On sparsity and overcompleteness in image models

Pietro Berkes, Richard E. Turner, and Maneesh Sahani

Sparse coding

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

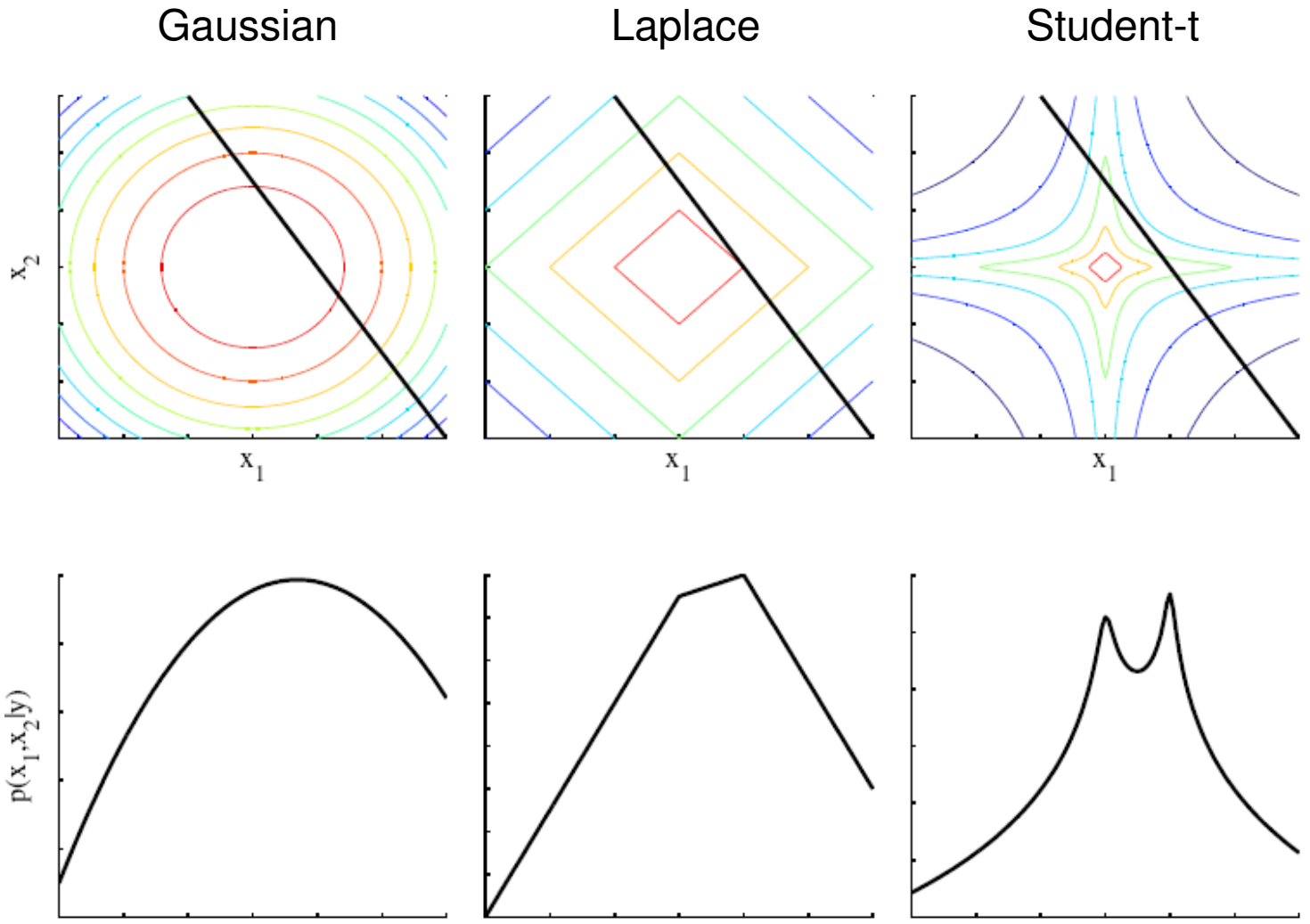
- widely used: BSS, model for early sensory processing in visual and auditory cortex
- motivation for sparse prior different in different communities; BSS: non-gaussianity; sensory cortex: generative model, each image/sound produced by small number of external causes

Overcomplete sparse coding

- Reasons for overcomplete representation:
  - there are more causes than input dimensions (e.g., two speakers, one microphone; more causes than pixels)
  - the brain does it
  - why not? completeness is a very special case
  - even in the noiseless case, inference becomes nonlinear (O&F, 1997)

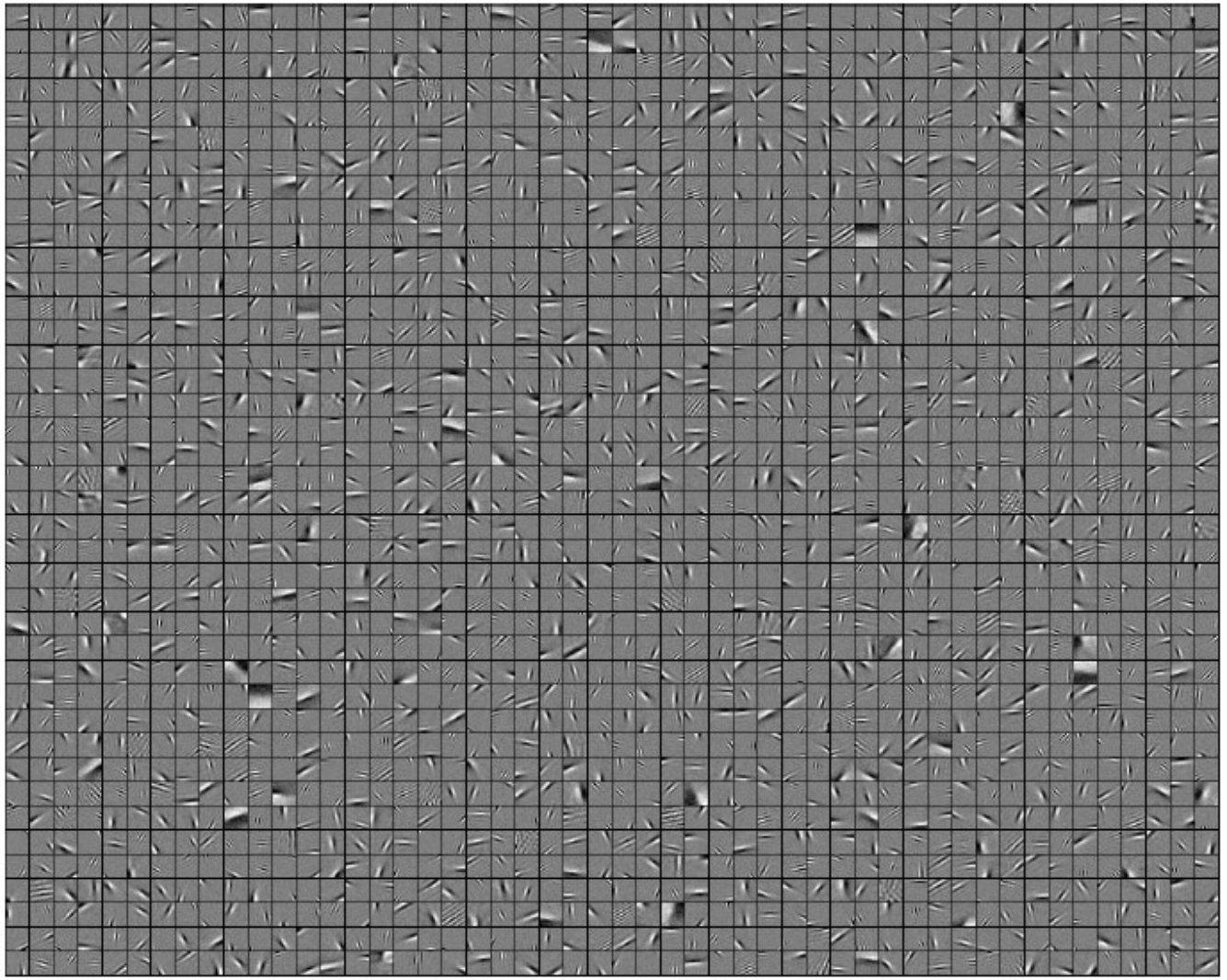
Why inference gets interesting

A 1D model with 2 components: $y = g_1 x_1 + g_2 x_2$



Overcomplete sparse coding

Previous work: sparse representation with fixed model size



5 times overcomplete (Lee et al., 2007)

Legitimate but neglected questions

- How sparse? Which family of distributions?
- How overcomplete? (might be overfitting)
- One might expect a trade-off between sparseness and overcompleteness
- Simple questions, require complicated machinery

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

Model selection in practice

- Tiling the sparseness/overcompleteness space is out of question
- Strategy:
  - use Automatic Relevance Determination (ARD) prior, Variational Bayes EM to determine the posterior over parameters and the overcompleteness
  - Unfortunately VB is biased (cannot use the free energy)
  - compute the marginal likelihood using Annealed Importance Sampling (AIS)

ARD

- Place a Gaussian prior over the components to favor small weights

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

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

VBEM is biased

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

Ξ : hyperparameters
V : latent variables
Θ : parameters

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.

Annealed Importance Sampling (AIS)

- One of the few methods for evaluating normalizing constants of intractable distributions
- Idea 1: Simulated annealing (Kirkpatrick et al., 1983)

$$p_j(x) = p_0(x)^{\beta_j} p_N(x)^{1-\beta_j}$$

$p_N(x)$: prior distribution

$p_0(x)$: unnormalized posterior distribution

$$0 = \beta_N > \ldots > \beta_0 = 1$$

The annealing process is a heuristic to avoid getting stuck in local modes. However, there is no guarantee about finding each mode with the right probability.

Annealed Importance Sampling (AIS)

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

$$p_N \sim x_{N-1} \overset{T_{N-1}}{\longrightarrow} x_{N-2} \overset{T_{N-2}}{\longrightarrow} \ldots \overset{T_1}{\longrightarrow} x_1 \overset{T_0}{\longrightarrow} x_0$$

$$x_{N-1} \underset{\tilde{T}_{N-1}}{\longleftarrow} x_{N-2} \underset{\tilde{T}_{N-2}}{\longleftarrow} \ldots \underset{\tilde{T}_1}{\longleftarrow} x_1 \underset{\tilde{T}_0}{\longleftarrow} x_0 \sim p_0$$

  Sample from $p_N \cdot T_{N-1} \cdot \ldots \cdot T_1 \cdot T_0$

  to get importance weights for $\tilde{T}_{N-1} \cdot \ldots \cdot \tilde{T}_1 \cdot \tilde{T}_0 \cdot p_0$ (marginal distribution is $p_0(x)$)

  - Guarantees asymptotic correctness (Neal, 2001)

In Summary

1) Use Automatic Relevance Determination (ARD) prior, Variational Bayes EM to determine the posterior over parameters and the overcompleteness
2) Compute the marginal likelihood using Annealed Importance Sampling (AIS)

Results (1): Student-t family

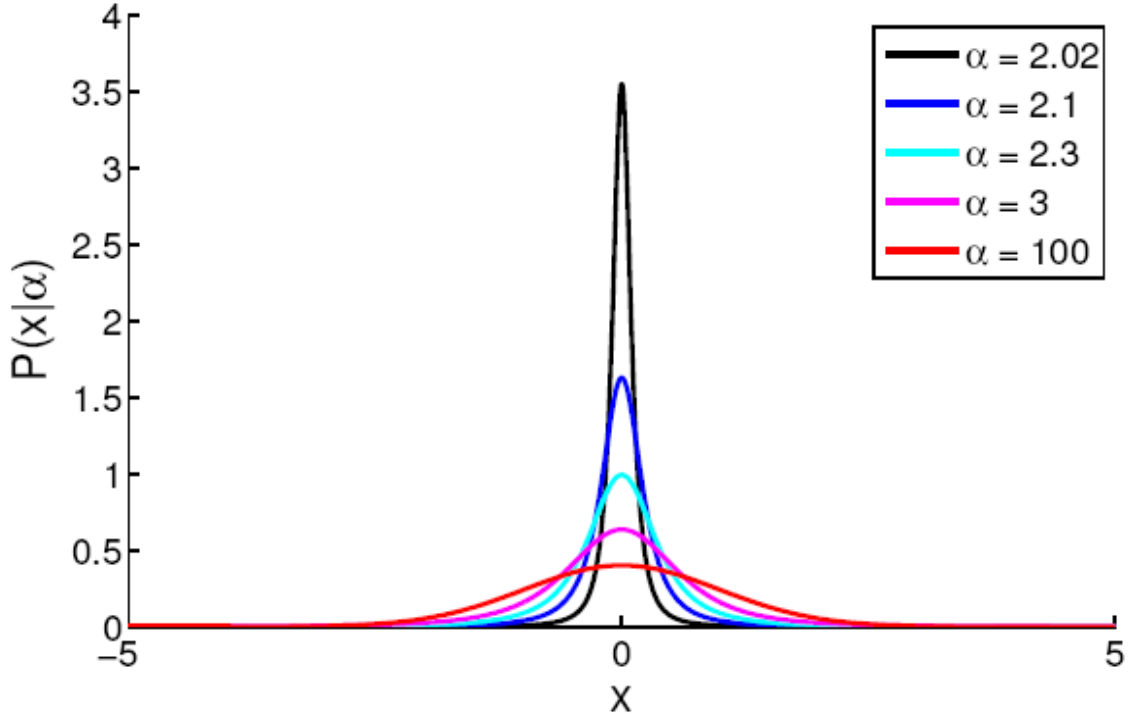
$$p(x_{t,k}|\alpha,\lambda) = \frac{1}{Z}\left(1+\frac{1}{\alpha}\left(\frac{x_{t,k}}{\lambda}\right)^2\right)^{-\frac{\alpha+1}{2}}$$

or

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

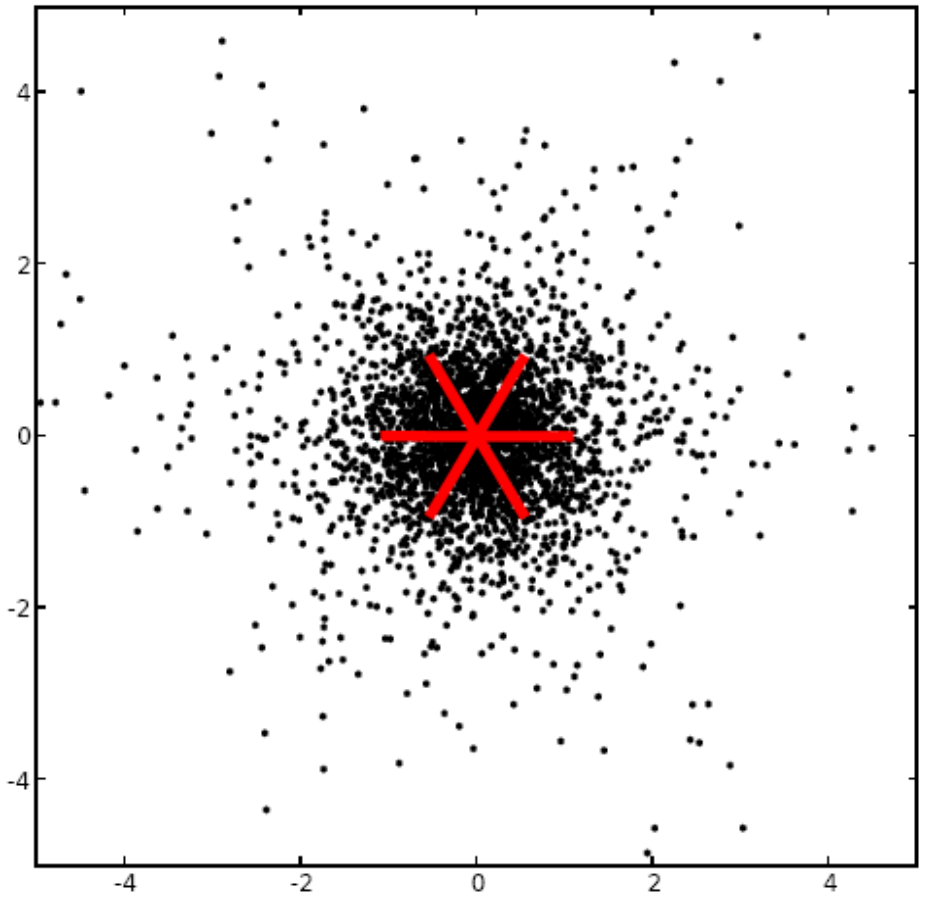
$$p(x_{t,k}|u_{t,k}) = \mathcal{N}_{x_{t,k}}\left(0,u_{t,k}^{-1}\right)$$

For given alpha, lambda chosen such that prior variance is 1



Artificial data

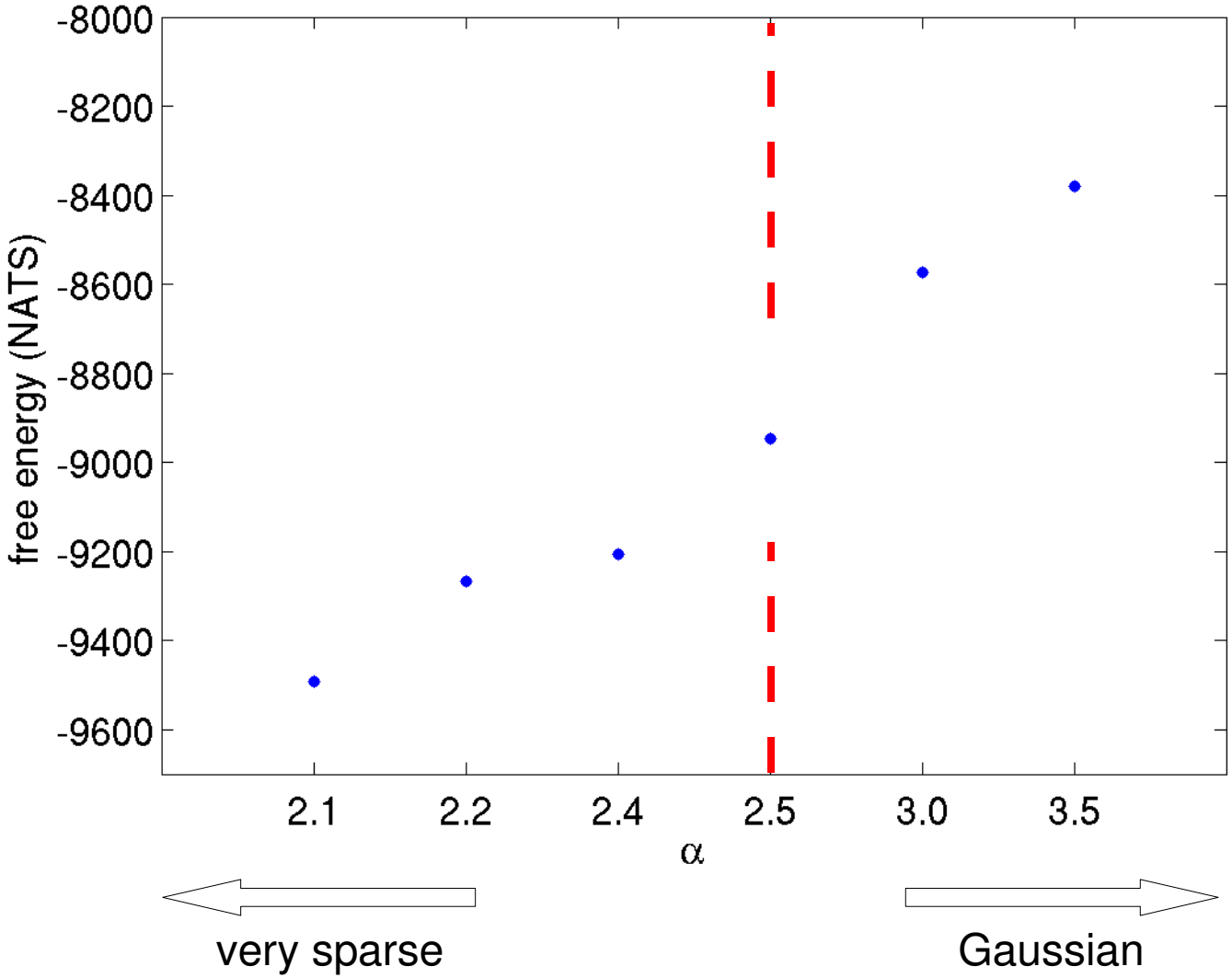
- 2 dimensions, 3 sources, alpha = 2.5
- model initialized with K=7 components
- 3000 data points



Artificial data – free energy

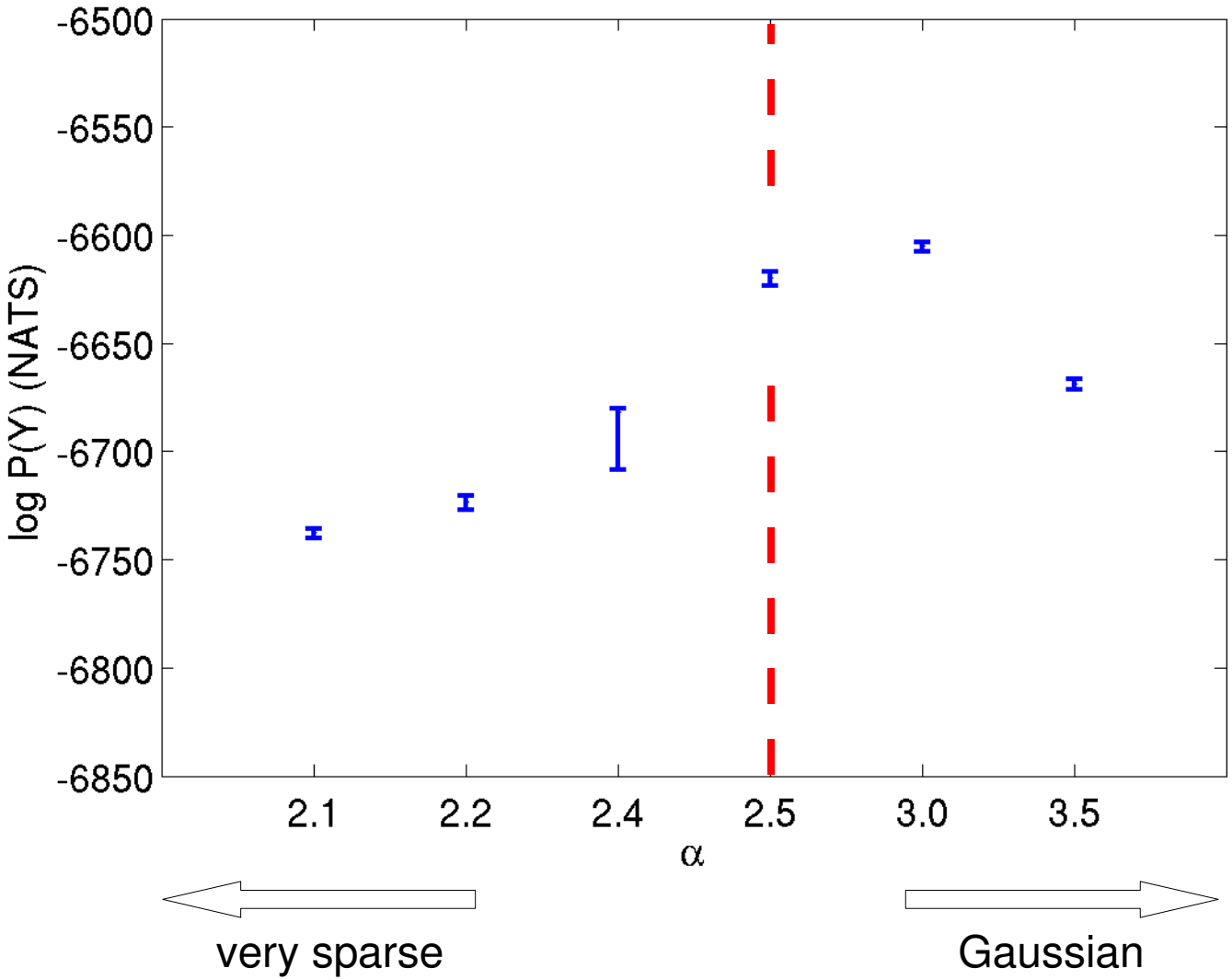
Results are from best of 3 simulations

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



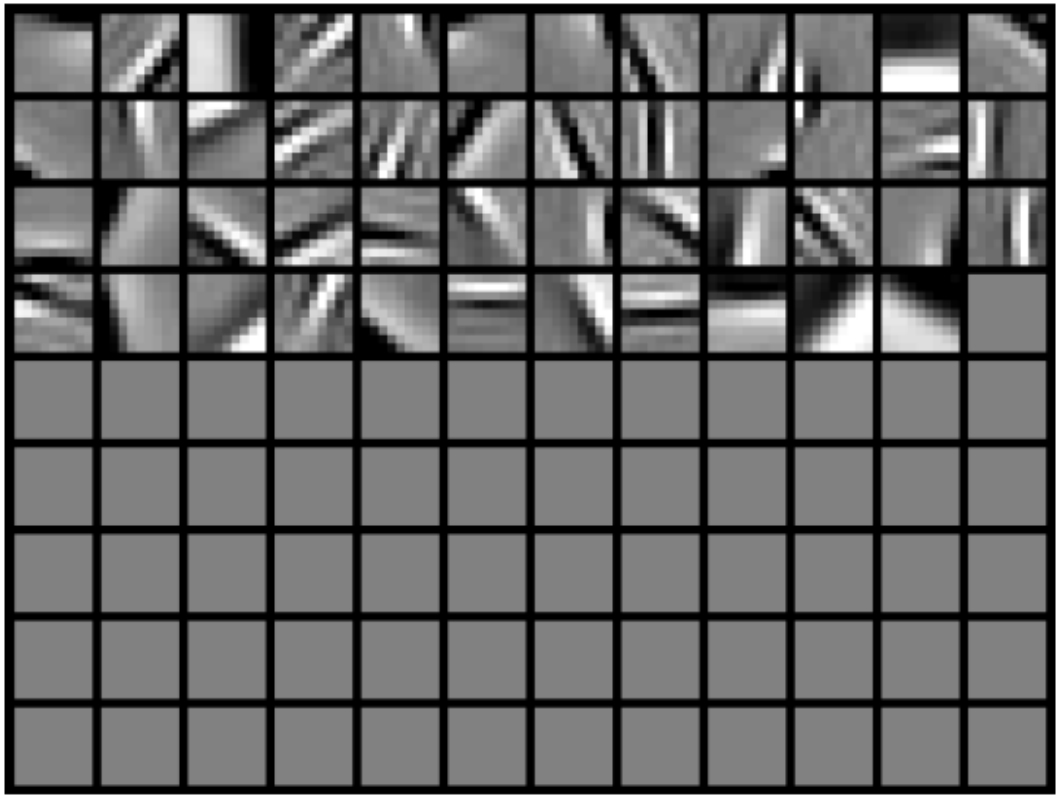
Artificial data – marginal likelihood

Error bars are always 3 times the estimated standard deviation

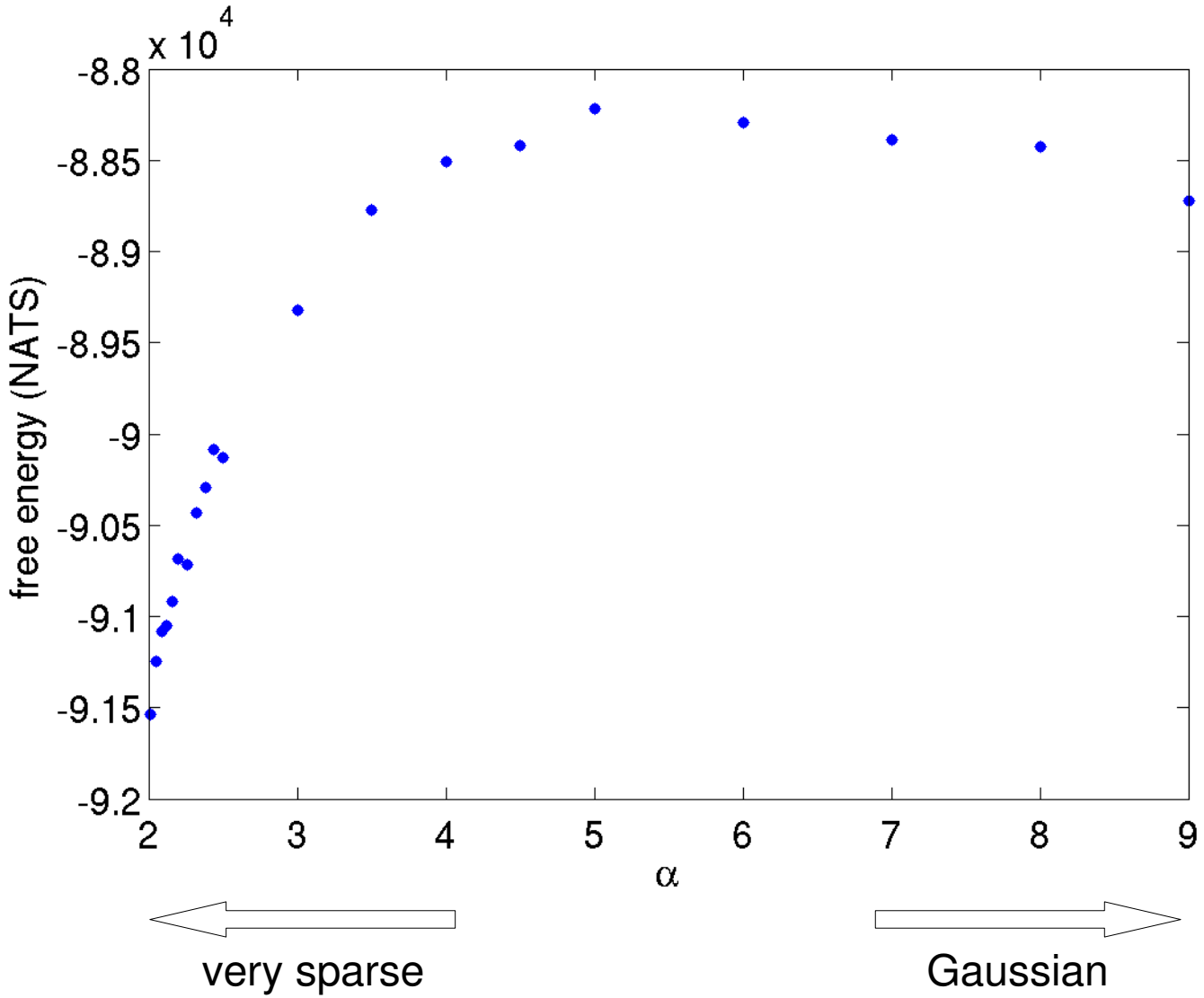


Natural images

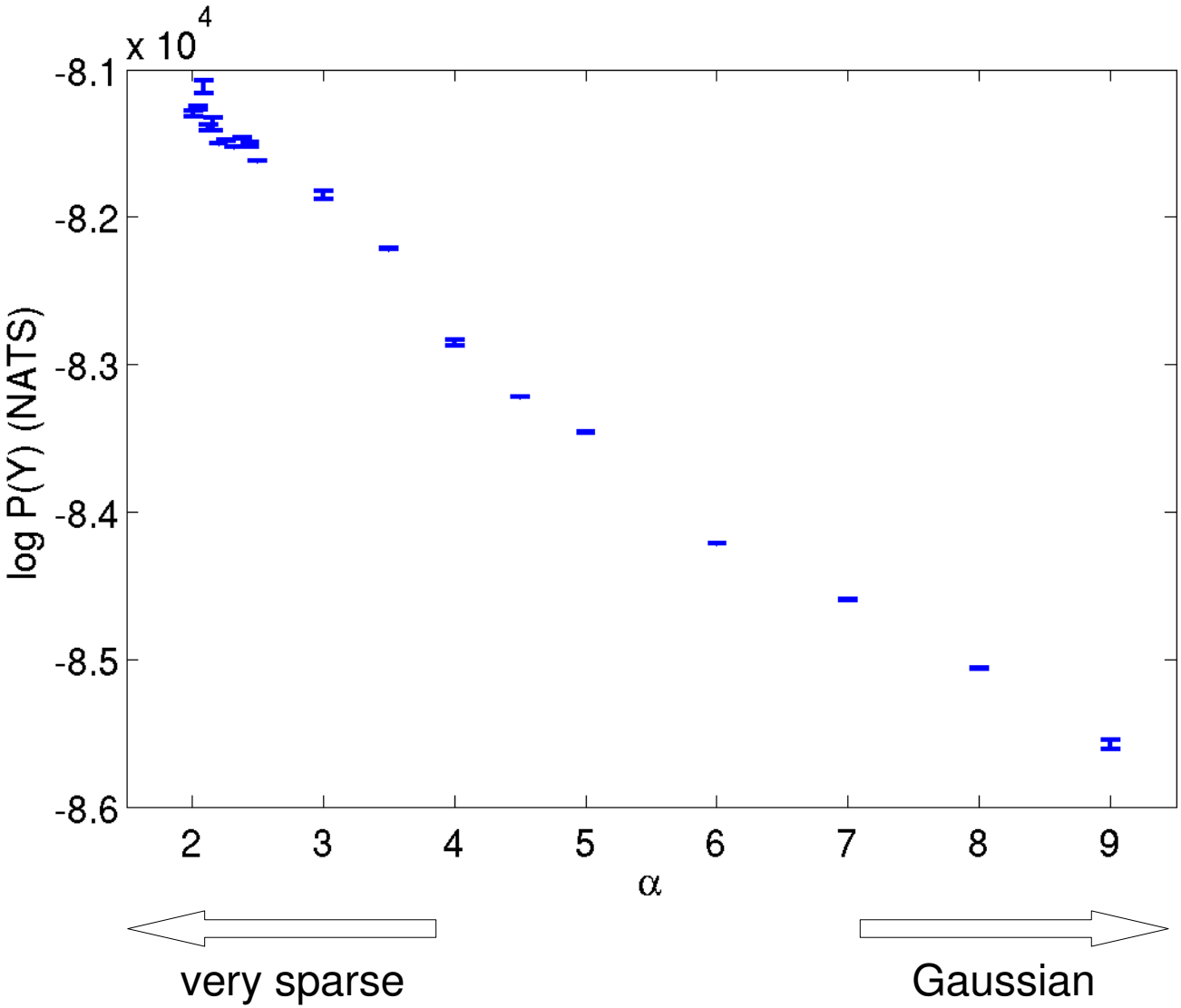
- 9x9 patches from 36 natural images (van Hateren's)
- dimensionality reduced to 36 by PCA
- model initialized with K=108 components
- new batch of 3600 patches at every iteration



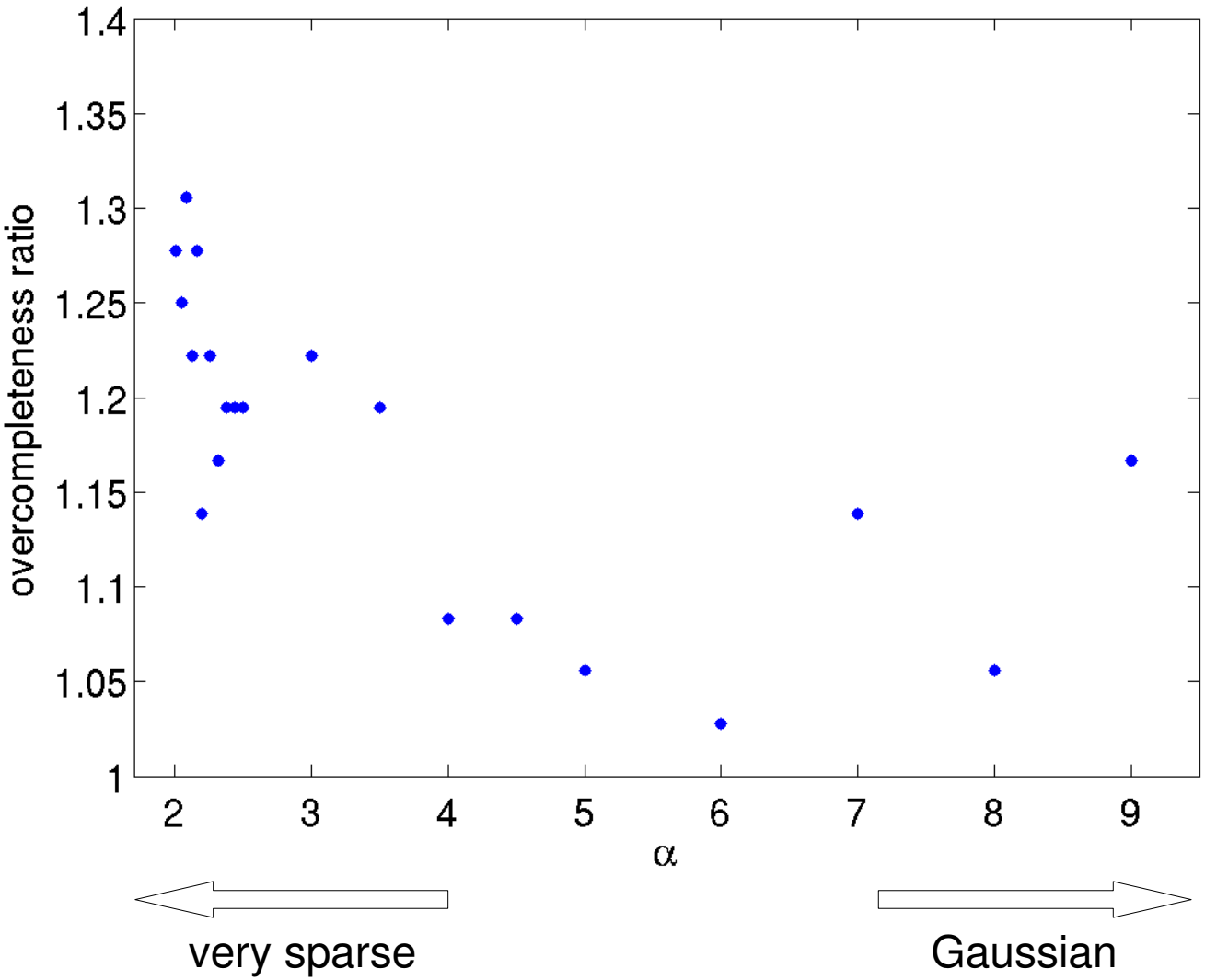
Natural images – free energy



Natural images – marginal likelihood



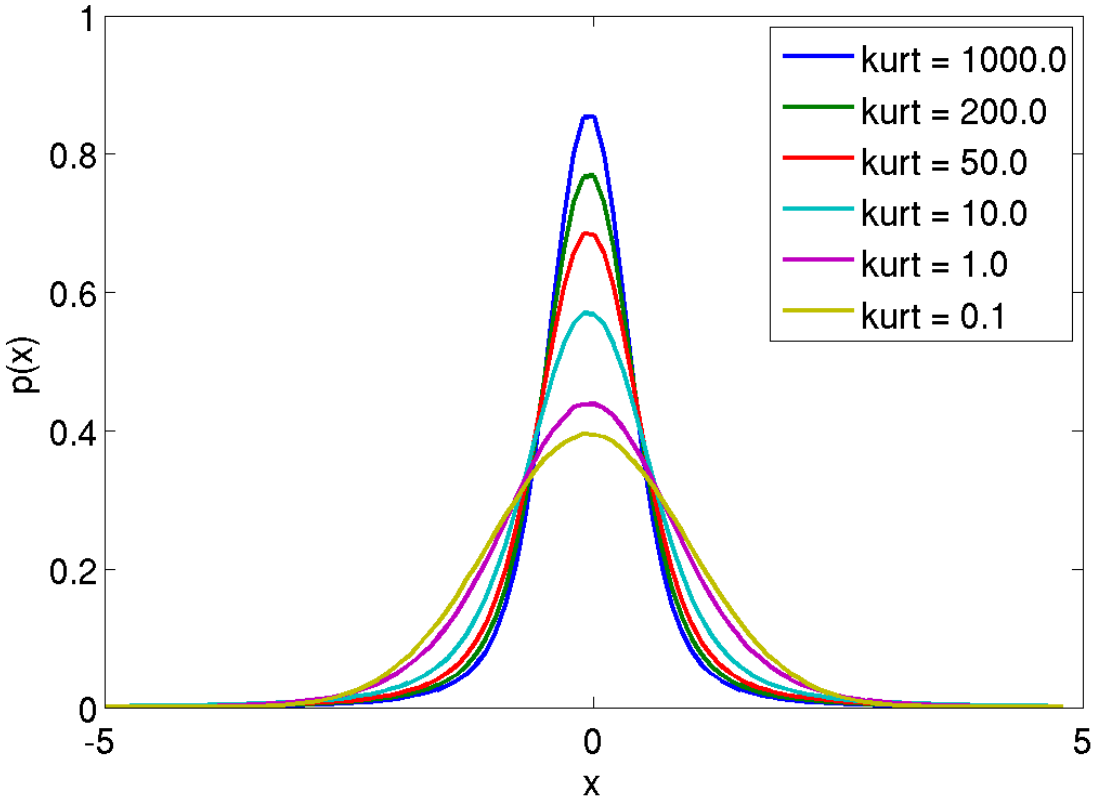
Natural images - overcompleteness



Results (2): “Maneesh's prior”

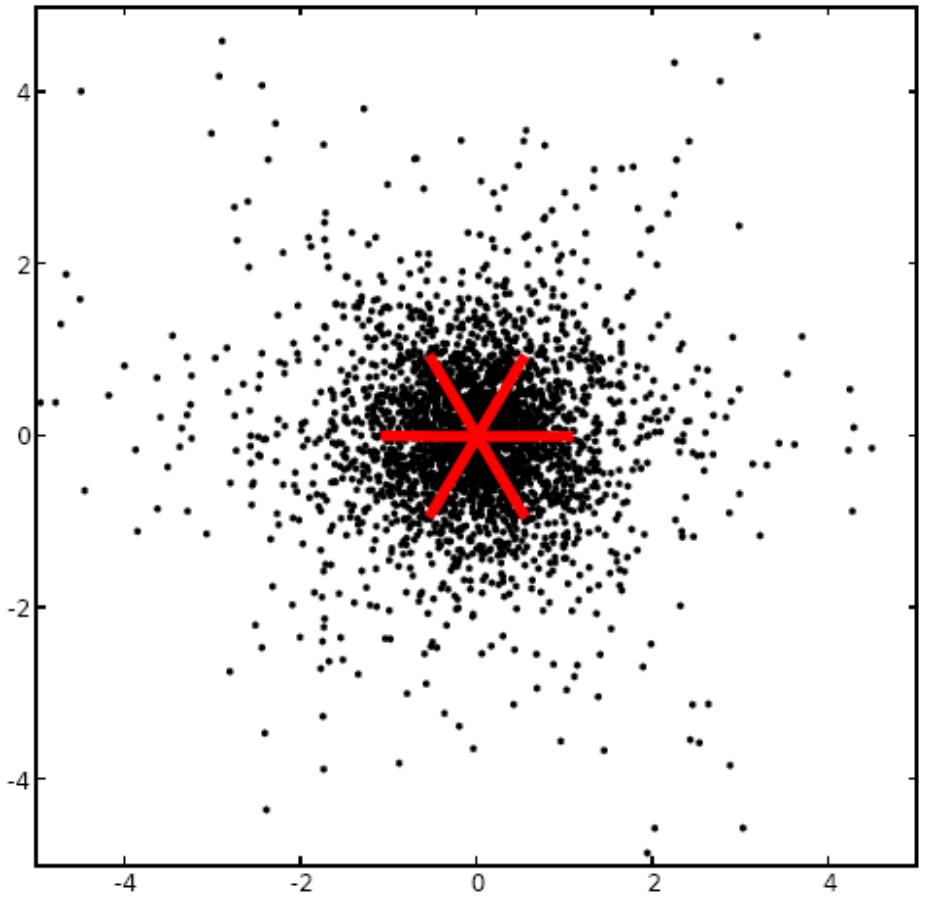
- Student-t might not be such a good prior
- alternative: $p(u_{t,k}|a,b) = \mathcal{U}_{u_{t,k}}(a,b)$

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



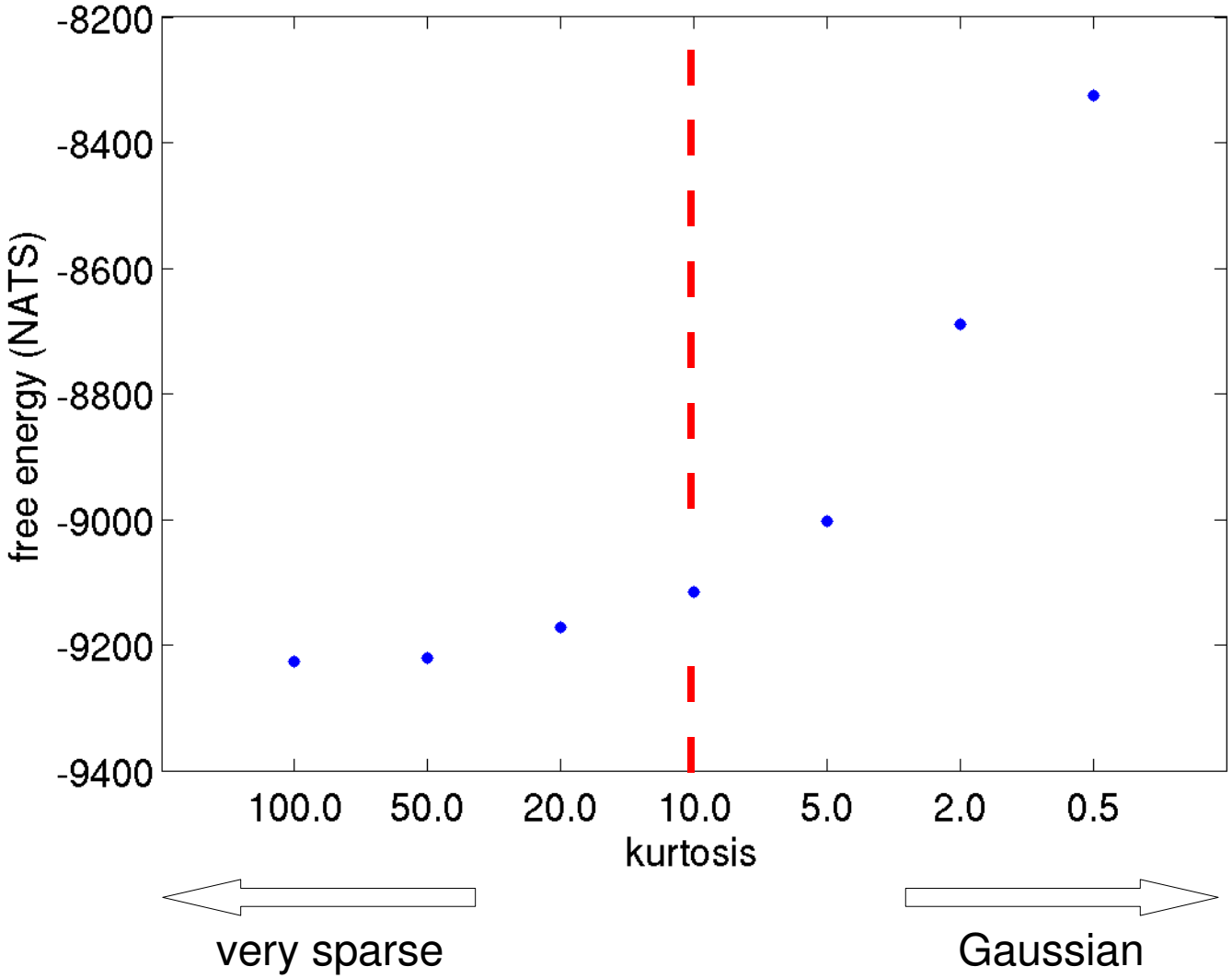
Artificial data

- 2 dimensions, 3 sources, kurtosis = 10.0
- model initialized with K=7 components
- 3000 data points

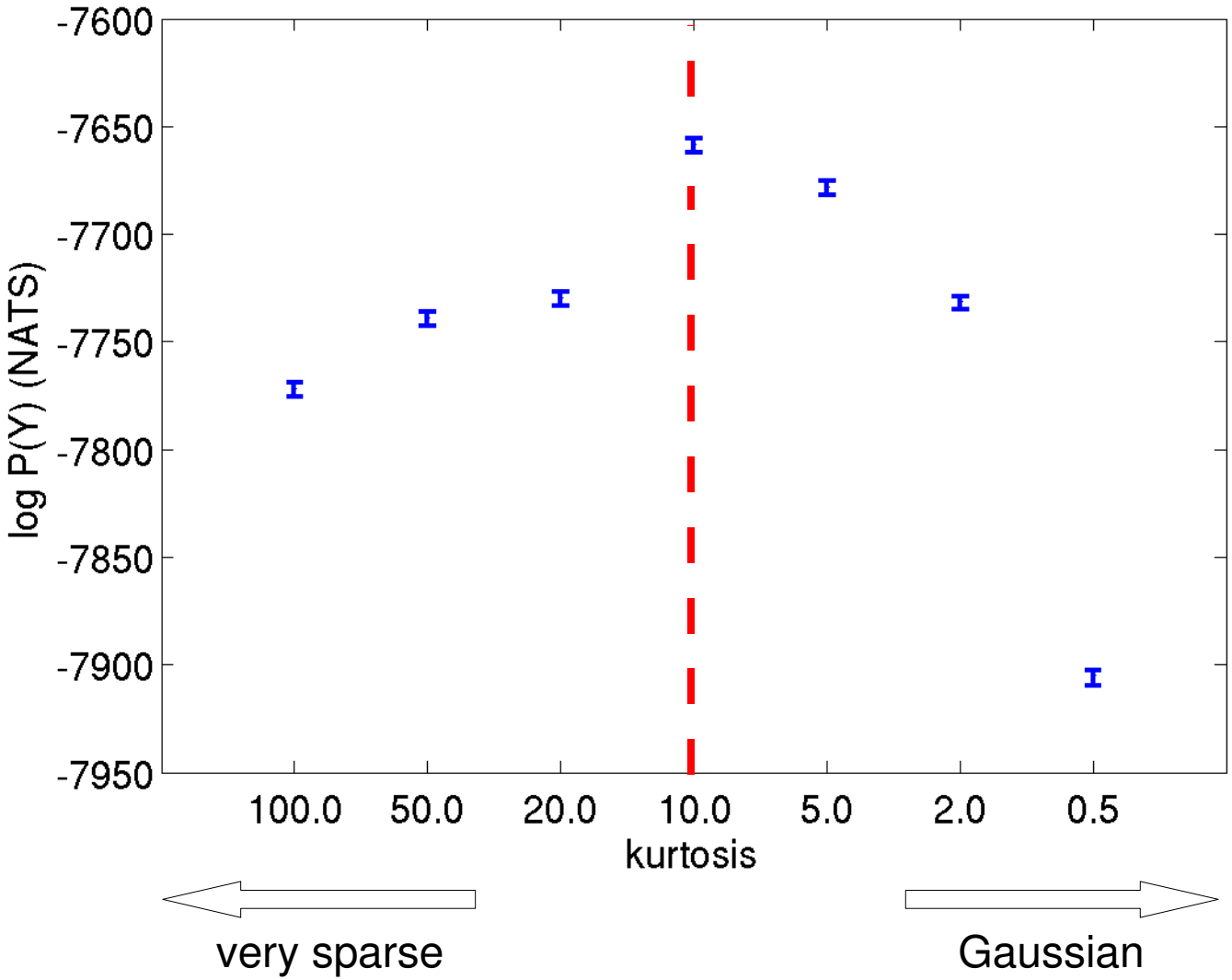


Artificial data – free energy

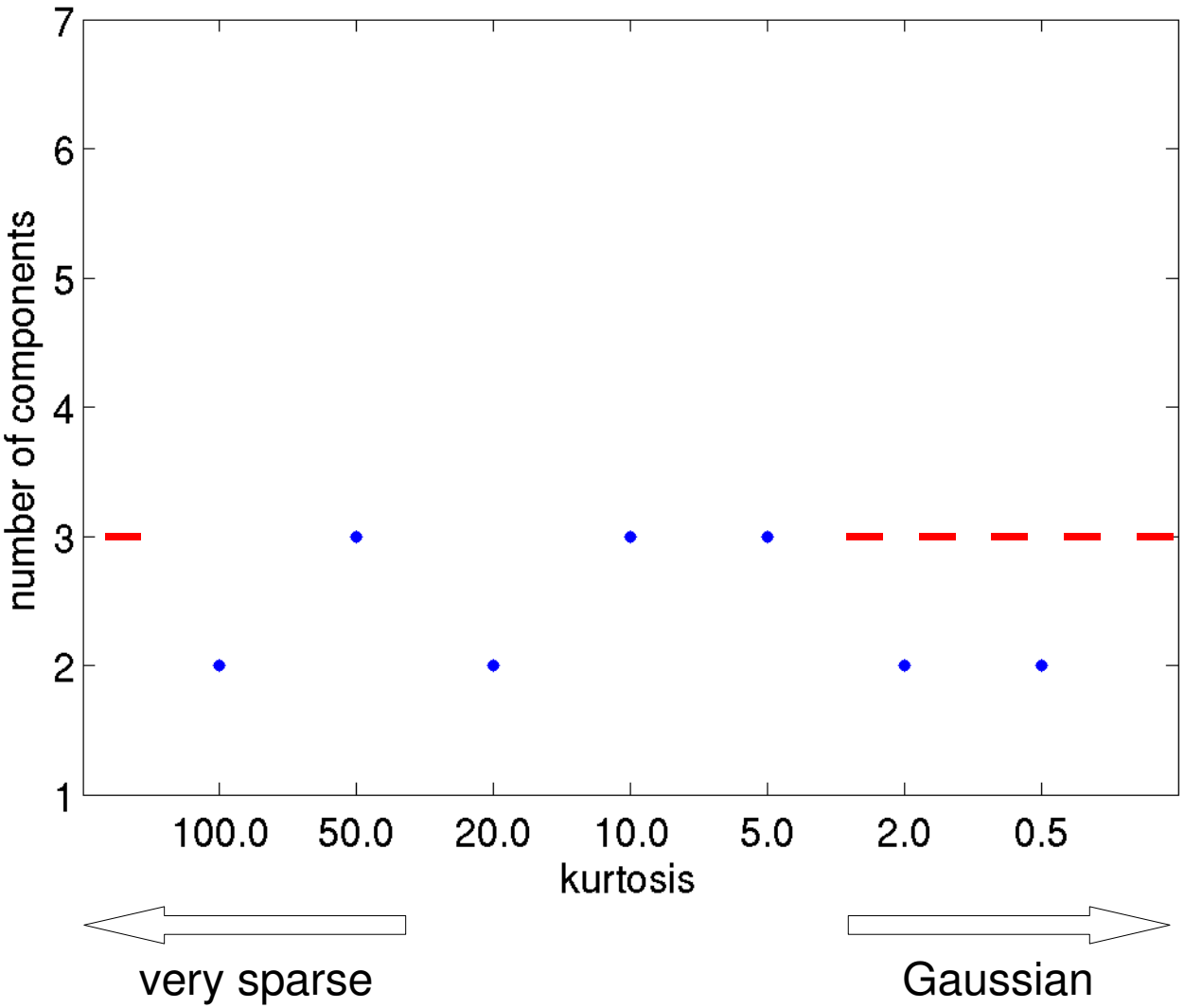
Results are from best of 10 simulations



Artificial data – marginal likelihood

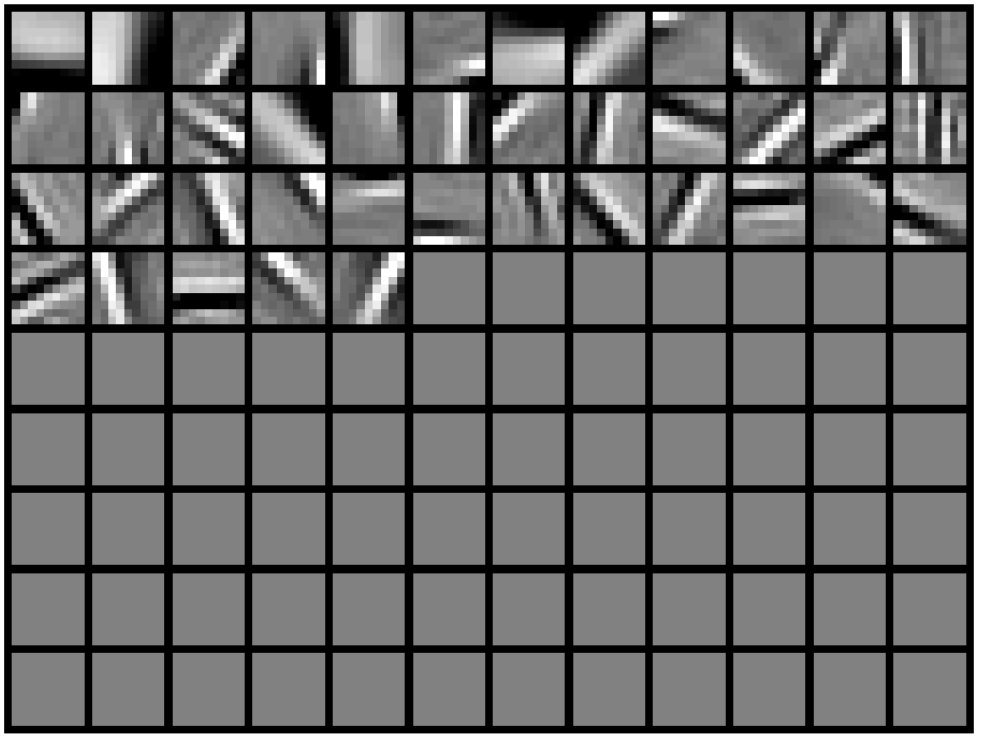


Artificial data – number of components

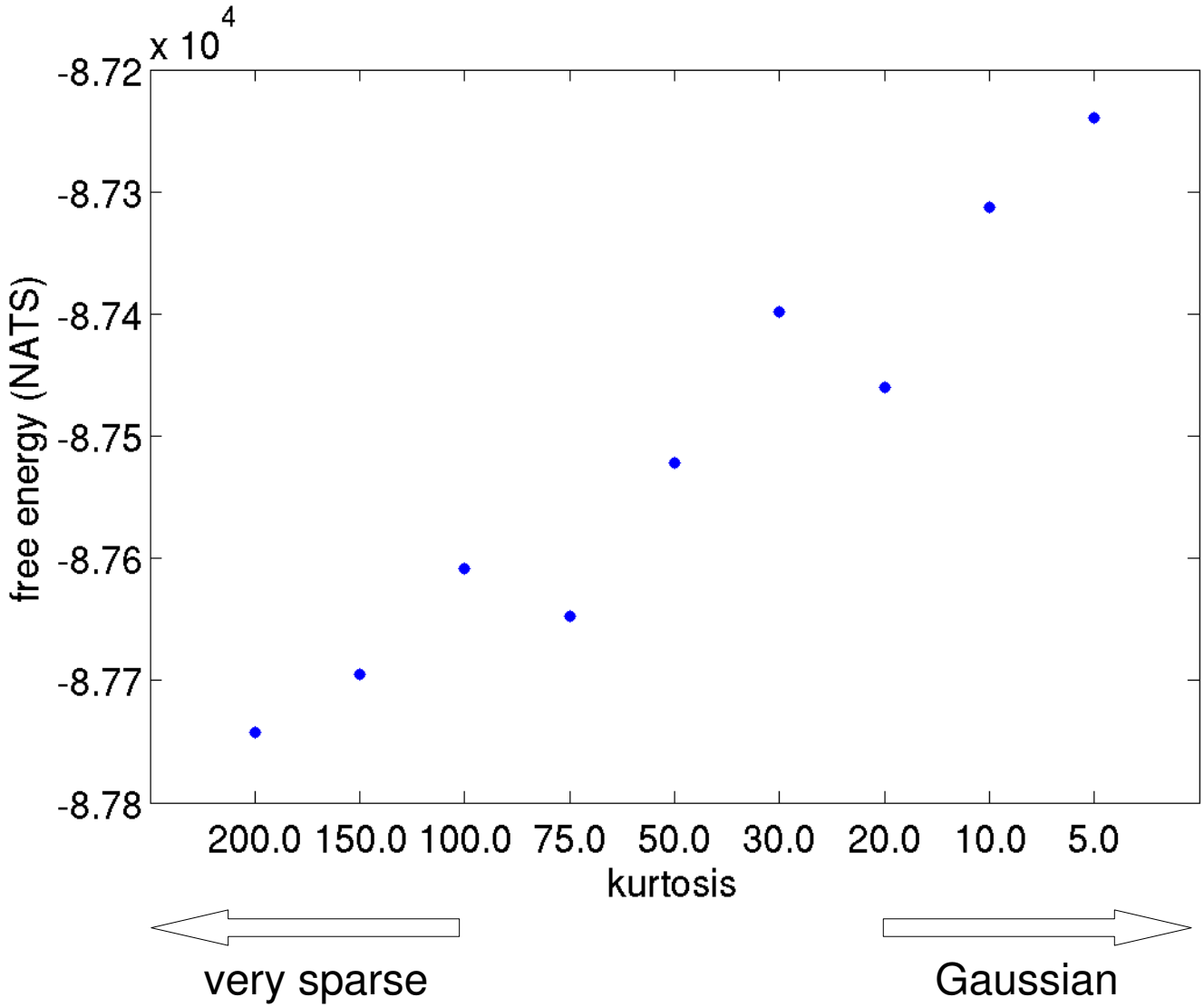


Natural images

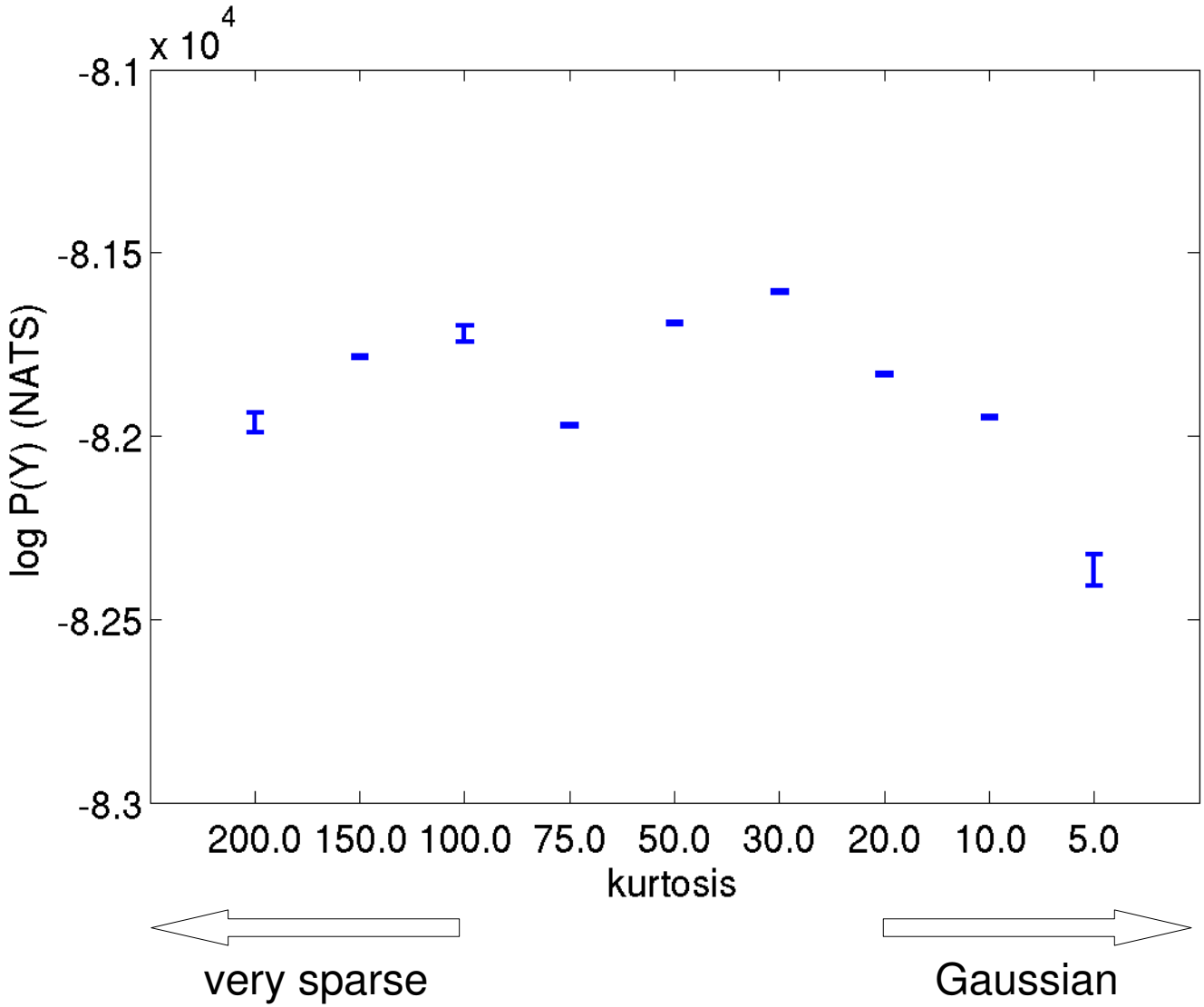
- 9x9 patches from 36 natural images (van Hateren's)
- dimensionality reduced to 36 by PCA
- model initialized with K=108 components
- new batch of 3600 patches at every iteration



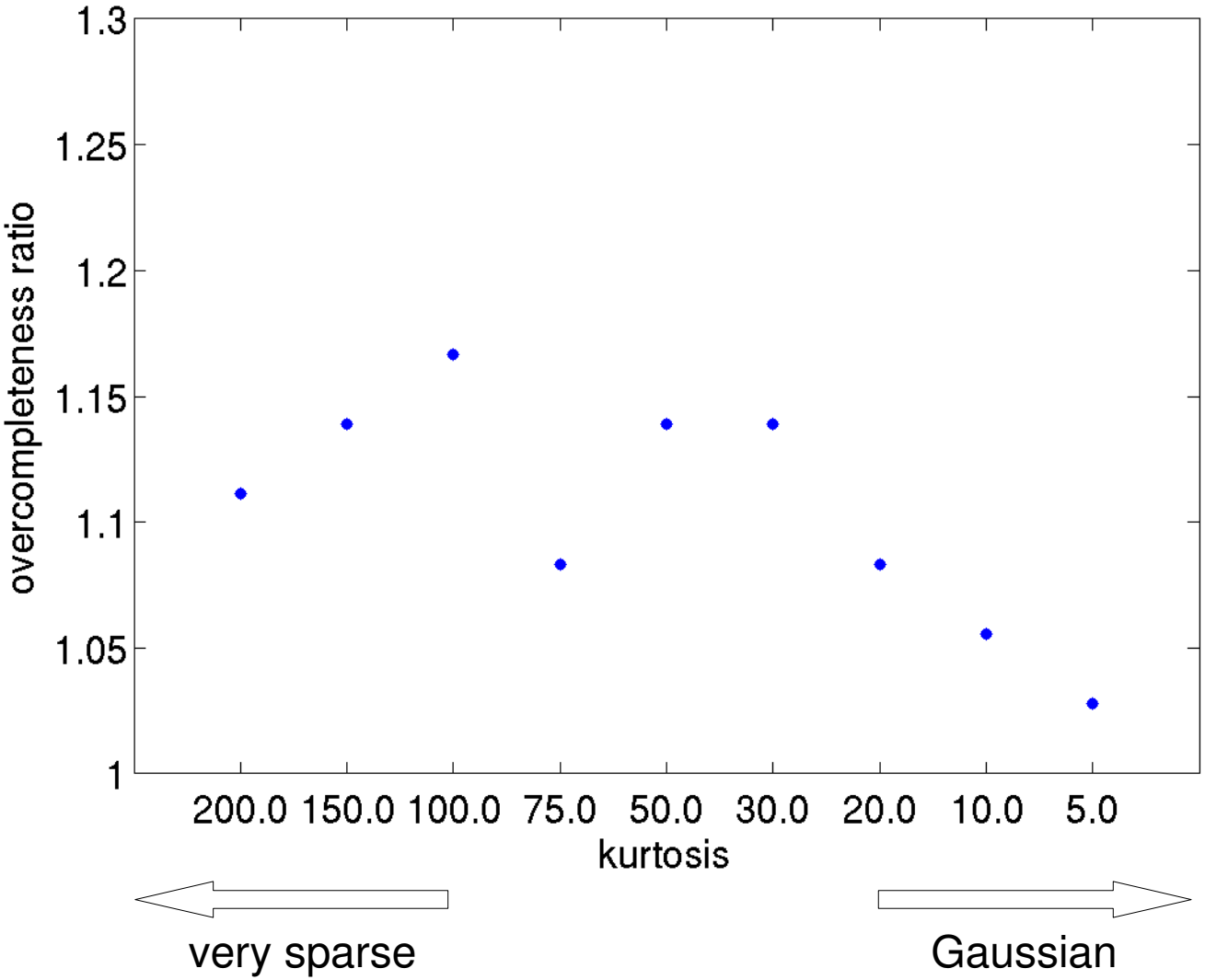
Natural images – free energy



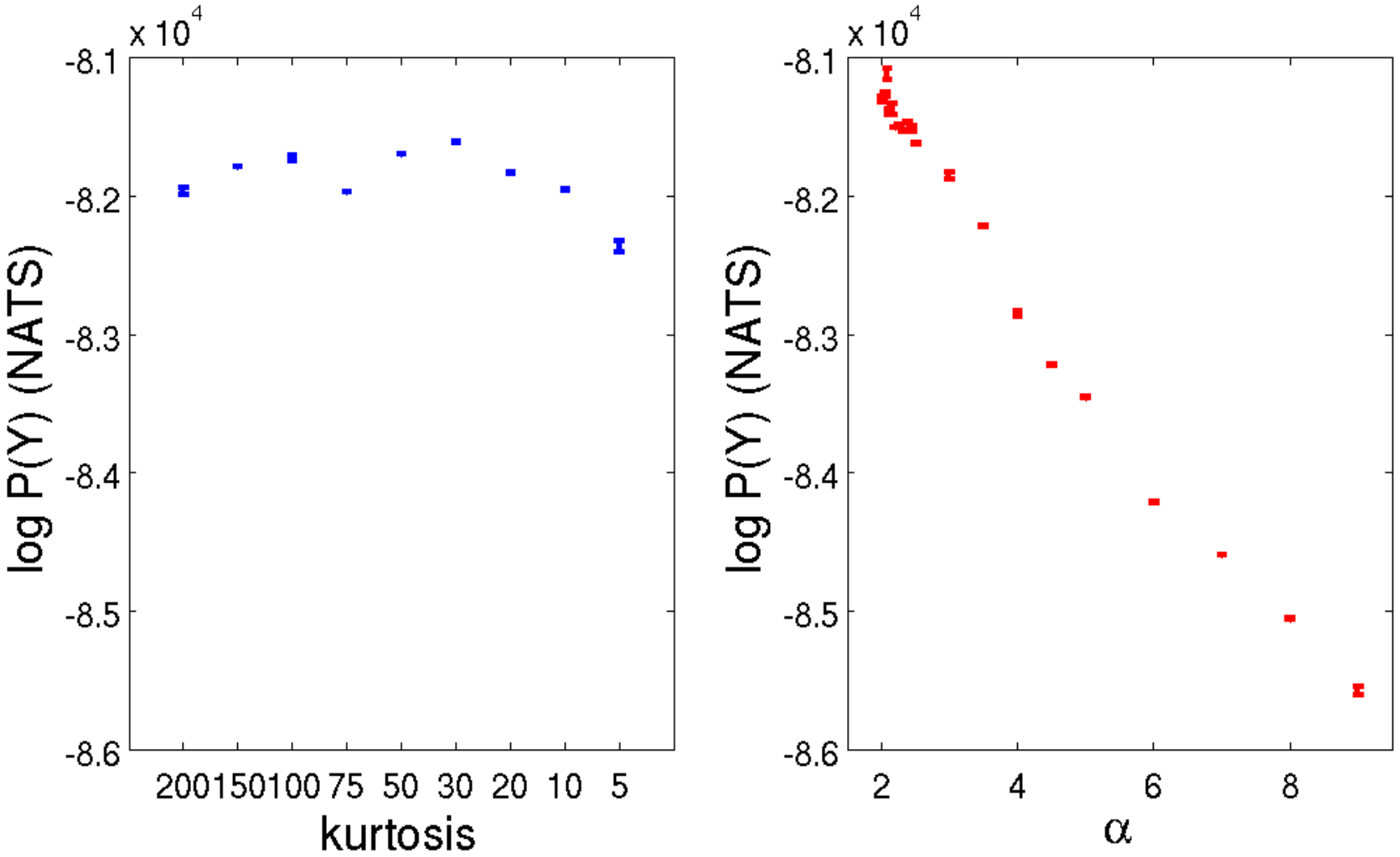
Natural images – marginal likelihood



Natural images - overcompleteness



Model comparison

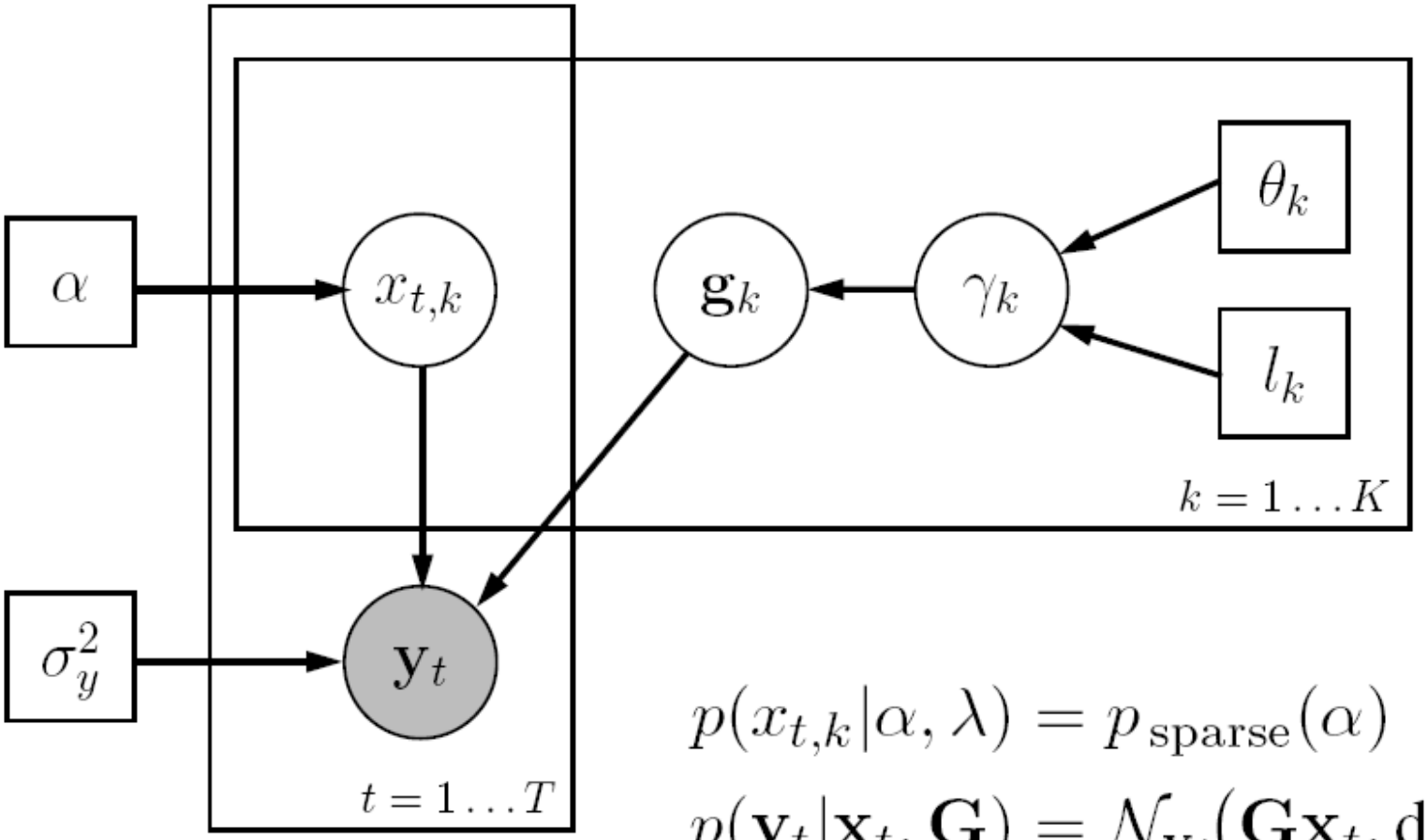


Conclusions

- What is the optimal level of sparseness and overcompleteness within a sparse coding model? Which model is optimal for, let's say, natural images?
  - Within the Student-t family and the Maneesh family, the optimal model for natural images is very sparse, but only modestly overcomplete
  - Very sparse Student-t distribution is a better prior for natural images than the Maneesh prior (todo: explore other priors, Laplace, gen. Gaussian)
- Of course, there might be other reasons to be overcomplete; todo: generate from an overcomplete, non-linear model and check linear solution

Many thanks to:
Yee Whye Teh
Iain Murray
David McKay

Complete model



$$p(x_{t,k}|\alpha, \lambda) = p_{\text{sparse}}(\alpha)$$

$$p(\mathbf{y}_t|\mathbf{x}_t, \mathbf{G}) = \mathcal{N}_{\mathbf{y}_t}\left(\mathbf{G}\mathbf{x}_t, \text{diag}\left(\sigma_y^2\right)\right)$$

$$p(\mathbf{g}_k|\gamma_k) = \mathcal{N}_{\mathbf{g}_k}\left(\mathbf{0}, \gamma_k^{-1}\right)$$

$$p(\gamma_k) = \mathcal{G}_{\gamma_k}(\theta_k, l_k) \ .$$