Hierarchical Sparse Bayesian Learning

Pierre Garrigues
UC Berkeley
Outline

• Motivation
• Description of the model
• Inference
• Learning
• Results
Motivation

• **Assumption**: sensory systems are adapted to the statistical properties of their inputs

• Our ability to extract statistical regularities of natural images help us perform complex visual tasks

• Building a better *statistical model of natural images* will help us improve algorithms for image processing
Sparse Coding Model

• Generative model [Olshausen, Field 96]:

\[ p(s) \propto e^{-|s|} \]

\[ x = As + \epsilon, \text{ where } A \in \mathbb{R}^{n \times m} \]

\[ \epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2 I_n) \]

• The model allows A to be overcomplete
Independent Component Analysis

- Find a linear transform such that the outputs are independent and have sparse distributions [Bell & Sejnowski 97]

\[ p(x) = |W| \prod_{i=1}^{n} q(w_i^T x) \]

\[ q(y) = \begin{cases} 
\frac{1}{2} e^{-|y|} & \text{Laplacian distribution} \\
\frac{1}{\pi(1+y^2)} & \text{Cauchy distribution}
\end{cases} \]
Learned transform

The learned filters resemble wavelets
Caveats of these models

• The independence assumption is violated for natural images

• The coefficients associated with quadrature pair or colinear Gabor filters are not independent

• The visual system probably makes use of these dependencies (e.g. for contour extraction)
Modeling the remaining dependencies

- **Existing work**
  - Gaussian Scale Mixtures [Wainwright & Simoncelli 01]
  - Density Components Models [Karklin & Lewicki 03]
  - Markov Random Fields [Hinton et al. 05]

- **Our Model:**
  - extends K&L to overcomplete setting
  - draws a connection with Sparse Bayesian Learning [Tipping 01]
Hierarchical Sparse Bayesian Learning

- We want to discover structure in the $\gamma$ pattern.
- Statistical model:
  - $x = As + n$, where $n \sim \mathcal{N}(0, \sigma^2)$
  - $s_i \sim \mathcal{N}(0, \gamma_i)$ for every $i$
  - $\gamma_i = \psi([Bv]_i)$
  - $d < m$
  - $d \geq m$ and sparse prior on $v$

$p(v) \propto e^{-|v|}$

Pierre Garrigues 4 July 28, 2006
Choice of the nonlinearity

\[ e^x \cdot 1[x < 0] + (1+x) \cdot 1[x \geq 0] \]

\[ e^x \cdot 1[x < 0] + (1+x+0.5x^2) \cdot 1[x \geq 0] \]
Intuition for B

• Our goal is to model the joint dependencies of the basis functions

\[ Bv = \sum_{i=1}^{d} v_i \begin{pmatrix} B_{1i} \\ \vdots \\ B_{mi} \end{pmatrix} \]

density component

• The relative signs within a density component model the excitation and inhibition
Inference of $v$

- As in SBL, we use the EM algorithm
  
  $$\hat{v} = \arg \max_v p(v|x) = \arg \max_v p(x|v)p(v)$$

- Expectation Step
  
  $$q(s|x, v^{(k)}) \sim \mathcal{N}(\mu, \Sigma)$$
  
  $$\left\{ \begin{array}{l} 
  \Sigma = (\sigma^{-2}A^TA + \Gamma^{-1})^{-1}, \\
  \Gamma = \text{diag}(\psi([Bv^{(k)}]_1), \ldots, \psi([Bv^{(k)}]_m)) \\
  \mu = \sigma^{-2}\Sigma A^T x \end{array} \right.$$ 

- Maximization Step
  
  $$v^{(k+1)} = \arg \max_v \mathbb{E}_{s \sim q}[\log p(x, s|v) + \log p(v)]$$
  
  $$= \arg \min_v \sum_{i=1}^{m} \left( \frac{1}{2} \log \psi([Bv]_i) + \frac{\mathbb{E}_{s \sim q}[s_i^2]}{2\psi([Bv]_i)} - \log p(v_i) \right)$$
Learning of $B$

- **MAP estimate**

- **Approximation** of the objective function

\[
p(x|B) = \int p(x, s, v|B) dsdv
= \int p(x|s)p(s|v, B)p(v) dsdv
\approx p(x|\hat{s})p(\hat{s}|\hat{v}, B)p(\hat{v})
\]

\[
\hat{v} = \arg\max p(v|x)
\]

\[
\hat{s} = \mathbb{E}[s|x, \hat{v}]
\]
Learning of B

• **MAP estimate** \( \hat{B} = \arg \min_{B} \sum_{i=1}^{N} - \log p(x^{(i)}|B) - \log p(B) \)

• **Approximation of the objective function**

\[
p(x|B) = \int p(x, s, v|B)dsdv
\]

\[
= \int p(x|s)p(s|v, B)p(v)dsdv
\]

\[
\simeq p(x|\hat{s})p(\hat{s}|\hat{v}, B)p(\hat{v})
\]

\( \hat{v} = \arg \max p(v|x) \)

\( \hat{s} = \mathbb{E}[s|x, \hat{v}] \)
Learning rule

• New objective function:

\[ \hat{B} = \arg \min_B \sum_{i=1}^{N} - \log p(\hat{s}^{(i)}|\hat{v}^{(i)}, B) = \arg \min_B J(B) \]

• Learning rule:

\[ B^{\text{new}} = B^{\text{old}} - \eta \nabla J(B) \]

\[ \frac{\partial J(B)}{\partial B_{ij}} = \frac{1}{2} \hat{v}_j \frac{\psi'([v]_i)}{\psi([Bv]_i)} \left( 1 - \frac{\hat{s}_i^2}{\psi([Bv]_i)} \right) + \frac{1}{2} B_{ij} \]
Results

- Settings:
  - $n = m = d = 144$
  - about 1000 iterations
  - The matrix $A$ was learned using ICA
Learned density components

They are hard to visualize!
Visualization w.r.t. spatial position of the basis functions
Visualization w.r.t. position in the Fourier domain
Sparsity of the coefficients

![Graph showing sparsity of coefficients with different methods: learned B, random B, SBL, and sparsenet. The x-axis represents the coefficient index, ranging from -50 to 50, and the y-axis represents the coefficient magnitude on a logarithmic scale, ranging from $10^{-1}$ to $10^7$. Each method is represented by a different line color or style, illustrating the distribution and sparsity of the coefficients.]
Sparsity index distribution

\[ \frac{\sqrt{m} - \|s\|_1}{\sqrt{m} - 1} \]
Conclusion

• We were able to reproduce similar results as K&L in the overcomplete setting

• Future work
  • results preliminary, still issues
  • denoising results
  • texture classification
  • MRF model
MRF model

Binary MRF

\[ s_i \mid u_i = 1 \sim \mathcal{N}(0, \sigma_i^2) \]
\[ s_i \mid u_i = 0 \sim \delta(s_i) \]

\[ x = As + \epsilon, \text{ where } A \in \mathbb{R}^{n \times m} \]

Apply similar algorithm as in [Hinton et al. 05]
Variance and mean for HSBL with learned B

Variance and mean of the density components