Bubbles: A unifying Framework for Low-Level Statistical Properties of Natural Image Sequences. Hyvarinen et al, J. Opt. Soc. Am. 2003

Richard Turner (turner@gatsby.ucl.ac.uk)

Gatsby Computational Neuroscience Unit, 03/03/2006

Motivation

- We want good **generative models** of **natural scenes**: $P(y) = \int dx P(y|x)P(x)$
- View **neurons as encoding some aspect of the posterior distribution over latents**: $P(x|y, \theta)$, like the mean, mode or width (inference)
- View **learning** and **adaptation** as learning the parameters of this generative model (by maximum-likelihood for example)
- Why would neurons do this?
  - Extraction of higher order statistical structure in the inputs into latent variables - **causes** - forms a **computationally useful representation**
  - Good generative models lead to **efficient codes** (Barlow)...

Goal of the paper

- Produce a generative model for **movies** with a prior over latent variable that combines three ideas:
  - **Sparse coding** (shown to produce simple cell like RFs)
  - **Topographic spatial dependencies** (computationally useful)
  - **Temporal slowness** (also shown to produce simple cell like RFs)

- **Proof of concept** rather than a finished piece of work

- Paper answers question: Why do both slowness and sparseness both produce simple cells?

- Because the latents are slow and sparse and one or both of these criteria can be used to infer them.

Sparse Coding

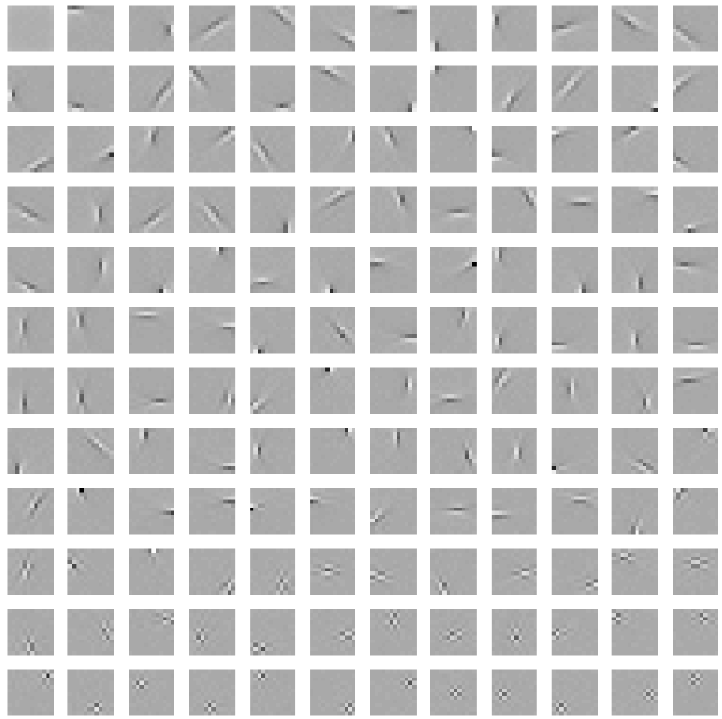
- Two motivations:
  - Redundancy reduction (Bell and Sejnowski, 1996; Olshausen and Field, 1996)
  - Latent variable modeling (Pearlmutter, 1999 and Mackay, 1999)

ICA Generative model	**Sparse Coding Generative model**
$P(\mathbf{x}) = \prod_{i=1}^{I} P(x_i)$ (1)	$P(\mathbf{x}) = \prod_{i=1}^{I} P(x_i)$ (4)
$P(x_i) = \text{sparse}$ (2)	$P(x_i) = \text{sparse}$ (5)
$P(\mathbf{y}\mid\mathbf{x}) = \delta(G\mathbf{x} - \mathbf{y})$ (3)	$P(\mathbf{y}\mid\mathbf{x}) = \text{Norm}(G\mathbf{x}, \sigma^2 I)$ (6)

- Typical choices for $P(x_i)$ are:
  - $1/$cosh (Bell), Cauchy (Olshausen), Biexponential, Student T (Osindero)

Projective fields: 144, 12 by 12 filters filters from ICA

Generative weights - columns of G - (projective fields) look like Gabors (or multi-scale wavelets)

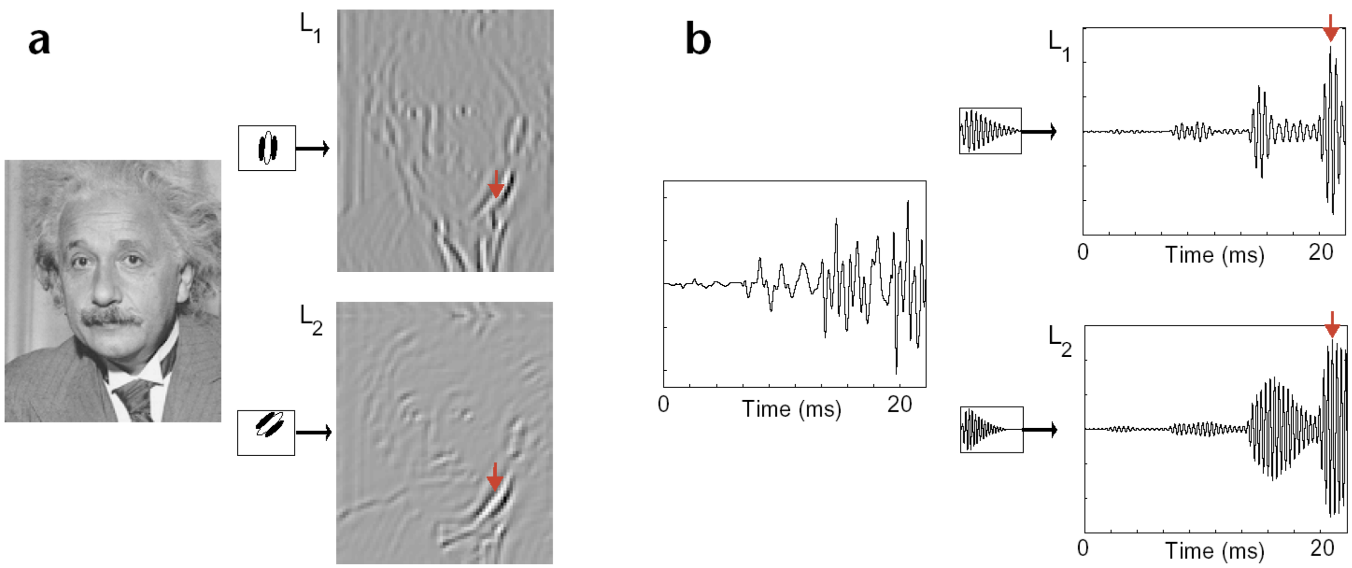


Receptive fields are also Gabor like.

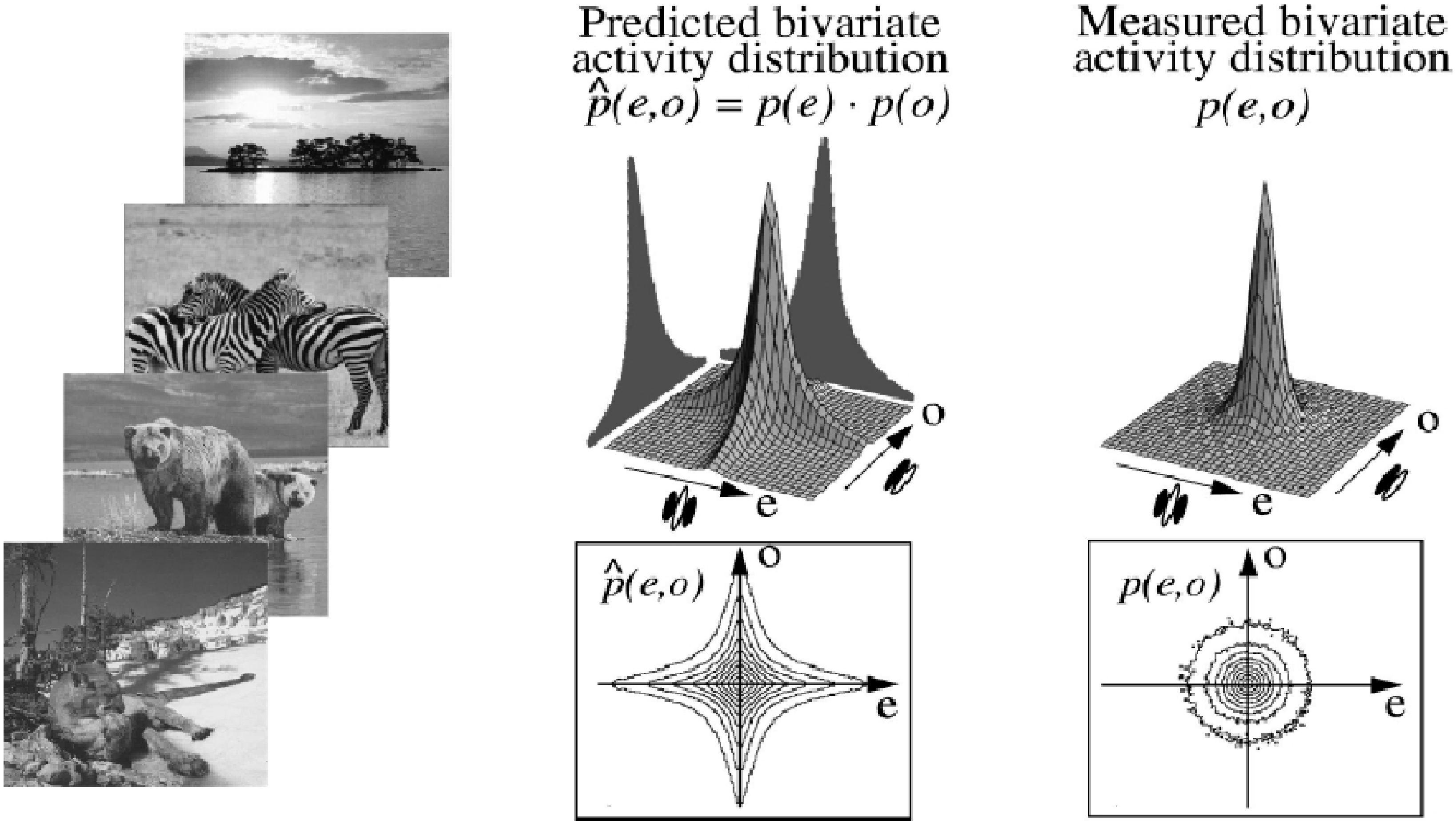
Are the "independent components" really independent?

- Seems unlikely that natural scenes could be explained by such low level causes.

stimulus $\longrightarrow$ 2 linear filters $\longrightarrow$ output
Stimulus: *Image or sound*
Filter pair: *Steerable pyramid shifted and rotated or Gammatone with different carrier frequencies*

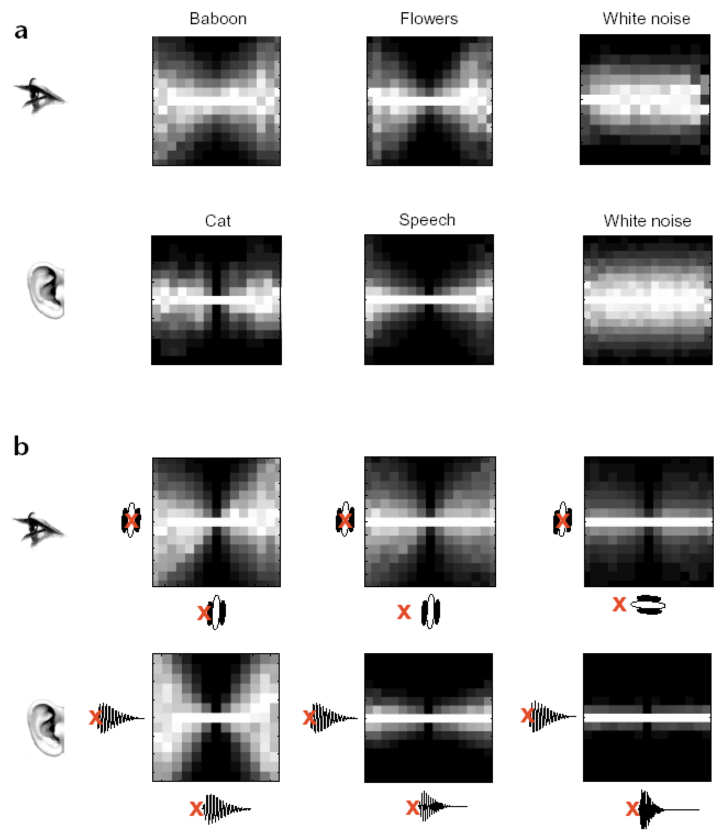


Spatial Statistics of linear filter responses, 2



Spatial Statistics of linear filter responses, 3

Conditional histograms. Vertical cross-sections are not identical. **Top**: Previous filter pairs with different images. The dependency is strong for natural stimuli but weak for white noise. **Bottom**: Fixed stimulus, different filters. The strength of the dependence depends on the filter pair.



How do we improve the model?

- We want to improve our generative model

- But still want the first layer to resemble simple cells (which are quite linear) - so fix the recognition distribution: $P(\mathbf{x}|\mathbf{y}) = \delta(\mathbf{x} - R\mathbf{y})$ (gave ok results for ICA)

- In the complete case, this fixes our generative distribution too: $P(\mathbf{y}|\mathbf{x}) = \delta(R^{-1}\mathbf{x} - \mathbf{y})$

- All we now need is a prior, chosen to match the statistics of images: $P(\mathbf{x}) = \int d\mathbf{y} P(\mathbf{y}) P(\mathbf{x}|\mathbf{y})$

- We have plenty of images (samples from $P(\mathbf{y})$), so we can approximate the integral by sampling over image patches: $P(\mathbf{x}) \simeq \frac{1}{N} \sum_{\mathbf{y}} \delta(\mathbf{x} - R\mathbf{y})$

- What type of distributions would make a suitable prior, that captured the features these histograms?

Gaussian Scale Mixtures (GSMs)

- $\mathbf{x} = \lambda \mathbf{u}$ [cf. product models, eg. Grimes and Rao, 2005]
  - $\lambda \geq 0$ a scalar random variable
  - $\mathbf{u} \sim G(0, Q)$
  - λ and $\mathbf{u}$ are independent

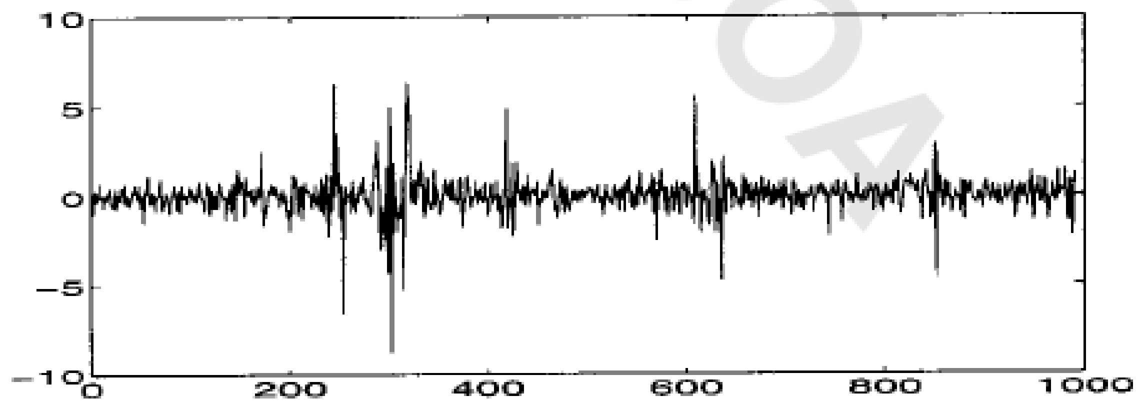
- density of these *semi-parametric* models can be expressed as an integral:

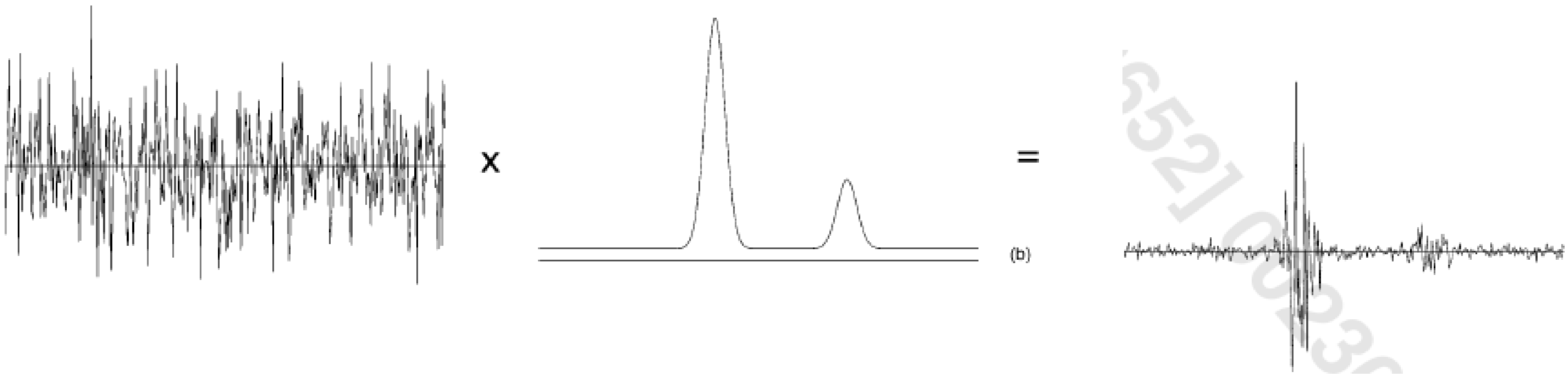
$$P(\mathbf{x}) = \int P(\mathbf{x}|\lambda)P(\lambda)d\lambda = \int |2\pi\lambda^2 Q|^{-1/2} \exp\left(-\frac{\mathbf{x}^T Q^{-1}\mathbf{x}}{2\lambda^2}\right)\psi(\lambda)d\lambda \qquad (7)$$

- One example is the MOG model [$\psi(\lambda)$ is discrete, components all 0 mean]

- Another is $\psi(\lambda) =$Gamma, [marginals are student T distributed]

Temporal Statistics of linear filter responses





Correlations in the energy of the latents through time, and between different latents over space

The bubbles model - a little odd

$$P(u_{i,t}) = \text{sparse} \quad (8)$$

$$\lambda_{i,t} = f[\sum_j h_{i,j}\Phi(t) \otimes u_j(t)] \quad (9)$$

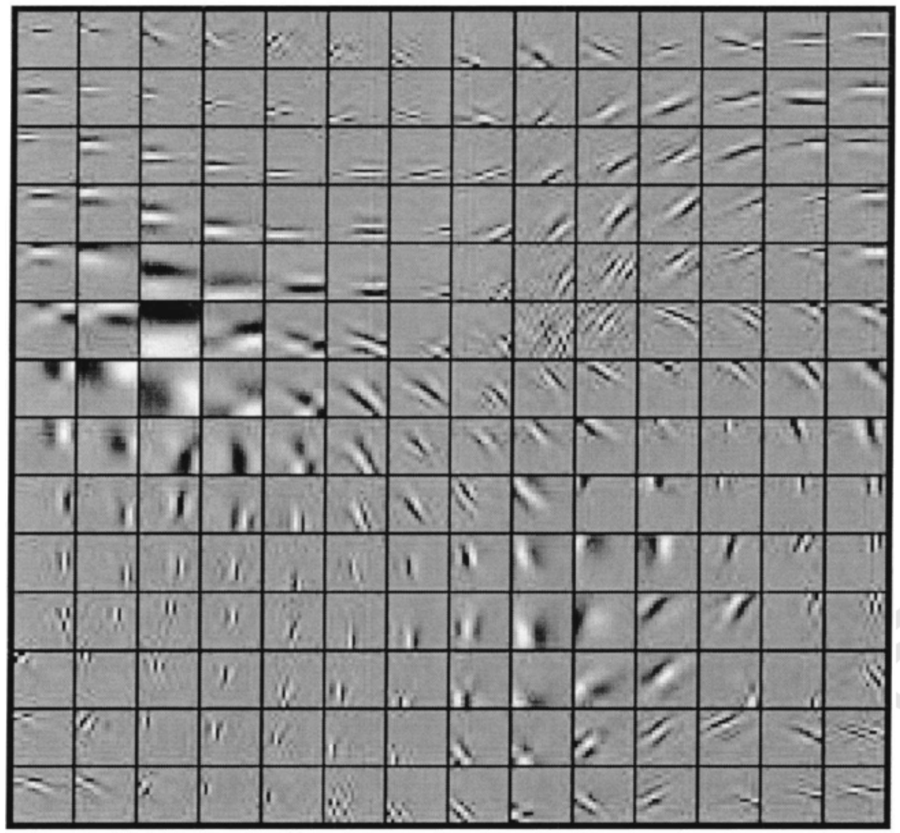
$$P(x_{i,t}|\lambda_{i,t}) = \text{Norm}(0, \lambda_{i,t}^2) \quad (10)$$

$$P(\mathbf{y}_t|G, \mathbf{x}_t) = \delta(G\mathbf{x}_t - \mathbf{y}_t) \quad (11)$$

- **Temporal correlations** between the multipliers are captured by the moving average $\Phi(t) \otimes u_j(t)$ (temporal smoothing)
- **Statistical dependencies** between filters depends on their separation (in space, scale, and orientation.): $h_{i,j}$
- Columns of $h_{i,j}$ fixed and change smoothly to induce **topographic structure** - **computationally useful**
- $\Rightarrow$ **Bubbles** of activity in latent space (both in space and time)

Results

- Learn filters using the **likelihood as a guide** to the sorts of terms you want in a cost function
- Orientation and location of generative weights change smoothly
- Low frequency patches segregate



Ideas for Future work

- **Tidy up the generative model** to have a more regular time series prior
- **Improve learning**
- Learn the neighbourhoods H (perhaps with some soft topological prior), and the temporal smoothing
- Investigate $\arg\max_\lambda P(\lambda|\mathbf{y})$ as complex-cell output - cf. **energy detector models**