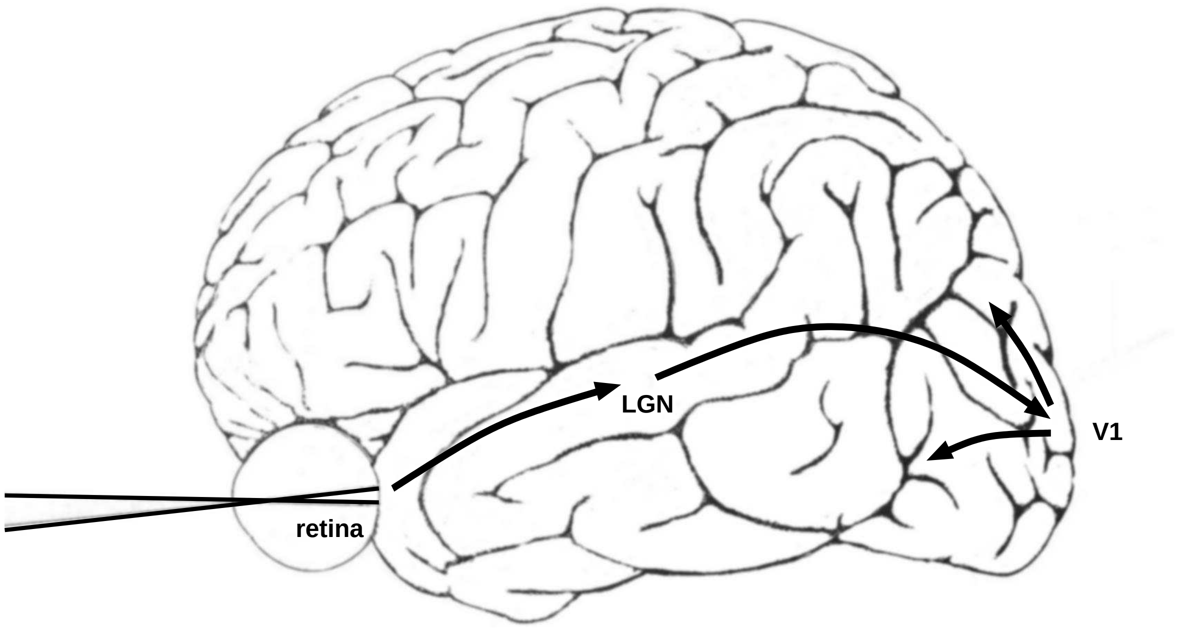
Natural Image Statistics

— A probabilistic approach to modelling early visual processing in the cortex

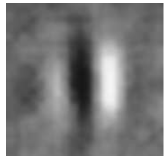
Aapo Hyvärinen
Dept of Computer Science
University of Helsinki

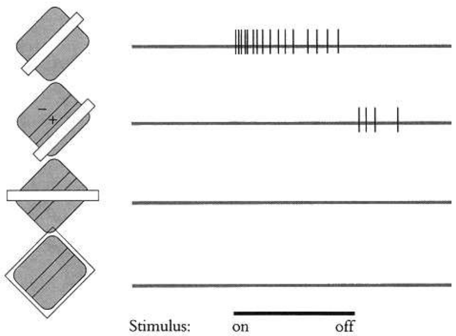
Early visual processing



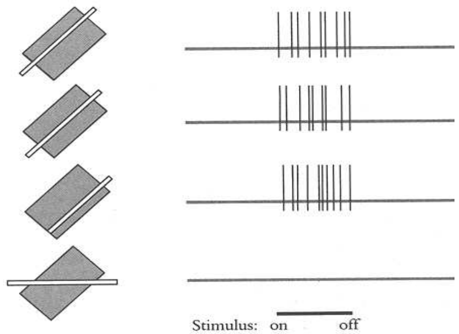
From the eye to the primary visual cortex (V1)

Simple and complex cells

- Basic dichotomy of neurons in primary visual cortex (V1)
- Simple cells modelled as linear functions of input: "Receptive fields" can be simply plotted 
- Complex cells considered strongly nonlinear
  - Invariance (tolerant) to location (phase) of input
  - Modelled as sum of squares of simple cell outputs



Simple cells:
Selective to orientation
and location of the bar



Complex cells:
Tolerant to exact location

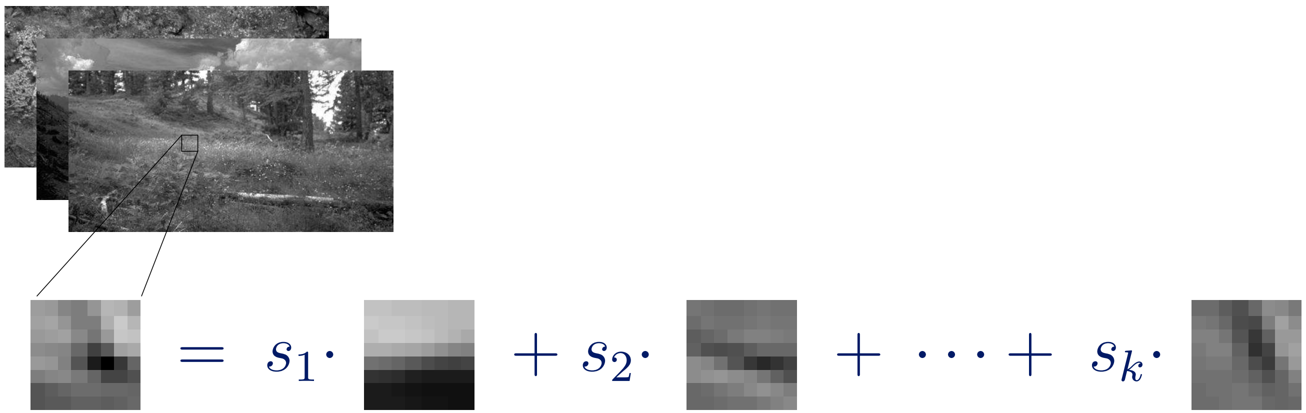
Theories of response properties of visual neurons

- Edge detection
- Joint localization in space and frequency
- Texture classification
- But: the above give only vague predictions.
- Here: Statistical-ecological approach (Barlow, 1972)
  - What is important in a real environment?
  - Natural images have statistical regularities.
  - Can we "explain" receptive fields by basic statistical properties of natural images?
  - Emergence : a lot of precise predictions from only a couple statistical assumptions.
- Extremely relevant to image processing /engineering

Outline of this talk:

- Statistical models that account for some properties of the (primary) visual cortex.
  - Independent Component Analysis / Sparse Coding
  - Various extensions
- Properties in visual cortex explained
  - simple cells
  - complex cells
  - spatial organization (topography)
- Multi-layer approach can predict properties beyond V1.

Linear statistical models of images



- Denote by $I(x, y)$ the gray-scale values of pixels.
- Model as a linear sum of basis vectors:

$$I(x, y) = \sum_i A_i(x, y) s_i \tag{1}$$

- What are the "best" basis vectors for natural images?

Independent Component Analysis (Jutten and Hérault, 1991)

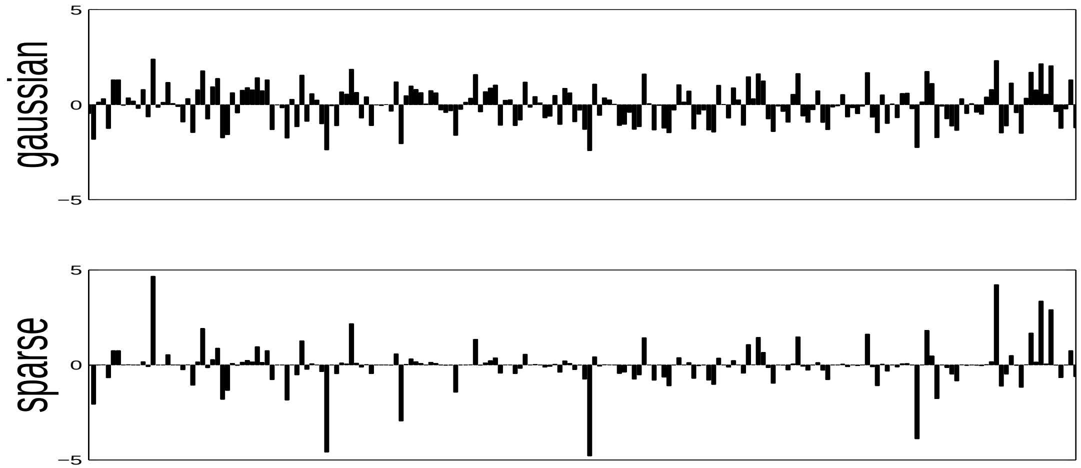
- Linear model:

$$I(x,y) = \sum_i A_i(x,y) s_i \tag{2}$$

- In ICA, we assume that
  - The s_i are mutually statistically independent
  - The s_i are nongaussian, e.g. sparse
  - For simplicity: number of A_i equals number of pixels
- Then, the actual basis vectors A_i can be estimated, if the data is actually generated using the linear model (Comon, 1994).
- Thus we get the best basis vectors from one statistical viewpoint.

Sparsity

- A form of nongaussianity often encountered in natural signals
- A random variable is "active" only rarely



- Outputs of linear filters are usually sparse when input is natural images.

Sparse coding and ICA

- Sparse coding (Barlow 1972): Find linear representation

$$I(x, y) = \sum_i A_i(x, y) s_i \tag{3}$$

  so that the s_i are as sparse as possible.

- Important property: a given data point is represented using only a limited number of "active" (clearly non-zero) components s_i.

- In contrast to PCA, active components change from image patch to patch.

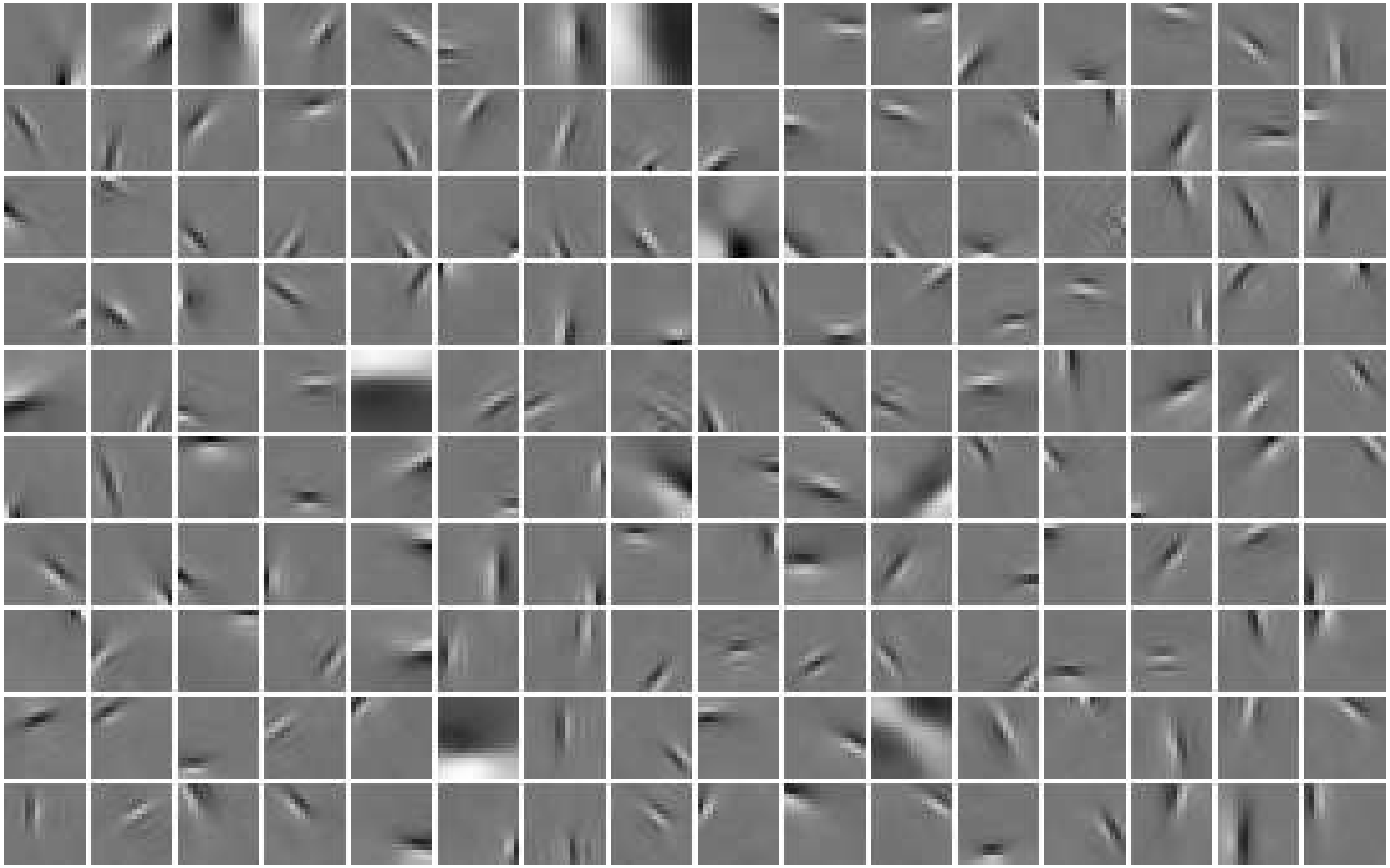
- Deep result: For images, ICA is sparse coding.

- Vectorizing whitened image as $\mathbf{x}$, denoting inverse system by $\mathbf{w}_i$:

$$\min_{\text{orthog } \mathbf{w}_1, \ldots, \mathbf{w}_n} \hat{E}\{|\mathbf{w}_i^T \mathbf{x}|\} \tag{4}$$

ICA / sparse coding of natural images

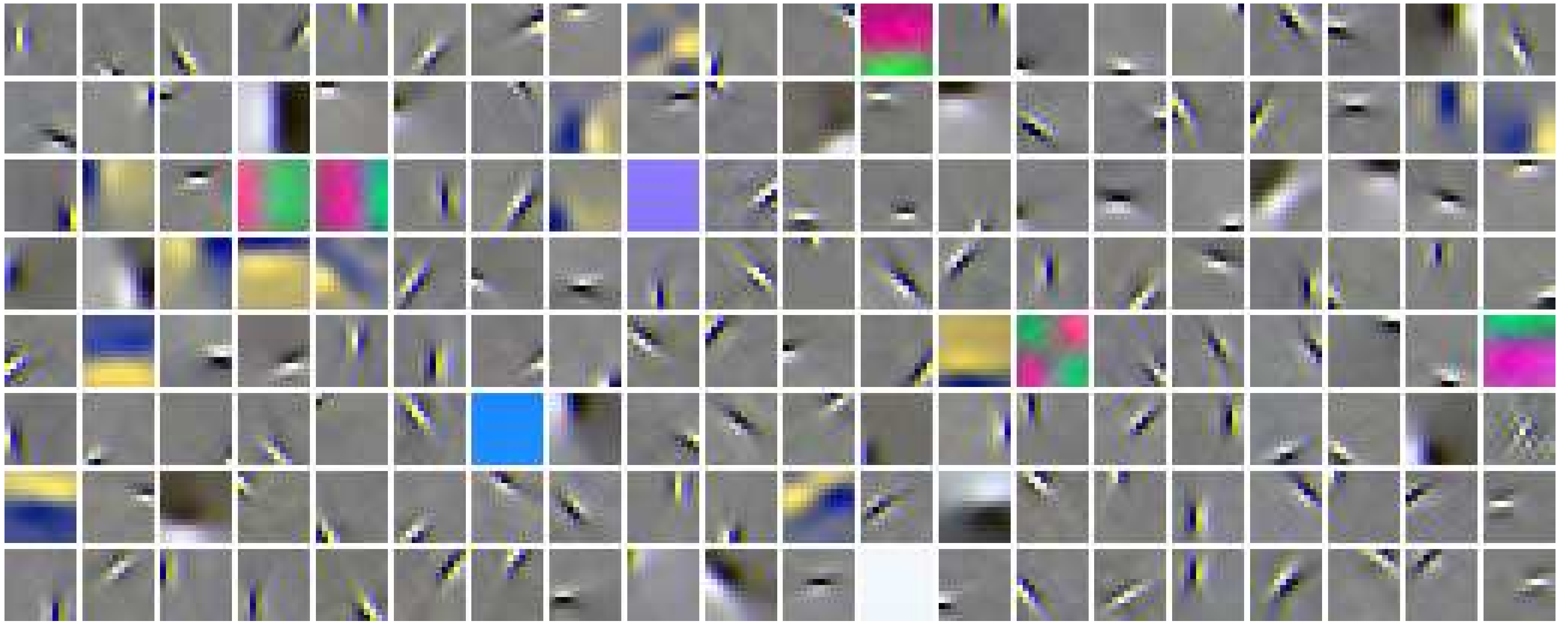
(Olshausen and Field, 1996; Bell and Sejnowski, 1997)



Using the FastICA algorithm (Hyvärinen, 1999)

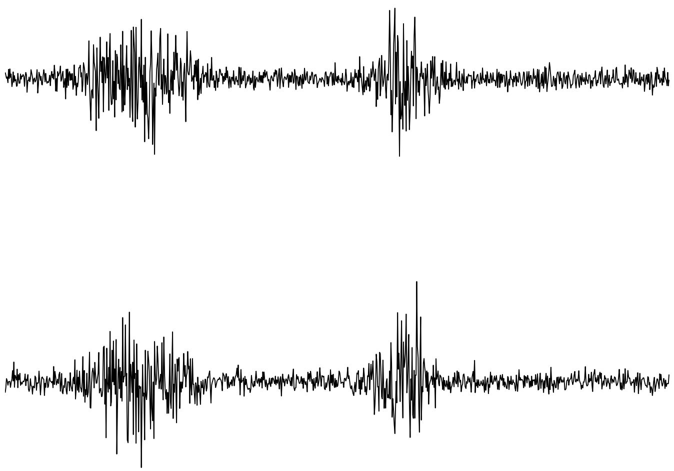
ICA of natural images with colour

(Hoyer and Hyvärinen, 2000)



Model II: Independent subspace analysis

- Components estimated from natural images are not really independent.
- The statistical structure much more complicated (of course!).
- Independent components cannot be found for most kinds of data: There are not enough free parameters.
- Dominant form of dependency after ICA is correlation of energies



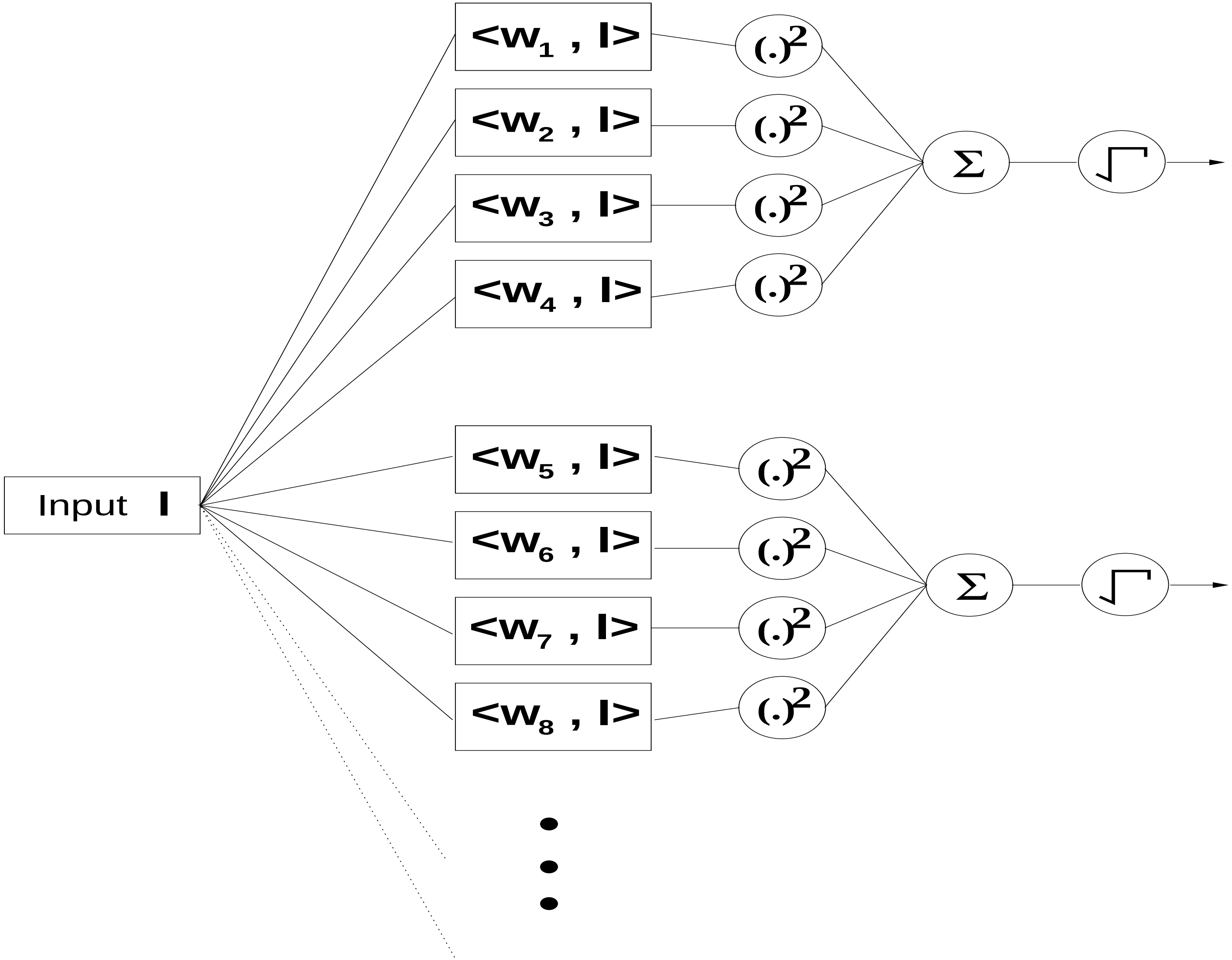
Signals which are uncorrelated but whose squares are correlated.

Using subspaces to model dependency

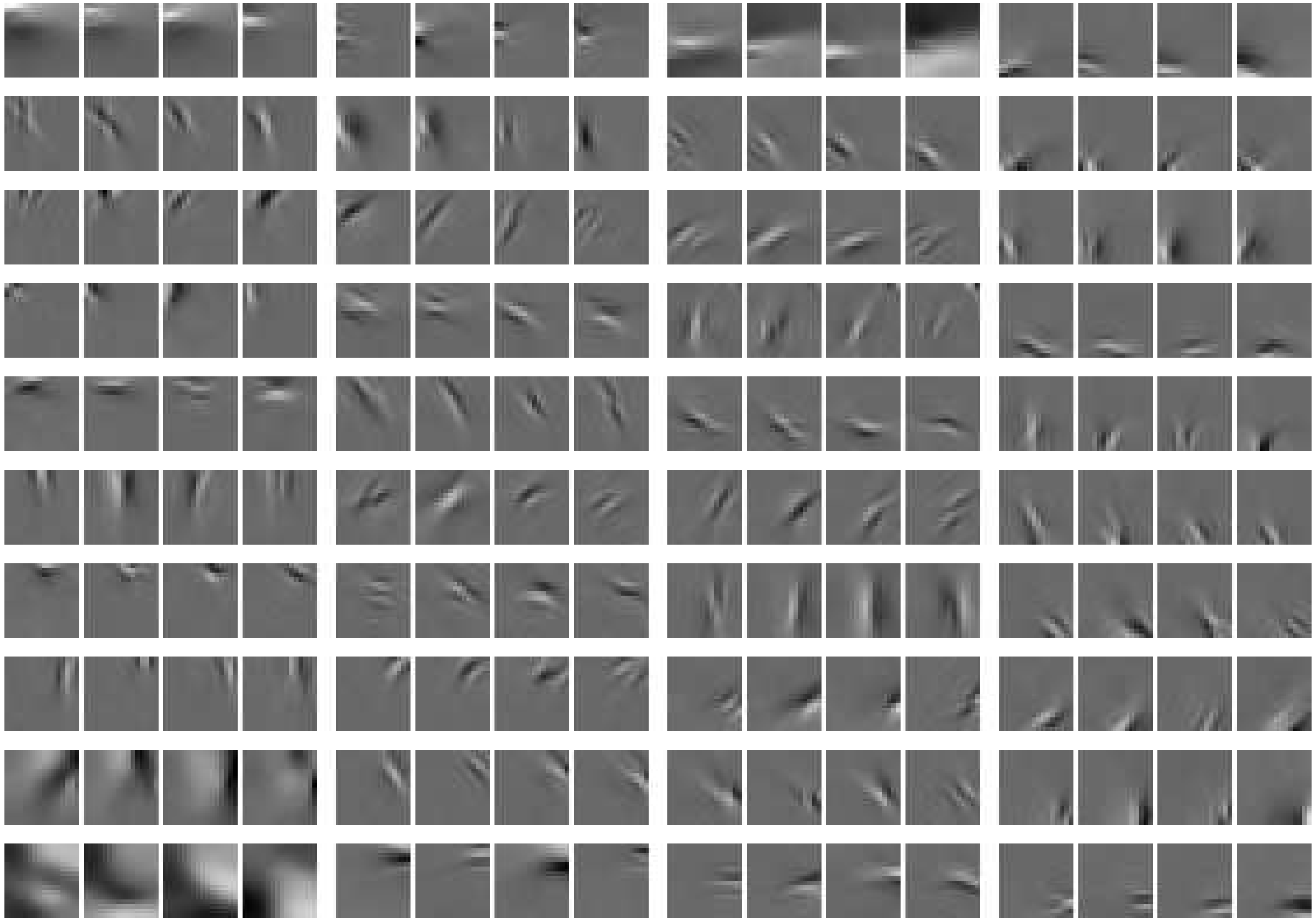
(Hyvärinen and Hoyer, 2000)

- Assumption: the s_i can be divided into groups or subspaces (Cardoso, 1998), such that
  - the s_i in the same group are dependent on each other
  - dependencies between different groups are not allowed.
- We also need to specify the distributions inside the groups
  - Use classic complex cell model, norm of projection in subspace
  - Leads to correlation of squares
  - Maximize independence / sparsity of complex cell output

Computation of features in independent subspace analysis



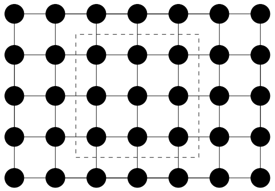
Independent subspaces of natural image patches



Each group of 4 basis vectors corresponds to one complex cell.

Model III: Spatial organization in V1

- In the brain, response properties mostly change continuously when moving on the cortical surface.
- We introduced Topographic ICA (Hyvärinen and Hoyer, 2001)
- Cells (components) are arranged on a two-dimensional lattice

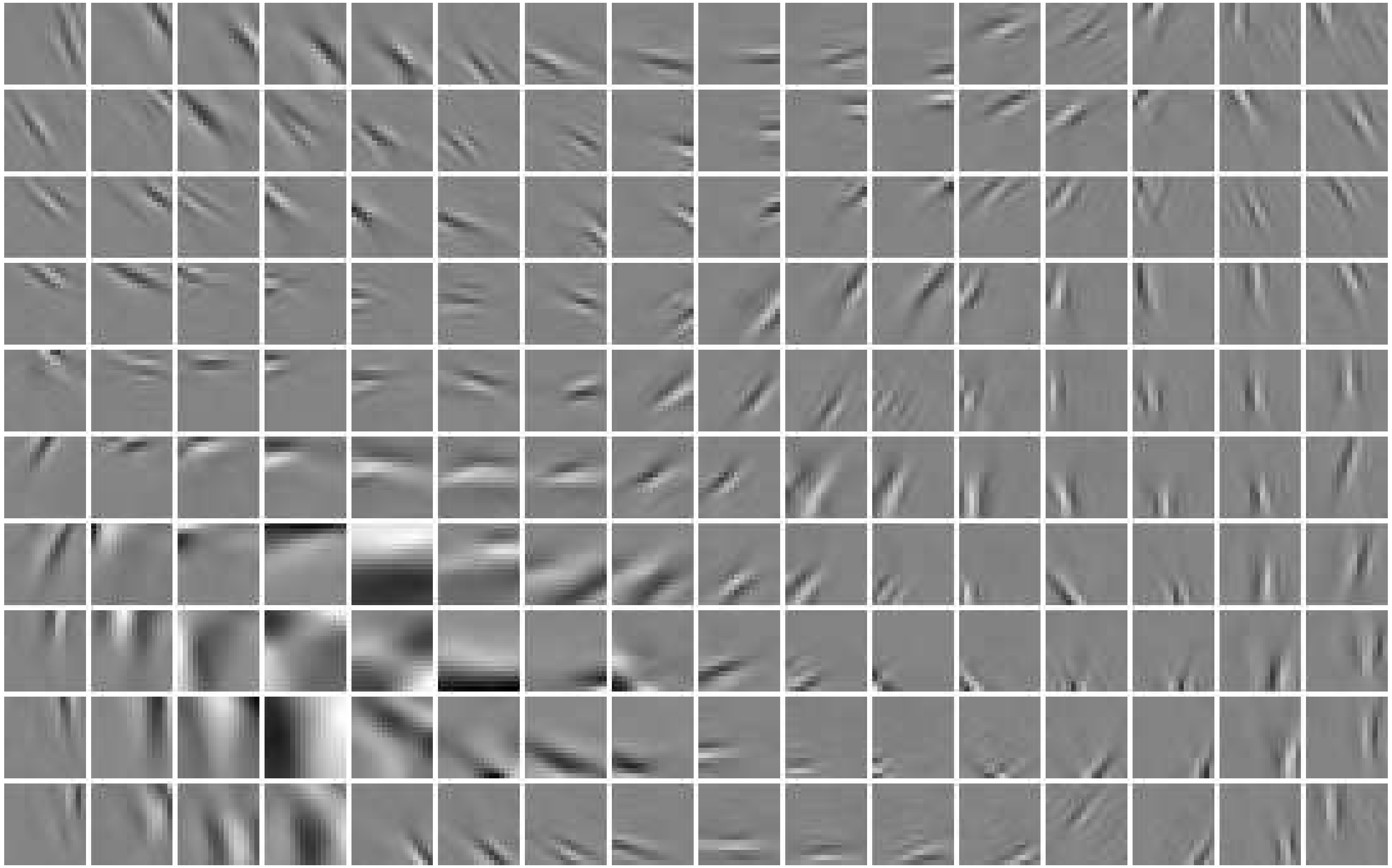


- Again, simple cell outputs are sparse, but not independent: Correlations of squares follows topography.
- Learn by maximizing likelihood which measures sparsity:

$$-\log p(\mathbf{x}|\mathbf{W}) = \sum_i \hat{E}\left\{\sqrt{\sum_j h_{ij}(\mathbf{w}_i^T\mathbf{x})^2}\right\} \tag{5}$$

where h is distance on topographic grid.

Topographic ICA on natural image patches



Basic vectors (simple cell RF's) with spatial organization

Digression: Generalizations leading to unnormalized models

- Generalize ISA and topographic ICA by estimating two layers

$$p(\mathbf{x}; \mathbf{W}, \mathbf{M}) = \frac{1}{Z(\mathbf{W}, \mathbf{M})} \exp[\sum_i G(\sum_i m_{ij} (\mathbf{w}_j^T \mathbf{x})^2)] \quad (6)$$

- This is an unnormalized model:
  Density function p is known only up to a multiplicative constant

$$Z(\mathbf{W}, \mathbf{M}) = \int \exp[\sum_i G(\sum_i m_{ij} (\mathbf{w}_j^T \mathbf{x})^2)] d\mathbf{x}$$

  which cannot be computed with reasonable computing time

- Common problem in non-Gaussian models, e.g. ICA with more components than observed variables

$$p(\mathbf{x}; \mathbf{W}) = \frac{1}{Z(\mathbf{W})} \exp[-\sum_i |\mathbf{w}_i^T \mathbf{x}|] \quad (7)$$

Recent methods for unnormalized statistical models

- Score matching (Hyvärinen, JMLR, 2005)
  - Take derivative of model log-density w.r.t. $\mathbf{x}$, so partition function disappears
  - Fit this derivative to the same derivative of data density
  - Easy to compute due to a integration-by-parts trick
- Noise-contrastive estimation
  (Gutmann and Hyvärinen, JMLR, 2012)
  - Learn to distinguish data from artificially generated noise
  - Logistic regression learns ratios of pdf's of data and noise
  - For known noise pdf, we have in fact learnt data pdf

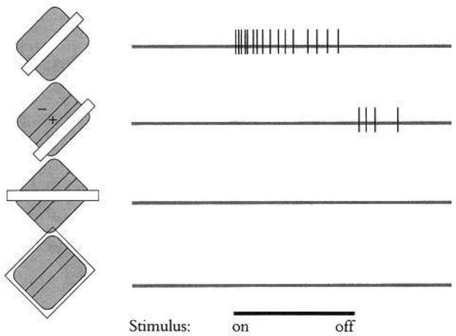
Another unnormalized model:
Linear correlations between components

- Many methods force components to be strictly uncorrelated
- So, any remaining linear correlations cannot be observed
- However, the “true” components could be correlated even linearly
- In ongoing work (Sasaki et al, 2013, 2014) we learn correlated components

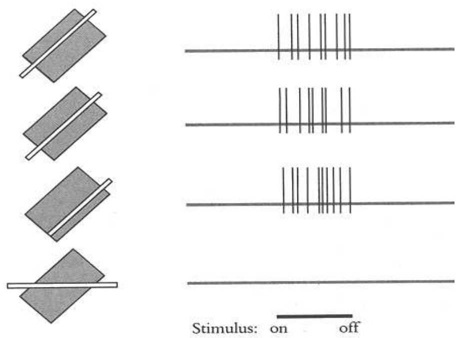
$$-\log p(\mathbf{x}|\mathbf{W}) = \sum_i \hat{E}\{|\mathbf{w}_i^T\mathbf{x}|\} + \sum_{i,j} \beta_{ij}\hat{E}\{|\mathbf{w}_i^T\mathbf{x} - \mathbf{w}_j^T\mathbf{x}|\} - \log Z(\beta_{ij}) \tag{8}$$

Model IV: Three-layer model

- Goal: Alternating selectivity and invariance in many layers
- Neurons are *selective* to certain properties of the stimulus:
  - response is strong when those properties take specific values
- Neurons are *tolerant* (invariant) to properties:
  - response does not change much when those properties change
- Example: Simple vs. complex cells in the primary visual cortex:



Simple cells:
Selective to orientation
and location of the bar



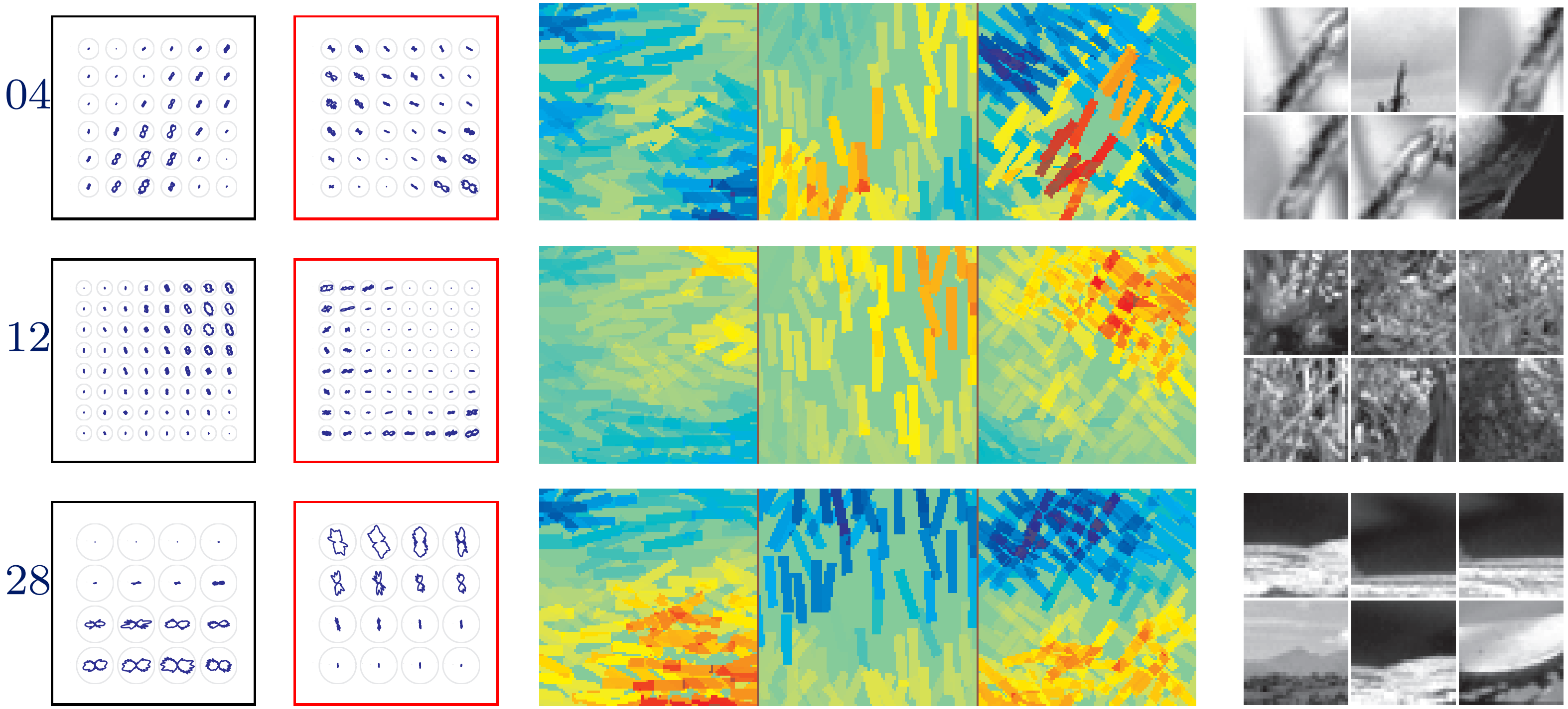
Complex cells:
Tolerant to exact location

Nonlinearities for neural selectivity and tolerance

- Selectivity has been modelled as
  - AND operation / MIN operation
- Tolerance (invariance) has been modelled as
  - OR operation / MAX operation
- Here, we use **convex** and **concave** nonlinearites
  - E.g. $(\sum_i x_i^4)^{1/4}$: similar to MAX
  - E.g. $(\sum_i x_i^{1/4})^4$: similar to MIN
- Two first layers similar to squaring + square root in complex cells
- Learning by maximization of sparsity in each layer

Layer three responses in three-layer model

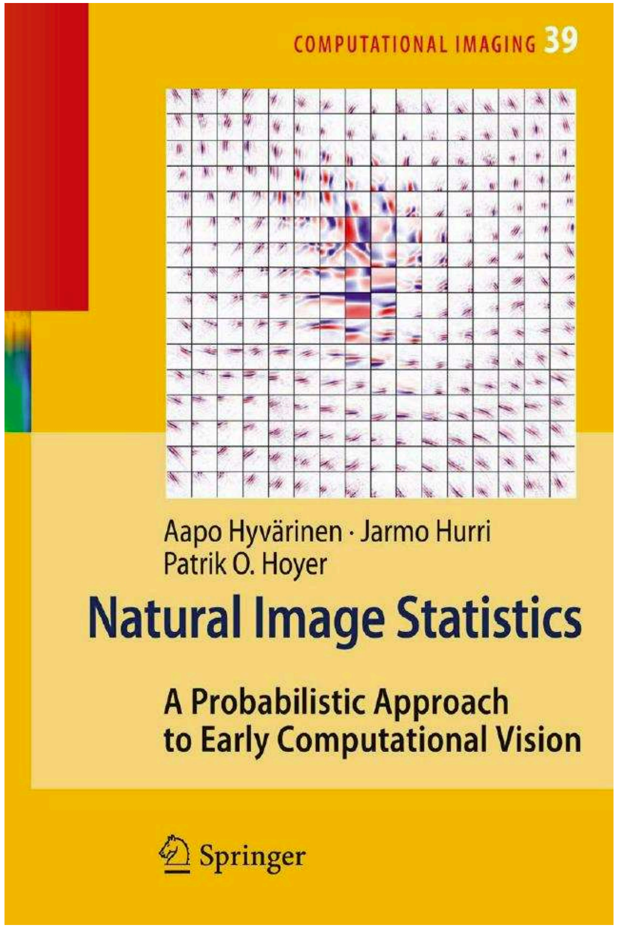
Receptive field (RF)	Inhibitory RF	Optimal 2nd layer activity (horizontal, vertical, diagonal)	Strongly activating images



Neurophysiological modelling vs. Deep learning

- Deep learning means neural network with many layers
- Result is often a black box: interpretation difficult
- For neurophysiological modelling, we would prefer a network where
  - The role of each unit is clear
  - All cell responses model biological responses
- Instead of blindly stacking many layers on top of each other, we must think about what each layer is doing

For more information on basic models



Conclusion

- Properties of visual neurons can be quantitatively modelled by statistical properties of natural images.
- Answers the Why question important in neuroscience
- Simple cell receptive fields can be learned by maximizing sparsity / independence of linear filters.
- By modelling dependencies between simple cell ouputs we can model complex cells and topography
- Three-layer models can use alternating selectivity and invariance
- Theoretical development on estimation of unnormalized models
- Many applications in image processing and computer vision