On Sparsity and Overcompleteness in Image Models

Richard Turner, Pietro Berkes, and Maneesh Sahani
{turner,berkes,maneesh}@gatsby.ucl.ac.uk
Gatsby Computational Neuroscience Unit, UCL, London

April 23, 2008
Motivation

Natural Scenes
Motivation
Motivation
Motivation
Two simple questions

How **overcomplete** should the dictionary be?

How **sparse** should the primitives be?
Trade-off between over-completeness and sparseness
Small dictionary, general purpose primitives, not sparse
Large dictionary, specific primitives, very sparse
Conventional wisdom and approach

Very over-complete and very sparse dictionaries are best

Evaluation via comparison of representations to V1
Take home message

Very over-complete and very sparse dictionaries are best

Evaluation via comparison of representations to V1

Modestly over-complete but very sparse dictionaries are best

Evaluation via probabilistic methods
Sparse coding model (or ICA)

\[
p(x_k | \alpha) = \text{Sparse}(\alpha)
\]

\[
p(y | x) = \text{Norm} \left( \sum_{k=1}^{K} w_k x_k, \sigma_y^2 I \right)
\]

- How sparse? What’s \( \alpha \)?
- What is the best family of distributions? Compare \( \text{Sparse}(\alpha) \)’s
- How overcomplete? How large is \( K \) versus \( \text{dim}(y) = D \)?
- Is there a trade-off between sparseness and over-completeness? \( \alpha \) versus \( \frac{K}{D} \)
The holy grail: $p(Y|K, \alpha)$
The holy grail: \( p(Y|K, \alpha) \)
The holy grail: $p(Y|K, \alpha)$

marginal likelihood contours

overcompleteness vs. sparsity

All Noise  Models equiprobable
The holy grail: $p(Y|K, \alpha)$

marginal likelihood contours

overcompleteness

sparsity

All Noise

Models equiprobable
General Approach

• **Hybrid approach** as speed and memory requirements are key
  – Lots of big image patches, highly over-complete models

• Use lots of **toy problems** where we know ground truth to check algorithms

• Use **multiple algorithms** and cross check results
Annealed importance sampling

- One of the few sampling methods for marginal likelihood evaluation
- Requires a (quick mixing) sampler (Gibbs)
- Anneal prior to the posterior
- Importance weights are collected at each temperature
- **Cons:** Too slow to tile the whole space
- **Pros:** Quite quick & reliable integrate over the latents, \( p(y|\theta) = \int dx p(y, x|\theta) \)
Variational Free-Energy Methods

• Returns a lower bound on the log-(marginal)likelihood

\[ F(q(z), \theta) = \log p(y|\theta) - KL(q(z)||p(z|y)) \]

• Factored approximation tractable and very like Gibbs sampling

• **Pros:** Although 10-100 times faster than Gibbs ...

• **Cons:** Biased when learning the sparsity
Variational Methods are Biased

\[ p(s_k) = \frac{1}{2} \]
\[ p(x_k | s_k) = \text{Norm} \left( 0, (1 - s_k)\sigma^2 + s_k(2 - \sigma^2) \right) \]
\[ \text{var}(x_k) = 1, \quad K(x) = 3(1 - \sigma^2)^2 \]
Variational Methods are Biased

\[ p(s_k) = \frac{1}{2} \]

\[ p(x_k | s_k) = \text{Norm} \left(0, (1 - s_k)\sigma^2 + s_k(2 - \sigma^2)\right) \]

\[ \text{var}(x_k) = 1, \quad K(x) = 3(1 - \sigma^2)^2 \]

\[ \sigma^2 = 1 \quad K = 0 \]
Variational Methods are Biased

\[ p(s_k) = \frac{1}{2} \]

\[ p(x_k | s_k) = \text{Norm} \left( 0, (1 - s_k)\sigma^2 + s_k(2 - \sigma^2) \right) \]

\[ \text{var}(x_k) = 1, \quad K(x) = 3(1 - \sigma^2)^2 \]
Variational Methods are Biased

\[ p(s_k) = \frac{1}{2} \]

\[ p(x_k|s_k) = \text{Norm}(0, (1 - s_k)\sigma^2 + s_k(2 - \sigma^2)) \]

\[ \text{var}(x_k) = 1, \quad K(x) = 3(1 - \sigma^2)^2 \]

\( \sigma^2 = 1 \quad K = 0 \)

\( \sigma^2 = 0.125 \quad K = 2.3 \)

\( \sigma^2 = 0.01 \quad K = 2.9 \)
Variational Methods are Biased

\[ p(s_k) = \frac{1}{2} \]

\[ p(x_k | s_k) = \text{Norm}(0, (1 - s_k)\sigma^2 + s_k(2 - \sigma^2)) \]

\[ \text{var}(x_k) = 1, \quad K(x) = 3(1 - \sigma^2)^2 \]
Variational Methods are Biased

\[ p(s_k) = \frac{1}{2} \]
\[ p(x_k|s_k) = \text{Norm}\left(0, (1 - s_k)\sigma^2 + s_k(2 - \sigma^2)\right) \]
\[ \text{var}(x_k) = 1, \quad K(x) = 3(1 - \sigma^2)^2 \]
\[ p(y|x_1, x_2) = \text{Norm}(x_1 + x_2, \sigma_y^2) \]
Variational Methods are Biased

\[
p(s_k) = \frac{1}{2}
\]
\[
p(x_k | s_k) = \text{Norm} \left( 0, (1 - s_k)\sigma^2 + s_k(2 - \sigma^2) \right)
\]
\[
\text{var}(x_k) = 1, \quad K(x) = 3(1 - \sigma^2)^2
\]
\[
p(y | x_1, x_2) = \text{Norm}(x_1 + x_2, \sigma_y^2)
\]
Variational Methods are Biased

\[ p(s_k) = \frac{1}{2} \]
\[ p(x_k | s_k) = \text{Norm} \left( 0, (1 - s_k)\sigma^2 + s_k(2 - \sigma^2) \right) \]
\[ \text{var}(x_k) = 1, \quad K(x) = 3(1 - \sigma^2)^2 \]
\[ p(y | x_1, x_2) = \text{Norm}(x_1 + x_2, \sigma_y^2) \]
High kurtosis $\implies$ $\text{KL}(q(x)\|p(x))$ large

\[
\begin{align*}
p(s_k) &= \frac{1}{2} \\
p(x_k|s_k) &= \text{Norm}(0, (1 - s_k)\sigma^2 + s_k(2 - \sigma^2)) \\
\text{var}(x_k) &= 1, \quad K(x) = 3(1 - \sigma^2)^2 \\
p(y|x_1, x_2) &= \text{Norm}(x_1 + x_2, \sigma_y^2)
\end{align*}
\]
Low kurtosis $\implies$ $\text{KL}(q(x)\|p(x))$ small

\[
p(s_k) = \frac{1}{2}
\]
\[
p(x_k|s_k) = \text{Norm}(0, (1 - s_k)\sigma^2 + s_k(2 - \sigma^2))
\]
\[
\text{var}(x_k) = 1, \quad K(x) = 3(1 - \sigma^2)^2
\]
\[
p(y|x_1, x_2) = \text{Norm}(x_1 + x_2, \sigma_y^2)
\]
Variational Methods are Biased

![Graph showing the relationship between Kurtosis and Likelihood and Free Energy with True Kurtosis marked as a blue dot. The graph demonstrates a decrease in likelihood and free energy as Kurtosis increases.](image-url)
Specific Approach

1. Learn a “good” weights using Variational Bayes with ARD prior
   - finds optimal model size
   - mean of the weights is used for ...

![Graph showing ARD sparsity and overcompleteness](image)
Specific Approach

1. Learn a “good” weights using Variational Bayes with ARD prior
   • finds optimal model size
   • mean of the weights is used for ...

2. Cross validation: Evaluate likelihood on held out data using AIS
   • Avoids having to integrate over the weights via sampling
Generic Sparse Coding

\[ p(x_{kn} | \alpha) = \text{studentt} (\alpha) \]

\[ p(y | x, \gamma^y, \{w_k\}_{k=1}^K) = \text{Norm} \left( Wx, \frac{1}{\gamma^y} I \right) \]
Exponential Family Version

\[ p(\gamma_{kn}^{(x)} | \alpha) = \text{Gamma} \left( \frac{1}{2} \alpha x, \frac{2}{\chi^2 \alpha} \right) \]

\[ p(x_{kn} | \gamma_{kn}^{(x)}) = \text{Norm} \left( 0, \frac{1}{\gamma_{kn}^{(x)}} \right) \]

\[ p(y | x, \gamma^y, \{w_k\}_{k=1}^{K}) = \text{Norm} \left( W x, \frac{1}{\gamma^y} I \right) \]
Automatic Relevance Determination Priors

\[ p(\gamma_k^{(w)}|\theta_w, l_w) = \text{Gamma}(\theta_w, l_w) \]

\[ p(w_k|\gamma_k^{(w)}) = \text{Norm}\left(0, \frac{1}{\gamma^{(w)}}I\right) \]

\[ p(\gamma_{kn}^{(x)}|\alpha) = \text{Gamma}\left(\frac{1}{2}\alpha x, \frac{2}{\lambda^2}\alpha\right) \]

\[ p(x_{kn}|\gamma_{kn}^{(x)}) = \text{Norm}\left(0, \frac{1}{\gamma_{kn}^{(x)}}\right) \]

\[ p(y|x, \gamma, \{w_k\}^K_{k=1}) = \text{Norm}\left(Wx, \frac{1}{\gamma_y}I\right) \]
\[ p(\gamma(y)|\theta_y, l_y) = \text{Gamma}(\theta_y, l_y) \]
\[ p(\gamma_k(w)|\theta_w, l_w) = \text{Gamma}(\theta_w, l_w) \]
\[ p(w_k|\gamma_k(w)) = \text{Norm}(0, \frac{1}{\gamma(w)}I) \]
\[ p(\gamma(x)|\alpha) = \text{Gamma}\left(\frac{1}{2}\alpha_x, \frac{2}{\chi^2\alpha}\right) \]
\[ p(x_{kn}|\gamma_{kn}(x)) = \text{Norm}\left(0, \frac{1}{\gamma_{kn}(x)}\right) \]
\[ p(y|x, \gamma_y, \{w_k\}_{k=1}^K) = \text{Norm}\left(Wx, \frac{1}{\gamma_y}I\right) \]
Variational Bayes and Annealed Importance Sampling
Toy model 1: Generative Model, $K=3$, $\alpha = 2.5$
Toy model 1: Free-energy, $K=3$, $\alpha = 2.5$
Toy model 1: Cross-validation Likelihood, $K=3$, $\alpha = 2.5$
Toy model 2: $\alpha = 2.5$
Toy model 2: $\alpha = 2.5$
Toy model 2: $\alpha = 2.5$
Natural data: Free-energy v. Sparsity

-8.8 \times 10^4

free energy (NATS)

\alpha
Natural data: Marginal-likelihood v. Sparsity

\[ \log P(Y) \text{ (NATS)} \]
Natural data: Trade-off

overcompleteness ratio

\[ \alpha \]
Variational EM and Annealed Importance Sampling
Toy model 2: $\alpha = 2.5$
Natural data: Marginal-likelihood v. Sparsity

-8.35
-8.3
-8.25
-8.2
-8.15
-8.1
-8.05
-8
-7.95
-7.9 x 10^4

Log likelihood

Degree of freedom

-8.35
-8.3
-8.25
-8.2
-8.15
-8.1
-8.05
-8
-7.95
-7.9 x 10^4
Natural data: Trade-off

![Trade-off Diagram]

- Overcompleteness vs Degree of Freedom
- Color scale indicates values in $10^4$
Conclusions

• The optimal sparse coding model for natural scenes is
  – Very sparse
  – Only modestly overcomplete
  – Seeger, 2008 JMLR, uses EP and finds under-complete optimal

• Important question: What does the above approach yield if the true model is non-linear/has higher order dependencies?
  – Sample from an overcomplete Gaussian Scale Mixture and see what we get

• Sophisticated probabilistic machinery is required to answer seemingly simple scientific questions