Higher Order Statistics

Matthias Hennig

Neural Information Processing
School of Informatics, University of Edinburgh

February 12, 2018

1

[0] Based on Mark van Rossum's and Chris Williams's old NIP slides
[1] version: February 12, 2018

Outline

- First, second and higher-order statistics
- Generative models, recognition models
- Sparse Coding
- Independent Components Analysis
- Convolutional Coding (temporal and spatio-temporal signals)

Redundancy Reduction

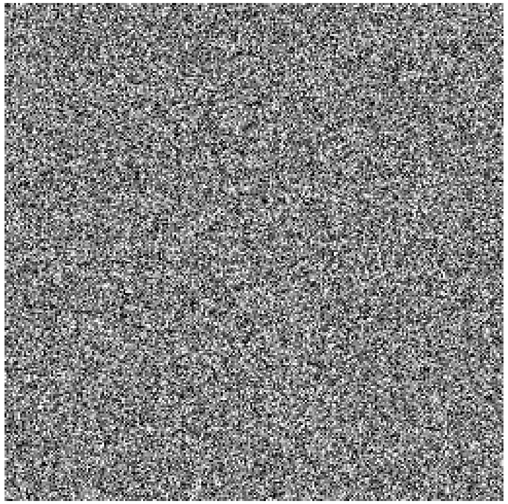
(Barlow, 1961; Attneave 1954)

- 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

Natural Image Statistics and Efficient Coding

- First-order statistics
  - Intensity/contrast histograms ⇒ e.g. histogram equalization
- Second-order statistics
  - Autocorrelation function ($1/f^2$ power spectrum)
  - Decorrelation/whitening
- Higher-order statistics
  - orientation, phase spectrum
  - Projection pursuit/sparse coding

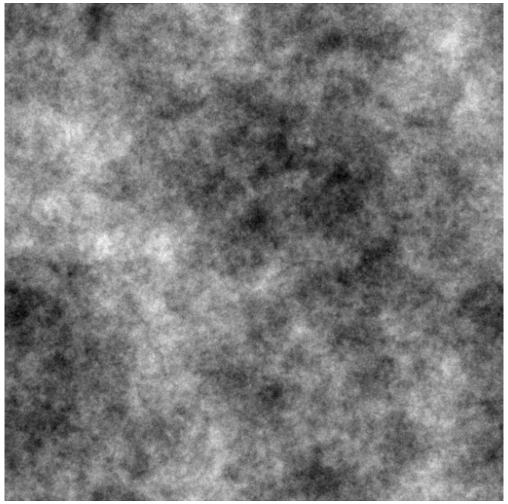
Image synthesis: First-order statistics



[Figure: Olshausen, 2005]

Log-normal distribution of intensities

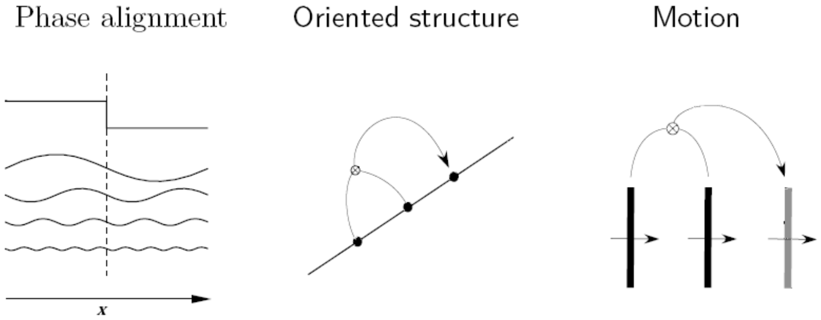
Image synthesis: Second-order statistics



[Figure: Olshausen, 2005]

Describe as correlated Gaussian statistics, or equivalently, power spectrum

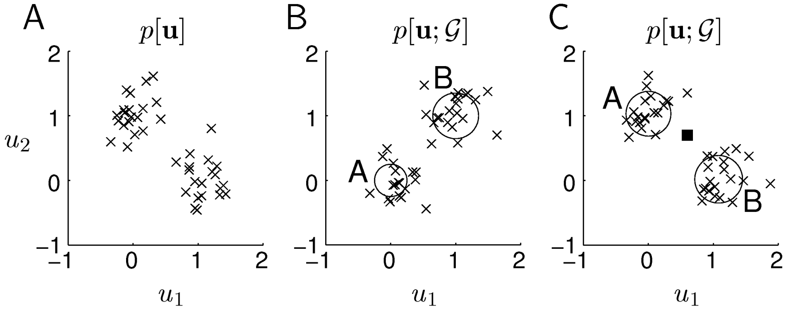
Higher-order statistics



[Figure: Olshausen, 2005]

Generative models, recognition models

(§10.1, Dayan and Abbott)



Left: observations. Middle: prior. Right: good model
In image processing one would want, e.g. A are cars, B are faces.
They would explain the image.

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)

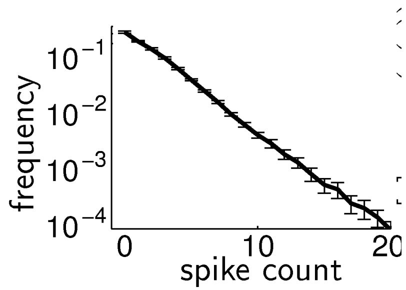
Examples of generative models

(§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 ≈ 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)

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

- 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

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: $g(h) = -\log(\beta^2 + h^2)$
- $\mathbf{n} \sim N(\mathbf{0}, \sigma^2 I)$
- 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 $g(h)$ corresponding to the Cauchy distribution, $p(\mathbf{h}|\mathbf{u}, \mathcal{G})$ is difficult to compute exactly
- Olshausen and Field (1996) used MAP approximation

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

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

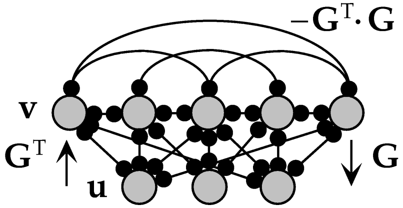
$$\sum_{b=1}^{N_h} \frac{1}{\sigma^2}[\mathbf{u} - G\hat{\mathbf{h}}]_b G_{ba} + g'(\hat{h}_a) = 0$$

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

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, $\mathbf{v} = \mathbf{h}$)



$-\mathbf{G}^{\mathrm{T}}\cdot\mathbf{G}$
v
$\mathbf{G}^{\mathrm{T}}$
u
G

Figure: Dayan and Abbott, 2001]

- Dynamics does gradient ascent on log posterior.
- Note inhibitory lateral term
- Process is guaranteed only to find a local (not global) maximum

Learning of the model

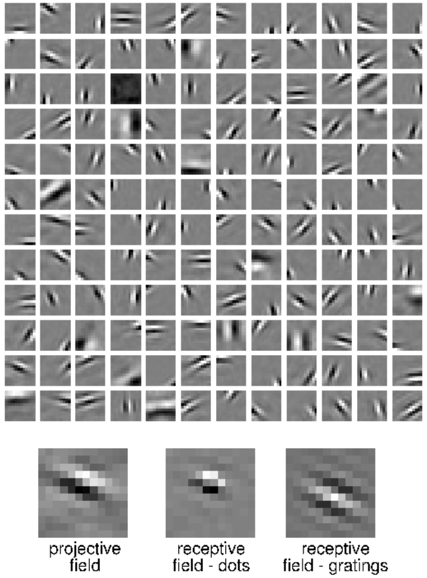
- Now we have $\mathbf{h}$, we can compare
- Log likelihood $L(\mathcal{G}) = \langle \log p(\mathbf{u}|\mathcal{G}) \rangle$. Learning rule:

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

- Basically linear regression (mean-square error cost)

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

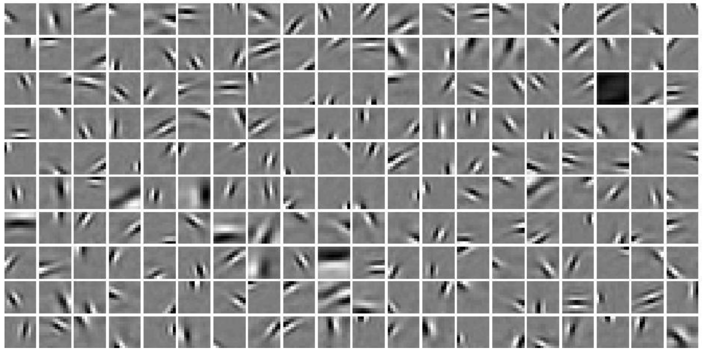
- Small values of h can be balanced by scaling up G. Hence impose constraint on $\sum_b G_{ba}^2$ for each cause a to encourage the variances of each h_a to be approximately equal
- It is common to whiten the inputs before learning (so that $\langle \mathbf{u} \rangle = \mathbf{0}$ and $\langle \mathbf{u}\mathbf{u}^T \rangle = I$), 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 × 12 patches [Figure: Olshausen, 2005]

Gabor functions

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

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

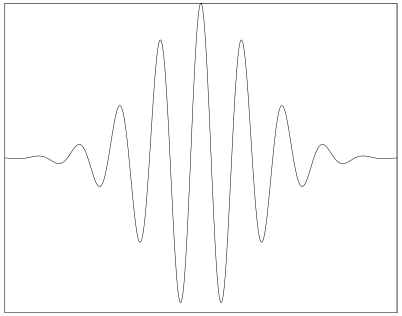
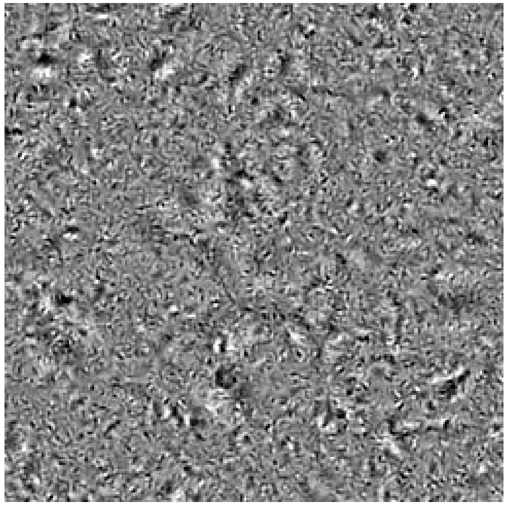


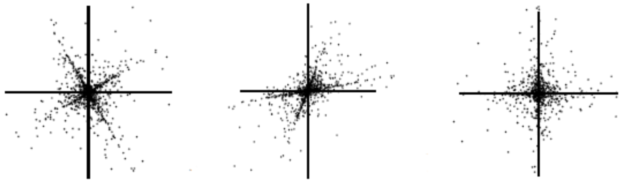
Image synthesis: sparse coding



[Figure: Olshausen, 2005]

ICA: Independent Components Analysis

- $H(h_1, h_2) = H(h_1) + H(h_2) - I(h_1, h_2)$
- Maximal entropy typically if $I(h_1, h_2) = 0$, i.e. $P(h_1, h_2) = P(h_1)P(h_2)$
- The more random variables are added, the more Gaussian. So 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)



ICA derivation, [Bell and Sejnowski, 1995]

- Linear network with output non-linearity $\mathbf{v} = W\mathbf{u}$, $y_j = f(h_j)$.
- Find weight matrix maximizing information between $\mathbf{u}$ and $\mathbf{y}$
- No noise (cf. Linsker), so $I(\mathbf{u}, \mathbf{y}) = H(y) - H(y|u) = H(y)$
  $H(y) = \langle \log p(\mathbf{y}) \rangle_y = \langle \log p(\mathbf{u}) / \det J \rangle_u$ with
  $J_{ji} = \frac{\partial y_j}{\partial u_i} = \frac{\partial h_j}{\partial u_i} \frac{\partial y_j}{\partial h_j} = w_{ij} \prod_j f'(h_j)$
  $H(y) = \log \det W + \langle \sum_j logf'(h_j) \rangle + const$
- 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. $1/(1 + \mathrm{e}^{-h})$.
- $\det W$ helps to insure independent components
- For $f(h) = 1/(1 + \mathrm{e}^{-h})$, $dH(y)/dW = (W^T)^{-1} + (\mathbf{1} - \mathbf{2y})\mathbf{x}^\mathbf{T}$

ICA: Independent Components Analysis

Derivation as generative model

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

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

  note Jacobian term
- 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

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

- For image patches, again Gabor-like RFs are obtained
- In the ICA case PFs and RFs can be readily computed

Beyond Patches

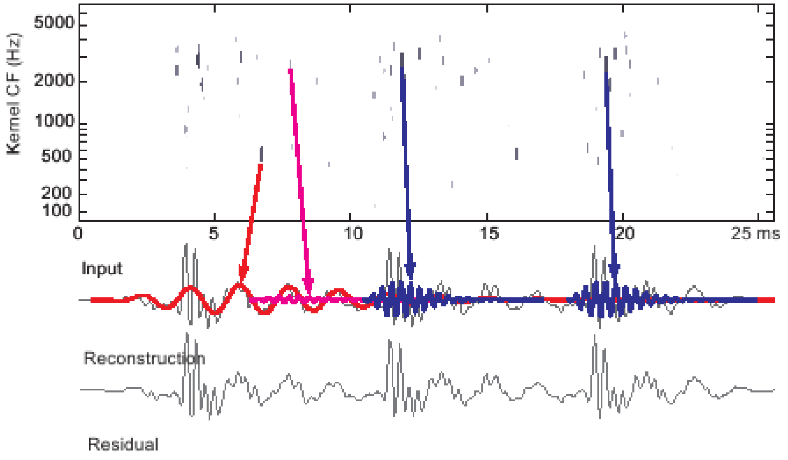
“Convolutional Coding” (Smith and Lewicki, 2005)

- For a time series, we don’t want to chop the signal up into arbitrary-length blocks and code those separately. Use the model

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

- τ_i^m and h_i^m are the temporal position and coefficient of the ith instance of basis function g_m
- Notice this basis is M-times *overcomplete*

- Want a *sparse* representation
- A signal is represented in terms of a set of discrete temporal events called a *spike code*, displayed as a *spikegram*
- Smith and Lewicki (2005) use matching pursuit (Mallat and Zhang, 1993) for inference
- Basis functions are gammatones (gamma modulated sinusoids), but can also be learned
- Zeiler et al (2010) use a similar idea to decompose images into sparse layers of feature activations. They used a Laplace prior on the h’s.



[Figure: Smith and Lewicki, NIPS 2004]

Spatio-temporal sparse coding

(Olshausen 2002)

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

- 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.
- `animate -resize 783x393 bfmovie.gif`
  (200 bases, 12 × 12 × 7)
  http://redwood.berkeley.edu/bruno/bfmovie/bfmovie.html

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 algorhithms [Stein & Gerstner, preprint])
- Differences
  - ICA: number of inputs = number of outputs. Sparse Coding: over-complete
  - 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.