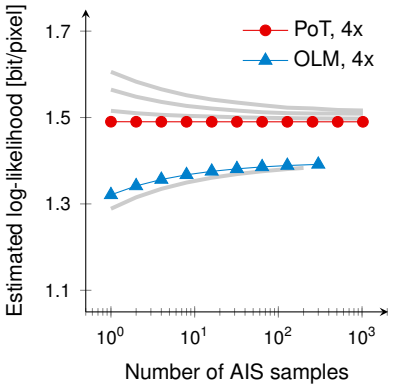

Figure 1: Convergence behavior of AIS estimates for four-times overcomplete models on 8×8 image patches. The number of intermediate distributions was 300 for the overcomplete linear model and 10^7 for the PoT model. Gray lines correspond to estimates with alternative numbers of intermediate distributions.

1 Energy trace plots and autocorrelation functions

To generate the trace plots for the toy model, we first jointly sampled 5000 hidden states $\mathbf{z}$ and visible states $\mathbf{x}$. We then sampled from the posterior $p(\mathbf{z} \mid \mathbf{x})$ using either Gibbs sampling or HMC. Markov chains z_t were initialized with $z_0 = BA^\top x$ for each visible state x. Finally, we measured the negative log-density (or energy) of posterior samples at each time step,

$$-\log p(x, z_t),$$

and averaged over posterior samples and data points. Since the hidden states jointly drawn from the prior can be treated as unbiased samples from the posterior, we can use them to obtain an unbiased estimate of the energy the Markov chains should converge to (dashed line in Figure 2A of main text). For the image model, we used 200 actual image patches taken from the van Hateren dataset instead of samples from the model and therefore could not compute an unbiased estimate of the energy.

For the autocorrelation plots, we used the following generalization of autocorrelation to multivariate Markov chains,

$$R(\tau) = \frac{E[(z_t - \mu)^\top (z_{t+\tau} - \mu) \mid x]}{E[(z_t - \mu)^\top (z_t - \mu) \mid x]},$$

where μ is $E[z_t \mid x]$. For the image model, the Markov chain used to estimate 80 seconds of the autocorrelation function was 20000 seconds long. For the toy model, we used a 10000 seconds long Markov chain to estimate 15 seconds.

2 Evaluating log-likelihoods

Computing the average log-likelihood of the overcomplete linear model (OLM) and the product of Student-t distributions (PoT) requires the estimation of analytically intractable integrals. Evaluating the OLM involves an integral over hidden states for each data point,

$$p_{\text{OLM}}(x) = \int p_{\text{OLM}}(x, z)\, dz. \tag{1}$$

To evaluate the PoT, on the other hand, we need to integrate over visible states to compute its normalization constant,

$$Z = \int p^*_{\text{PoT}}(x)\, dx, \tag{2}$$

where $p^*_{\text{PoT}}(x)$ is the unnormalized density of the PoT, which can be evaluated easily.

Both quantities can be estimated using a form of importance sampling, for instance,

$$Z = \int q(x) \frac{p^*_{\text{PoT}}(x)}{q(x)}\, dx \approx \frac{1}{K} \sum_k \frac{p^*_{\text{PoT}}(x^{(k)})}{q(x^{(k)})} = \hat{Z}, \quad x^{(k)} \sim q, \tag{3}$$

where q is some tractable distribution. This yields us unbiased estimates of a PoT's partition function, $\hat{Z}$, and unbiased estimates of the likelihood of the OLM, $\hat{p}_{\text{PoT}}(x)$. However, using *Jensen's inequality*, it is easy to see that this leads to biased estimates of each model's log-likelihood,

$$E\left[\log\left(\frac{p^*_{\text{PoT}}(x)}{\hat{Z}}\right)\right] \geq \log p_{\text{PoT}}(x), \tag{4}$$

$$E\left[\log \hat{p}_{\text{OLM}}(x)\right] \leq \log p_{\text{OLM}}(x). \tag{5}$$

Because the partition function is in the denominator, the estimates are biased in different directions. The behavior of both estimates as a function of the number of samples is shown in Figure 1.

The number of samples and intermediate distributions required for the annealed importance sampling (AIS) procedure to converge to a stable estimate depends on the distribution being sampled and the transition operator being used to implement MCMC. Since the partition function needs to be evaluated only once, we can choose the parameters more generously in this case than for evaluating the OLM.

3 GSM with separately modeled DC component

Here we briefly describe the model corresponding to "GSM" in Figure 4 of the main text. We first used the discrete cosine transform to separate the DC component,

$$x_{\text{DC}} = \frac{1}{\sqrt{N}} \sum_i x_i,$$

from the remaining components $x_\perp \in \mathbb{R}^{N-1}$ of an image patch x. The DC component was modeled using a univariate mixture of Gaussian with 20 components, p_{DC}, while $x_\perp$ was modeled with a multivariate Gaussian scale mixture, p_{GSM}. That is, the joint distribution of the model is given by

$$p(x) \propto p_{\text{DC}}(x_{\text{DC}}) p_{\text{GSM}}(x_\perp). \tag{6}$$