Finding the optimal sparse, overcomplete model for natural images by model selection

Pietro Berkes, Richard Turner and Maneesh Sahani {berkes,turner,maneesh}@gatsby.ucl.ac.uk

Gatsby Computational Neuroscience Unit, Alexandra House, 17 Queen Square, London.

Abstract

Computational models of visual cortex, and in particular those based on sparse coding, have enjoyed much recent attention. Despite this currency, the question of how sparse or how over-complete a sparse representation should be, has gone without principled answer. Here, we use Bayesian model-selection methods to address these questions for a sparse-coding models based on a Student-t prior and on a Gaussian scale mixture model with uniform prior on precision. We find that natural images are indeed best modeled by extremely sparse distributions, although for these priors, the associated optimal basis size is only modestly over-complete (Berkes *et al.,* 2008).

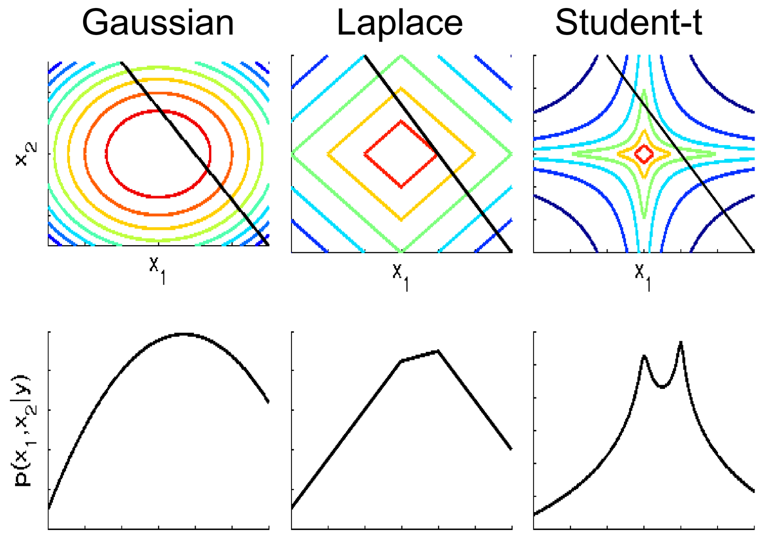
Linear, sparse coding model

$$\mathbf{y}_t = \sum_i \mathbf{g}_i x_{i,t} + \epsilon_t\,, \qquad p(x_{i,t}|\alpha) = p_{\text{sparse}}(\alpha)$$

(Olshausen & Field, 1996, 1997)

Overcomplete case: # latent variables > # input dimensions
An overcomplete 1D model with 2 components:

$$y = g_1 x_1 + g_2 x_2$$



The posterior distribution in very sparse, overcomplete models are complex and multimodal.

Addressed questions

- **How sparse? Which family of distributions?**
- **How overcomplete? (overfitting)** (see also Olshausen and Millman, 2000)
- One might expect a **tradeoff** between sparseness and overcompleteness

Model selection

- One possibility is to implement different models and find the one which is most "similar" to visual processing
- Bayesian perspective:
  - compare marginal likelihood of the model

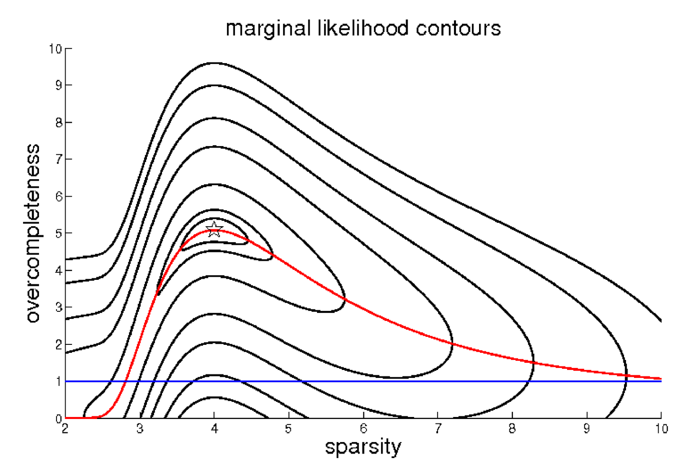
$$\frac{p(\mathcal{M}_1, \Xi_1|Y)}{p(\mathcal{M}_2, \Xi_2|Y)} = \frac{p(Y|\mathcal{M}_1, \Xi_1)\, P(\mathcal{M}_1, \Xi_1)}{p(Y|\mathcal{M}_2, \Xi_2)\, P(\mathcal{M}_2, \Xi_2)} = \frac{P(Y|\mathcal{M}_1, \Xi_1)}{P(Y|\mathcal{M}_2, \Xi_2)}$$

  - automatic Occam's razor
  - natural if hypothesis is that the visual system implements an optimal generative model

References:
P. Berkes, R. Turner, and M. Sahani, On sparsity and overcompleteness in image models. In *Advances in Neural Information Processing Systems*, 20, 2008.
M.J. Beal. *Variational Algorithms for Approximate Bayesian Inference.* PhD thesis, Gatsby Computational Neuroscience Unit, University College London, 2003.
C.M. Bishop. Variational principal components. In *ICANN 1999 Proceedings*, 509–514, 1999.
B.A. Olshausen and D.J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature,* 381(6583):607–609, 1996.
B.A. Olshausen and D.J. Field. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research,* 37:3311–3325, 1997.
B.A. Olshausen and K.J. Millman, Learning sparse codes with a Mixture-of-Gaussians prior. In *Advances in Neural Information Processing Systems,* 12, 2000.
R.M. Neal. Annealed importance sampling. *Statistics and Computing,* 11:125–139, 2001.

Model selection in practice

•Strategy:
- Use Automatic Relevance Determination (ARD) prior, Variational Bayes Expectation Maximization (VBEM) to determine the posterior over parameters and the overcompleteness
- Unfortunately VBEM is biased (cannot use the free energy)
- Compute the likelihood of the learned model using Annealed Importance Sampling (AIS)



ARD (Bishop, 1999; Beal, 2003)

- Gaussian prior over the components that favors small weights; hyperprior over the precisions to keep the prior uninformative

$$p(\mathbf{g}_k|\gamma_k) = \mathcal{N}_{\mathbf{g}_k}(\mathbf{0}, \gamma_k^{-1})\,,$$
$$p(\gamma_k) = \mathcal{G}_{\gamma_k}(\theta_k, l_k)\,.$$

- Start with excess of components, let the inference process prune the weights which are unnecessary
- Learning using VBEM

Why VBEM is biased

Ξ : hyperparameters
V : latent variables
Θ : parameters

$$\log p(Y|\mathcal{M},\Xi) \geq \int \mathrm{d}V \mathrm{d}\Theta\; q(V,\Theta) \log \frac{p(Y,V,\Theta|\mathcal{M},\Xi)}{q(V,\Theta)} =: \mathcal{F}(q(V,\Theta))$$
$$= \log p(Y|\mathcal{M},\Xi) - KL(q(V,\Theta)||p(V,\Theta|Y))$$

The free energy bound is tightest where q(V,Theta) is a good match to the true posterior. At high sparsities, the true posterior is multimodal and highly non-Gaussian. At low sparsities, the true posterior is Gaussian-like and unimodal. q(V,Theta) is always unimodal.
There is an additional bias toward complete solutions, due to the fact that the posterior of an overcomplete solution is more multimodal and thus less Gaussian. This is investigated in a set of simulations with artificial data. The simulations confirm that the bias exist, but the solution found are still considerably more overcomplete than those found in the main simulations.

Annealed Importance Sampling (Neal, 2001)

Idea 1: Simulated annealing

$$p_j(x) = p_0(x)^{\beta_j} p_N(x)^{1-\beta_j} \qquad p_N(x) : \text{prior distribution}$$
$$0 = \beta_N > \ldots > \beta_0 = 1 \qquad p_0(x) : \text{unnormalized posterior distribution}$$

Idea 2: Importance sampling: View annealing as defining an importance sampling distribution over $(x_0, \ldots, x_{N-1})$

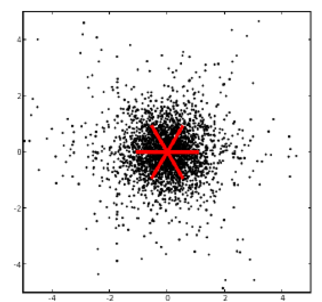
$$p_N \sim x_{N-1} \xrightarrow{T_{N-1}} x_{N-2} \xrightarrow{T_{N-2}} \ldots \xrightarrow{T_1} x_1 \xrightarrow{T_0} x_0$$
$$x_{N-1} \xleftarrow[\tilde{T}_{N-1}]{} x_{N-2} \xleftarrow[\tilde{T}_{N-2}]{} \ldots \xleftarrow[\tilde{T}_1]{} x_1 \xleftarrow[\tilde{T}_0]{} x_0 \sim p_0$$

Guarantees asymptotic correctness.

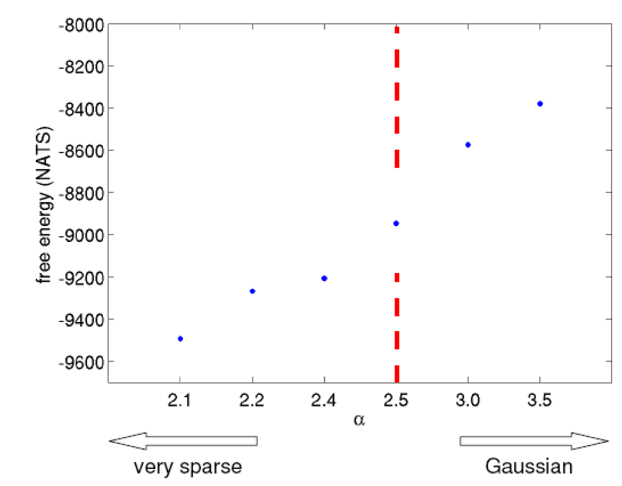
Artificial data

- 2D, 3 components, alpha = 2.5
- 3000 data points
- model initialized with K=7 components

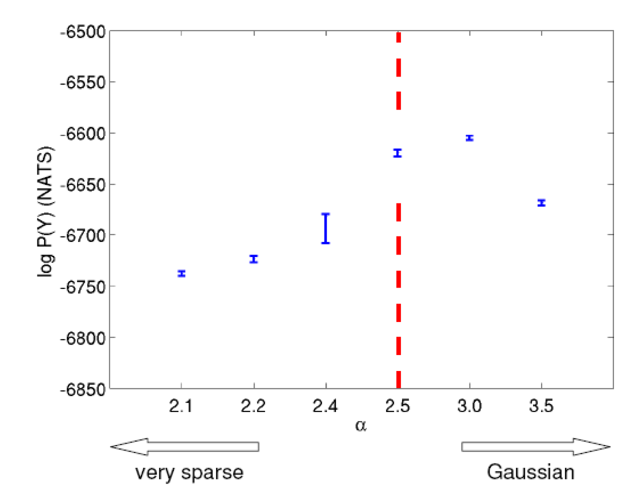
Student-t [1]



Free energy



Marginal likelihood

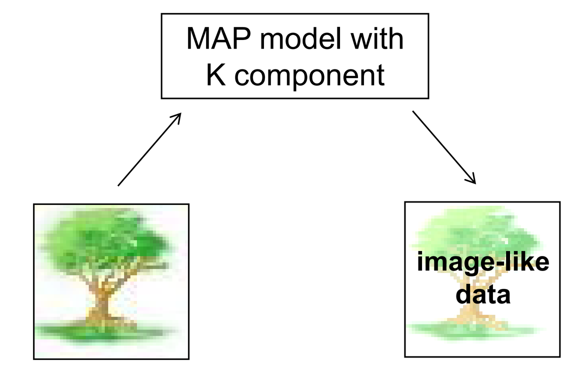


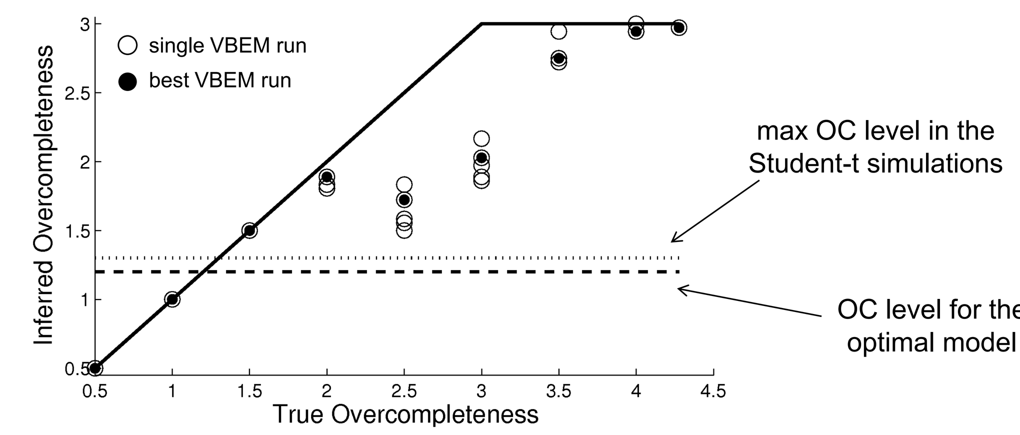
Overcompleteness

Correct number of sources always recovered, except in sparsest case (5)

Image-like data

- Artificial data generated from sparse, Maximum-A-Posteriori model of natural images
- Fixed number of components, alpha=2.5
- Has scale and statistical structure of main simulations





Natural data

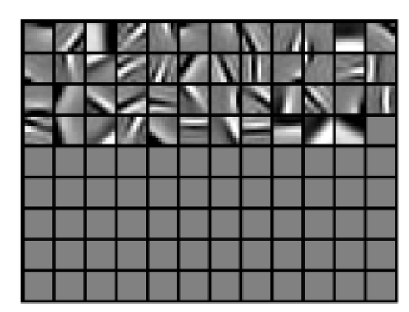
- 9x9 patches from 36 natural images (van Hateren's)
- dimensionality reduced to 36 by PCA
- model initialized to K=108 components
- batch of 3600 patches at every iteration

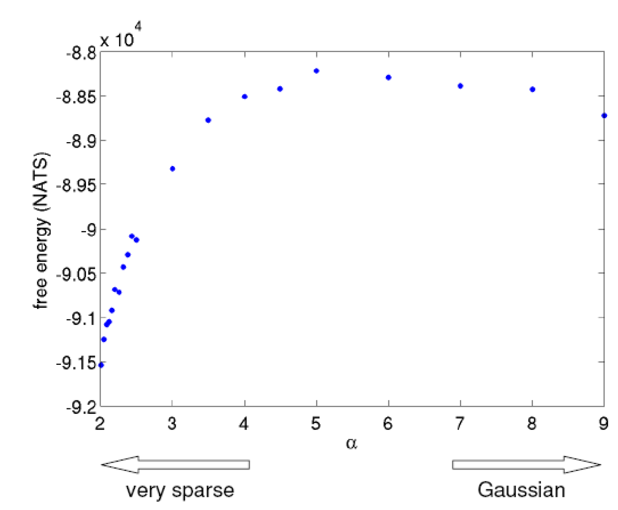
Free energy

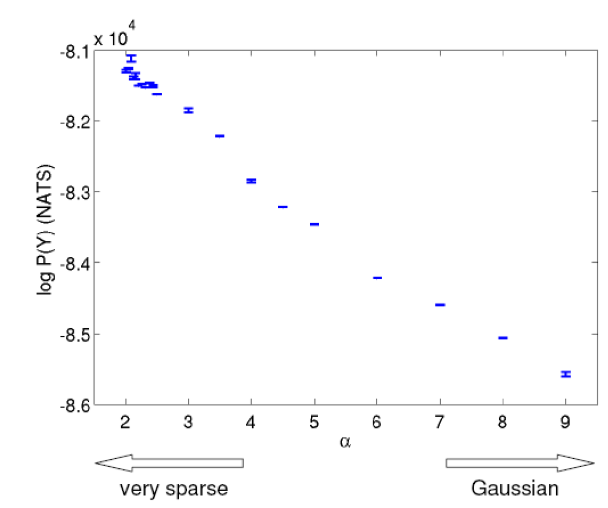
Marginal likelihood

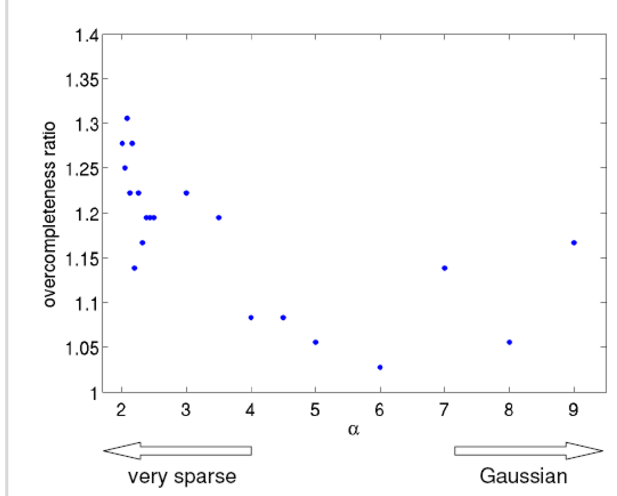
Overcompleteness

Student-t [1]

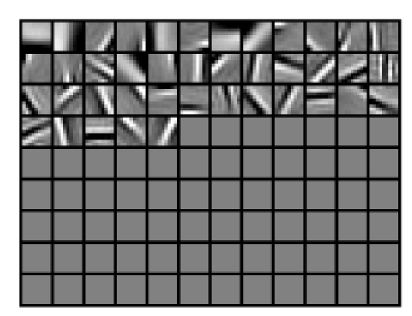


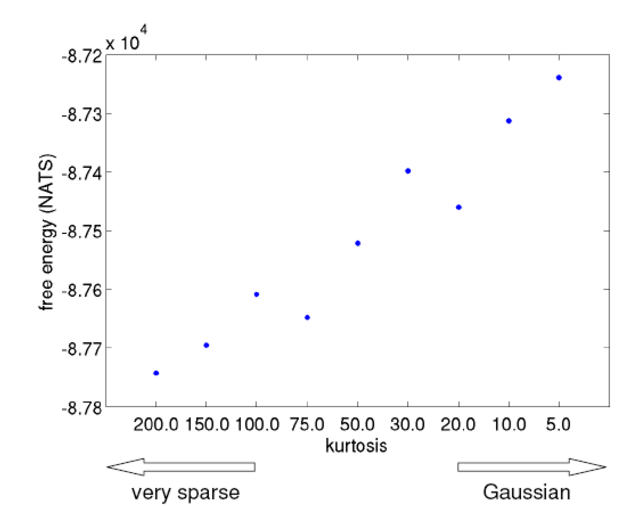


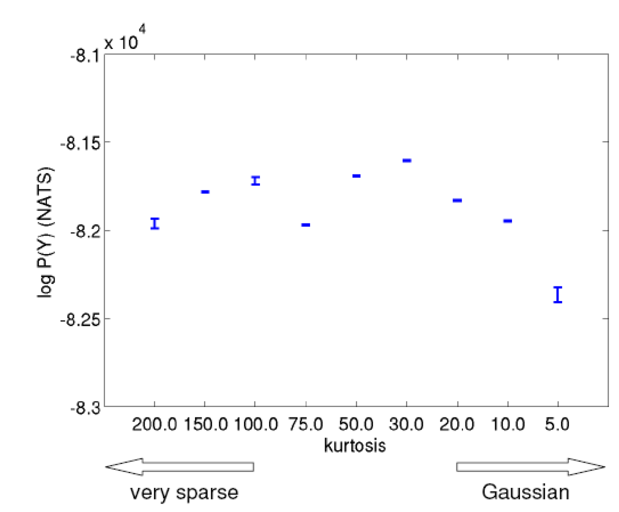


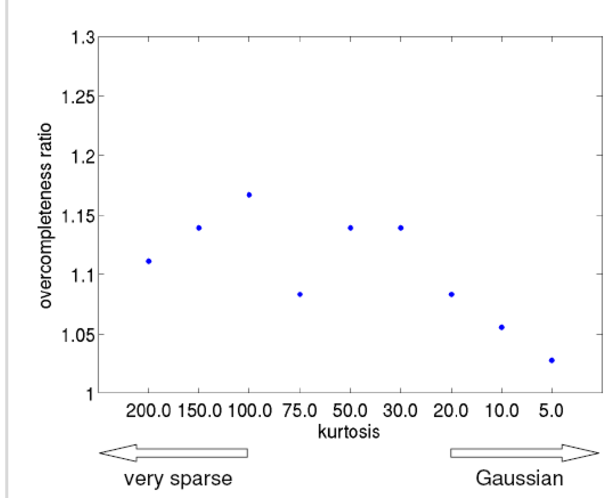


Uniform GSM prior [2]



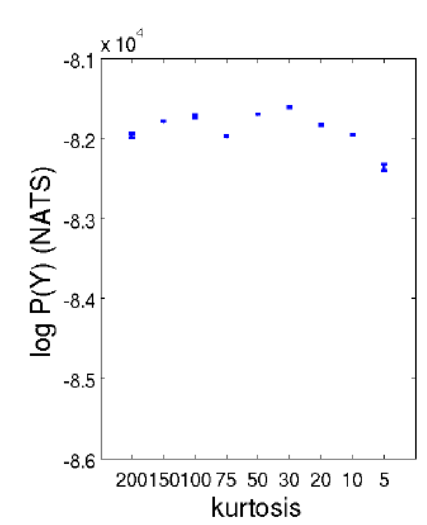


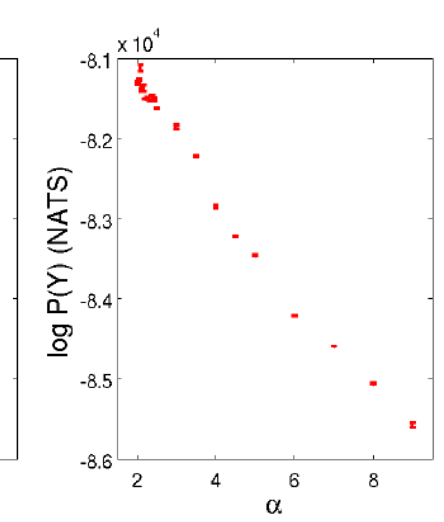




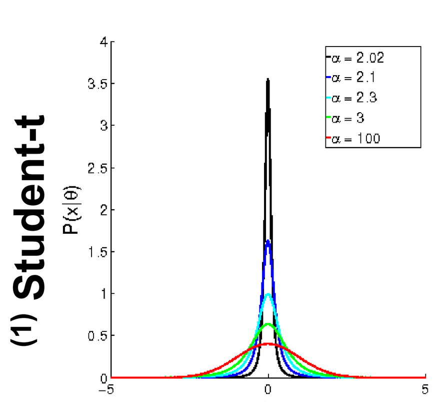
Model comparison

- Within the Student-t family and the uniform GSM family, the optimal model for natural images is very sparse, but only modestly overcomplete
- Very sparse Student-t distribution is a better prior than the uniform GSM prior



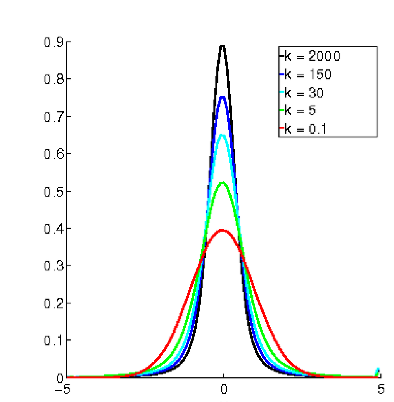


(1) Student-t



$$p(u_{t,k}|\alpha,\lambda) = \mathcal{G}_{u_{t,k}}\left(\frac{\alpha}{2}, \frac{2}{\alpha\lambda^2}\right)$$
$$p(x_{t,k}|u_{t,k}) = \mathcal{N}_{x_{t,k}}\left(0, u_{t,k}^{-1}\right)$$

(2) Uniform GSM prior



$$p(u_{t,k}|a,b) = \mathcal{U}_{u_{t,k}}(a,b)$$
$$p(x_{t,k}|u_{t,k}) = \mathcal{N}_{x_{t,k}}\left(0, u_{t,k}^{-1}\right)$$