_A - \langle u\rangle_A^2 = \Sigma_A$$

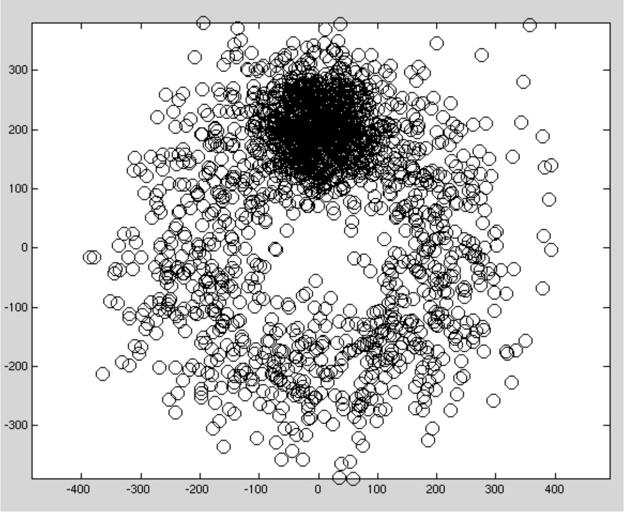
$$\frac{1}{1-\gamma_A}\int (u-\langle u\rangle_B)^2\,p(B|u)\,p(u)\,du = \frac{1}{1-\gamma_A}\int (u-\langle u\rangle_B)^2\,p(B;u)\,du = \langle u^2\rangle_B - \langle u\rangle_B^2 = \Sigma_B$$



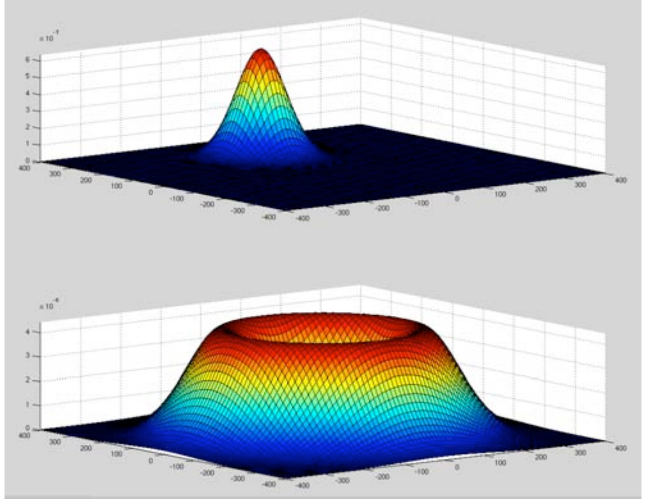


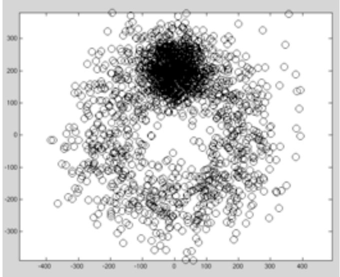
$$G = \{\gamma_A, \boldsymbol{g}_A, \boldsymbol{g}_B, \Sigma_A, \Sigma_B\}$$

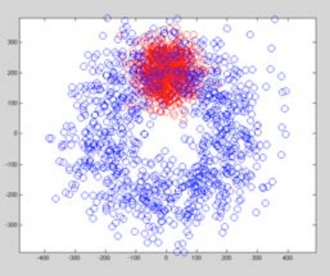
Example: 2D observations with non-Gaussian causes

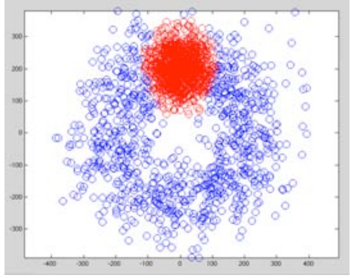


Generative model, with heuristics




Observations


True causes


Most probable causes

3 Mixture of Gaussians

A generative model with N_v separate Gaussian distributions is called "mixture of Gaussians". For observations $\boldsymbol{u}$ with N_u dimensions, the model is defined by

Generative model

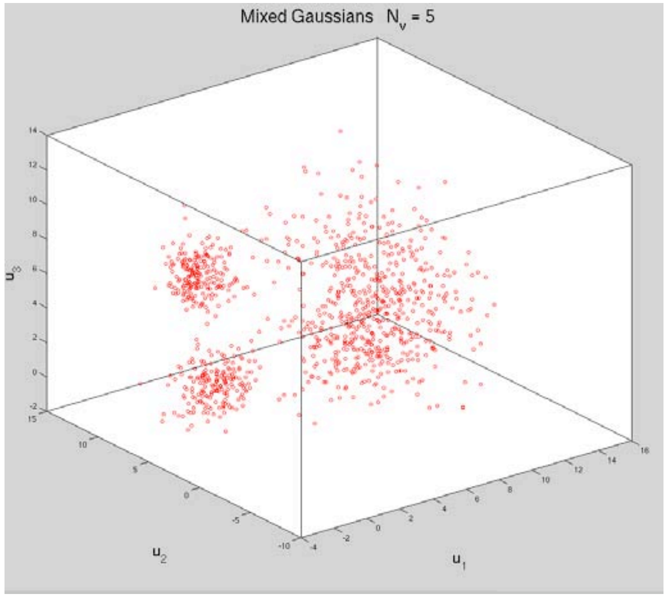
$$P[v;G] = \gamma_v \qquad\qquad p[\boldsymbol{u}|v;G] = N(\boldsymbol{u};\boldsymbol{g}_v,\Sigma_v)$$

$$N(\boldsymbol{u};\boldsymbol{g},\Sigma) = \frac{1}{(2\pi\,\Sigma)^{N_u/2}}\exp\left(-\frac{|\boldsymbol{u}-\boldsymbol{g}|^2}{2\Sigma}\right)$$

Here, $N()$ is an N_u-dimensional Gaussian distribution with mean $\boldsymbol{g}$ and identical variance Σ in all dimensions. Once this generative model has been optimized, the associated recognition model is given by:

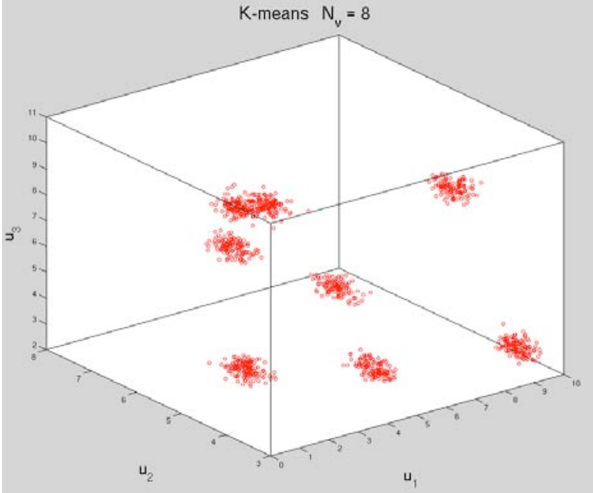
Recognition model

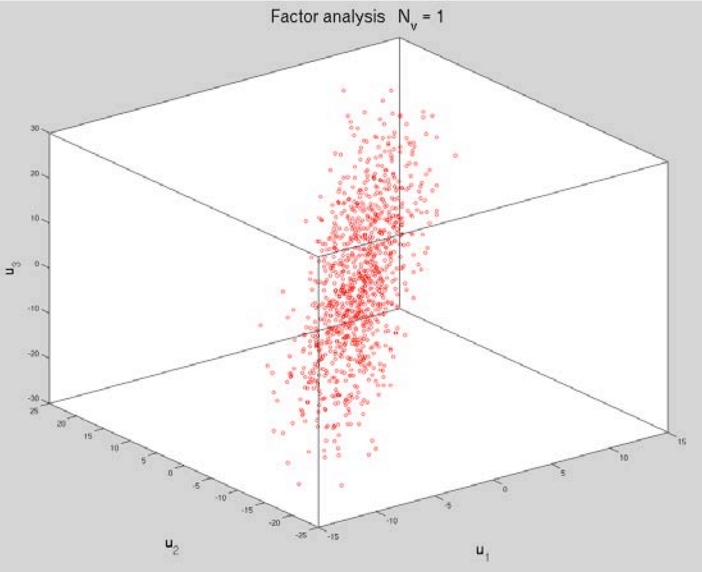
$$P[v|\boldsymbol{u};G] = \frac{\gamma_v\,N(\boldsymbol{u};\boldsymbol{g}_v,\Sigma_v)}{\sum_{v'}\gamma_{v'}\,N(\boldsymbol{u};\boldsymbol{g}_{v'},\Sigma_{v'})}$$



3.1 Sparseness of causes

In a "mixture of Gaussian" model, the causes are discrete and therefore sparse. In the degenerate case of $\Sigma \to 0$ the recognition model becomes increasingly deterministic.





4 Factor analysis

What if our causes are continuously distributed? The illustration on the next page suggests that observations are distributed along a line. We can capture this by postulating a normal distribution of causes v and generative distributions of which the mean depends linearly on v:

Generative model

$$\boldsymbol{u} = \{u_1, u_2, u_3\}$$

$$p[v; G] = \frac{1}{\sqrt{2\pi}} \exp\left(-v^2/2\right)$$

$$p[\boldsymbol{u}|v; G] = \frac{1}{\sqrt{(2\pi)^3\, \Sigma_1\, \Sigma_2\, \Sigma_3}} \exp\left(-\frac{(u_1 - v g_1)^2}{2\,\Sigma_1} - \frac{(u_2 - v g_2)^2}{2\,\Sigma_2} - \frac{(u_3 - v g_3)^2}{2\,\Sigma_3}\right)$$

Recognition model

$$p[v|\boldsymbol{u}; G] = \frac{1}{(2\pi)^{3/2}\, \Psi_1\, \Psi_2\, \Psi_3} \exp\left(-\frac{[v_1 - W_1(\boldsymbol{u})]^2}{2\,\Psi_1} - \frac{[v_2 - W_2(\boldsymbol{u})]^2}{2\,\Psi_2} - \frac{[v_3 - W_3(\boldsymbol{u})]^2}{2\,\Psi_3}\right)$$

$$\begin{aligned} W_1 &= w_{11}\, u_1 + w_{12}\, u_2 + w_{13}\, u_3 \\ W_2 &= w_{21}\, u_1 + w_{22}\, u_2 + w_{23}\, u_3 \\ W_3 &= w_{31}\, u_1 + w_{32}\, u_2 + w_{33}\, u_3 \end{aligned}$$

4.1 General case

In general, factor analysis uses an N_v-dimensional vector of causes v, drawn from a Gaussian prior distribution. The generative distributions are also Gaussian, with a mean that depends linearly on v and variances that are fixed.

Generative model

$$p[\boldsymbol{v}; G] = N(\boldsymbol{v}; \boldsymbol{0}, 1) \qquad\qquad p[\boldsymbol{u}|\boldsymbol{v}; G] = N(\boldsymbol{u}; \boldsymbol{G}\cdot\boldsymbol{v}, \boldsymbol{\Sigma})$$

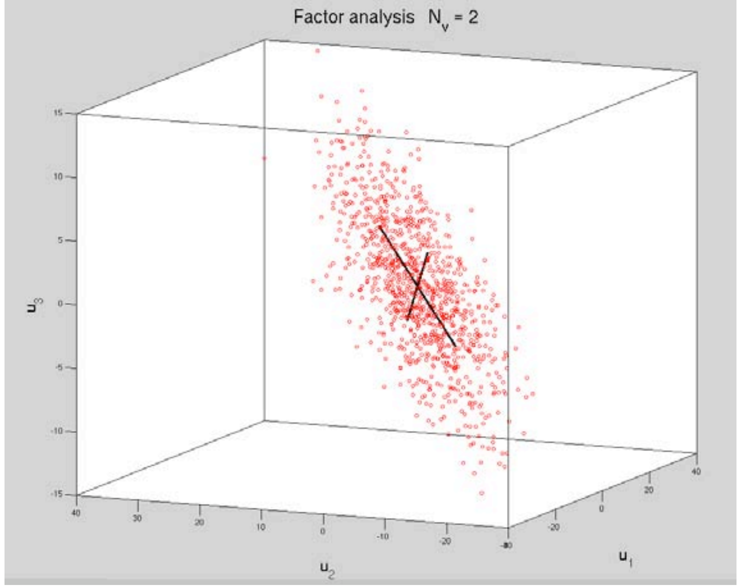
$$N(\boldsymbol{u}; \boldsymbol{g}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^{N_u}\, \Sigma_1 \ldots \Sigma_{N_u}}} \exp\left(-\frac{1}{2}\,(\boldsymbol{u} - \boldsymbol{g})\cdot\boldsymbol{\Sigma}^{-1}\cdot(\boldsymbol{u} - \boldsymbol{g})\right)$$

$$\boldsymbol{\Sigma} = diag\,(\Sigma_1, \ldots, \Sigma_{N_u}) \qquad\qquad \boldsymbol{\Sigma}^{-1} = diag\,(1/\Sigma_1, \ldots, 1/\Sigma_{N_u})$$

Recognition model

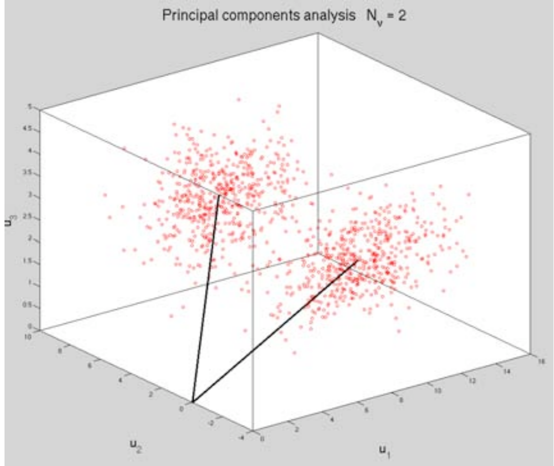
$$p[\boldsymbol{v}|\boldsymbol{u}; G] = N(\boldsymbol{v}; \boldsymbol{W}\cdot\boldsymbol{u}, \boldsymbol{\Psi}$$

$$\boldsymbol{\Psi} = \left(\boldsymbol{I} + \boldsymbol{G}^T\cdot\boldsymbol{\Sigma}^{-1}\cdot\boldsymbol{G}\right)^{-1} \qquad\qquad \boldsymbol{W} = \boldsymbol{\Psi}\cdot\boldsymbol{G}^T\cdot\boldsymbol{\Sigma}^{-1}$$



4.2 Degenerate case: Principal components analysis

In the degenerate case of $\Sigma \to 0$, the distribution of causes becomes 'sparse'. This case is better known under the name of "principal components analysis", or "PCA".



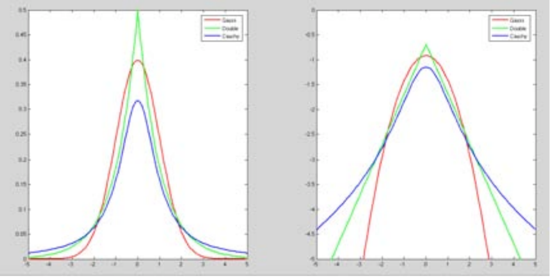
5 Sparse causes of images (Sparse coding)

5.1 Sparse distributions

A distribution is called 'sparse' when it generates values near zero and values far from zero *more often* than a comparable Gaussian distribution. It follows that a 'sparse' distribution generates intermediate values *less often* than a Gaussian distribution. Distributions of this type are also called 'heavy-tailed'.

As an example, consider the following three distributions:

Gaussian	$p(v) \propto \exp\left(-\frac{v^2}{2}\right)$	$k = 0$
Double exp	$p(v) \propto \exp\left(-\lvert v\rvert\right)$	$k = 3$
Cauchy	$p(v) \propto \frac{1}{1+v^2}$	$k = \infty$



5.2 Causes of natural images

To model the generation of natural images, we can choose causes of different complexity.

5.2.1 Single pixels

For example, we can apply one generative distribution to each image pixel, specifying its luminance value. In this case, we assume as many 'causes' as there are pixels. For natural scenes, an exponential distribution is more appropriate than a Gaussian distribution.

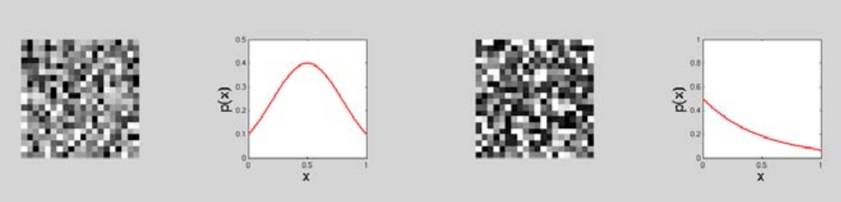
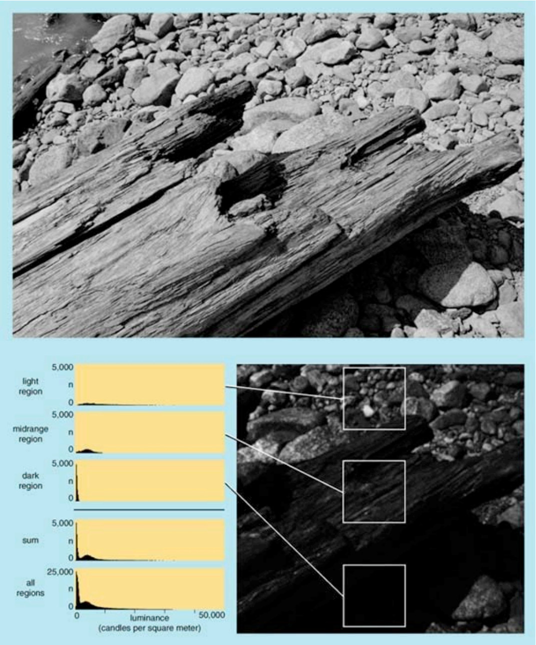
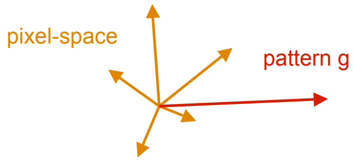


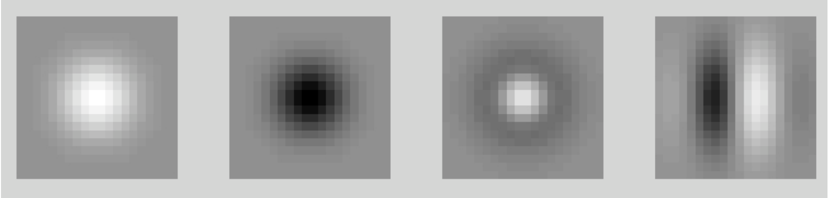
Figure 1. Images of the natural environment, such as this view of a log resting on a stony embankment (top), exhibit a surprising degree of statistical similarity. To investigate these qualities, the authors had first to remove the effects of the photographic process from their images, yielding estimates for the actual brightness (luminance) in each pixel. Because luminance spans an enormous range—it varies from about 100 to 40,000 candles per square meter in this image—linearly scaling these values to the shades that can be printed makes the scene look strangely dim and stark (lower right). Histograms of pixel intensity (yellow panels) show that the distribution of luminance values is short and wide in a light region, whereas it is narrow and peaked in a dark area. Summing the results from the three sample regions (white boxes) produces a distribution skewed toward low values, one that matches the shape of the histogram obtained for the image as a whole.



5.2.2 Filters

A slightly better way to generate natural images is to use a population of linear filters. Filters overlap, so that each contributes to many pixels. Formally, we treat the image as a high-dimensional vector $\boldsymbol{u}$, with as many components as there are pixels. To keep things simple, we ensure $\langle \boldsymbol{u} \rangle = \mathbf{0}$. In this case, each cause v is associated with a particular direction $\boldsymbol{g}_v$ in $\boldsymbol{u}$-space, which specifies ratios of luminance values for different pixels. By choosing an appropriate direction, we can generate any pattern of luminance values we choose.





6 Sparse version of factor analysis

Olshausen and Field (1997) suggested a non-linear version of factor analysis. In this approach, the generative distribution of $\boldsymbol{u}$ given $\boldsymbol{v}$ is still Gaussian, but the prior distribution over causes is sparse:

$$p[\boldsymbol{v};G] \propto \prod_{a=1}^{N_v} \exp[-\alpha|v_a|] \qquad\qquad p[\boldsymbol{u}|\boldsymbol{v};G] = N(\boldsymbol{u};\boldsymbol{G}\cdot\boldsymbol{v},\boldsymbol{\Sigma})$$

There is no simple way to invert this generative model into a recognition model. In consequence, there is also no simple way to adjust the parameters of the generative model. To iterate, E and M steps, Olshausen and Field used a neural network between $\boldsymbol{u}$ and $\boldsymbol{v}$ layers to set recurrent weights $\boldsymbol{G}$.

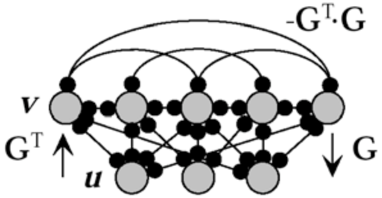
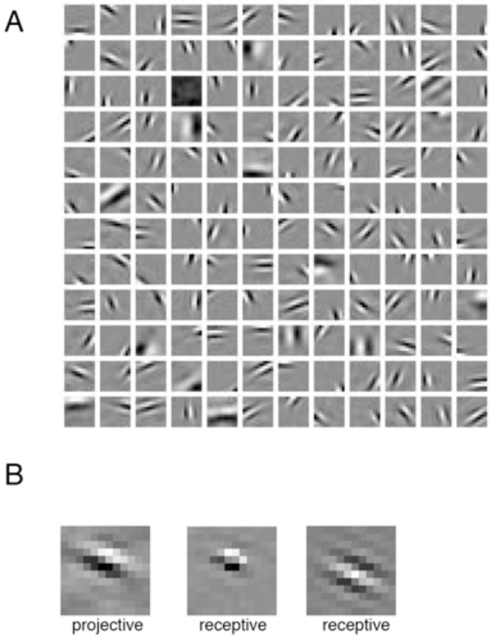


Figure 10.5: A network for sparse coding. This network reproduces equation (10.31) using recurrent weights $-\mathbf{G}^T \cdot \mathbf{G}$ in the $\mathbf{v}$ layer and weights connecting the input units to this layer that are given by the transpose of the matrix $\mathbf{G}$. The reverse connections from the $\mathbf{v}$ layer to the input layer indicate how the mean of the recognition distribution is computed.



The true causes of images are surfaces and objects, not low-level causes such as Gabor patterns. Presumably, the fact that the generative model has only one level biases the results towards Gabor-patterns. Moreover, if one analyzes the natural images in terms of Gabor patterns, one finds that different causes (patterns) are not actually independent. This suggests that low-level causes such as Gabor patterns are in turn caused by higher-level causes (more complex patterns, surfaces, objects, etc).

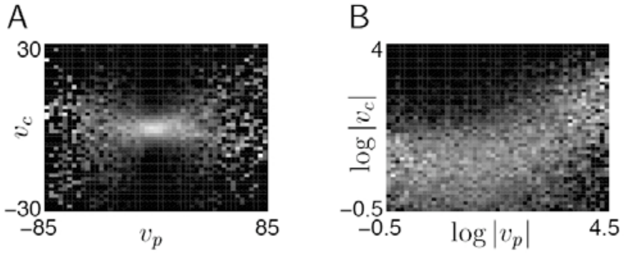


Figure 10.11: A) Gray-scale plot of the conditional distribution of the output of a filter at the finest spatial scale (v_c) given the output of a courser filter (v_p) with the same position and orientation (using the picture in figure 10.9A as input data). Each column is separately normalized. The plot has a characteristic bow-tie shape. B) The same data plotted as the conditional distribution of $\ln|v_c|$ given $\ln|v_p|$. (Adapted from Simoncelli & Adelson, 1990; Simoncelli & Schwartz, 1999.)

7 Neuronal coding

We often characterize neurons in terms of simple, Gabor-like receptive fields. This seems to suggest that neurons decompose natural scenes in terms of Gabor-like 'causes'. However, a close look shows that neuronal responses to natural scenes are not at all well explained by Gabor-like receptive fields. (Their responses to laboratory scenes such as bars and gratings are explained much better!). Apparently, neurons even in early visual areas represent more complex causes than Gabor-patterns.

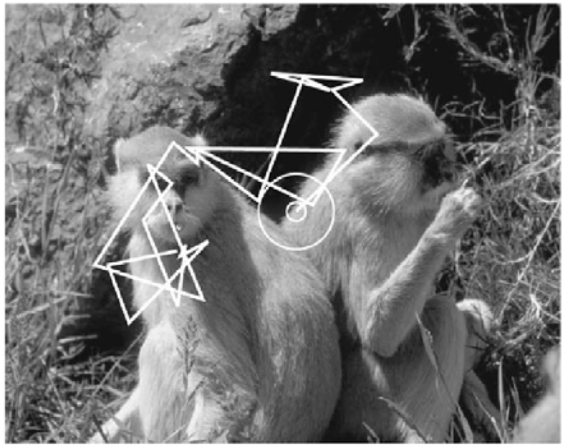


Figure 1. Natural-vision movies reproduce the stimulation that occurs during free viewing of natural scenes. To construct a natural vision movie, a saccadic scan path (*white line*) is generated using a model derived from previously recorded eye movements. Image patches centered on the scan path coordinates (*white circles*) are then extracted from the underlying image. Image patches were from one to four times the size of the CRF. (The *small circle* indicates 1 × CRF diameter, whereas the *large circle* indicates 4 × CRF diameter.) Note that although nCRF stimulation varied substantially with stimulus size, the stimulus falling on the CRF was the same for all sizes.

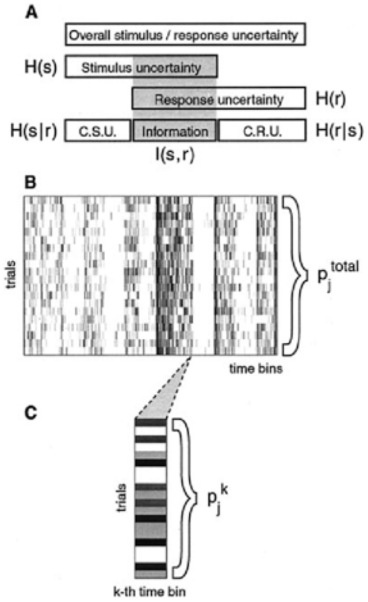
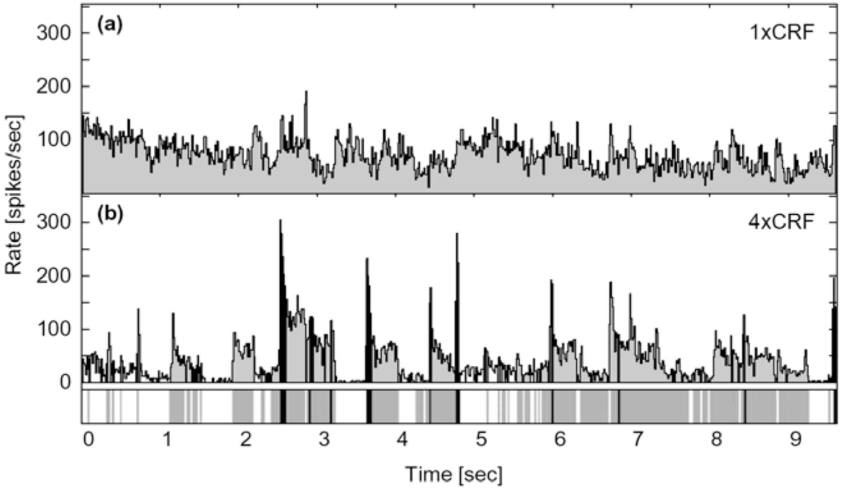


Figure 2. Sensory neurons transmit information when their responses allow an observer to reduce uncertainty regarding the nature of the stimulus. *A*, Diagram illustrating relationship between uncertainty and information. The *first rectangle* symbolizes the total uncertainty present in the set of all stimulus–response pairings for a given neuron; the *second rectangle* represents the observer's *a priori* uncertainty about the stimuli in a natural-vision movie; the *third rectangle* represents the uncertainty in the observed responses of the neuron. These uncertainties can be translated into entropies by means of Equation 5. The single number that summarizes overall stimulus uncertainty is the total stimulus entropy, $H(s)$, while the total response entropy is $H(r)$. The remaining *rectangles* are the conditional stimulus uncertainty (*C.S.U.*) and the conditional response uncertainty (*C.R.U.*) (quantified by the entropies $H(s|r)$ and $H(r|s)$, respectively). The *gray-shaded region* denotes correlations between the stimulus and the responses of the neuron; this correlation is what allows information, $I(s, r)$, to be transmitted. If every stimulus evokes a unique and repeatable response, then response uncertainty will be entirely determined by stimulus uncertainty. In this case the gray-shaded region would completely overlap both stimulus and response uncertainties. In real neurons, repeated presentation of a stimulus produced a range of responses, so $H(r) > I(s, r)$. The remaining uncertainty, $H(r|s)$, is attributable to noise in the encoding and transmission process. *B*, Grayscale rastergram of single neuron responses to repeated movie presentations. *Rows* represent repeated presentations of the movie, whereas *columns* represent individual time bins. Each time bin contains a single response word whose identity is determined by the number of action potentials (identity is indicated by the shading of each bin). The total response entropy, $H(r)$, is a function of the frequency with which each word is observed, p_j^{total}. *C*, Magnified view of responses to one stimulus repeated 20 times. Variation in the identity of the response words is clearly visible across trials and is quantified as noise entropy, $H(r|s = k)$. Noise entropy is a function of the probability that each word occurs in response to the *k*th stimulus, p_j^k.

context sparsifies responses in V1



Responses to natural scenes. Context in natural scenes sparsifies responses of V1 neurons. Shown is the average response of a neuron to multiple repetitions of a natural vision movie played just within the receptive field of the neuron (top) or the same movie but with additional spatial context extending into the receptive field surround (bottom). Context appears to make the neuron more selective to certain episodes within the movie sequence. Taken from [42••], with permission. Copyright 2002 by the Society for Neuroscience.