GP-select: Accelerating EM using adaptive subspace preselection

Jacquelyn A. Shelton[1], **Jan Gasthaus**[2, 3], **Zhenwen Dai**[4], **Jörg Lücke**[5], **and Arthur Gretton**[2]

[1] Technical University Berlin, Marchstrasse 23, 10587 Berlin, Germany.

[2] University College London, Gatsby Unit, 25 Howland Street, London W1T 4JG, UK.

[3] Amazon Development Center, Karl-Liebknecht-Str. 5, 10178 Berlin, Germany.

[4] University of Sheffield, Western Bank, Sheffield, South Yorkshire S10 2TN, UK.

[5] University of Oldenburg, Ammerländer Heerstr. 114, 26129 Oldenburg, Germany.

Keywords: Approximate inference, generative graphical models, latent variable models, Expectation Maximization, EM acceleration, variable preselection

Abstract

We propose a nonparametric procedure to achieve fast inference in generative graphical models when the number of latent states is very large. The approach is based on iterative latent variable preselection, where we alternate between learning a 'selection function'

to reveal the relevant latent variables, and using this to obtain a compact approximation of the posterior distribution for EM; this can make inference possible where the number of possible latent states is e.g. exponential in the number of latent variables, whereas an exact approach would be computationally infeasible. We learn the selection function entirely from the observed data and current EM state via Gaussian process regression. This is by contrast with earlier approaches, where selection functions were manually-designed for each problem setting. We show that our approach performs as well as these bespoke selection functions on a wide variety of inference problems: in particular, for the challenging case of a hierarchical model for object localization with occlusion, we achieve results that match a customized state-of-the-art selection method, at a far lower computational cost.

1 Introduction

Inference in probabilistic graphical models can be challenging in situations where there are a large number of hidden variables, each of which may take on one of several state values. The Expectation Maximization (EM) algorithm is widely applied for inference when hidden variables are present, however inference can quickly become intractable as the dimensionality of hidden states increases.

Expectation truncation (ET) (Lücke and Eggert, 2010) is a meta algorithm for accelerating EM learning in graphical models, which restricts the inference performed during the E-step to an "interesting" subset of states of the latent variables, chosen per data point according to a *selection function*. This subspace reduction can lead to a sig-

nificant decrease in computation with very little loss of accuracy (compared with the full model). In previous work, functions to select states of high posterior mass were derived individually for each model of interest, e.g. by taking upper bounds or noiseless limits. (Lücke and Eggert, 2010; Shelton et al., 2012; Bornschein et al., 2013; Sheikh et al., 2014). The crucial underlying assumption remains that when EM has converged, the posterior mass is concentrated in small volumes of the latent state space. This property is observed to hold in many visual data, auditory data, and general pattern recognition settings.

For more complex graphical models, notably the hierarchical model of Dai and Lücke (2014), the design of suitable selection functions is extremely challenging: it requires both expert knowledge on the problem domain and considerable computational resources to implement (indeed, the design of such functions for particular problems has been a major contribution in previous work on the topic).

In the present work, we propose a generalization of the ET approach, where we avoid completely the challenge of problem-specific selection function design. Instead, we learn selection functions adaptively and nonparametrically from the data, while learning the model parameters simultaneously using EM. We emphasize that the selection function is used only to "guide" the underlying base inference algorithm to regions of high posterior probability, but is not itself used as an approximation to the posterior distribution. As such, the learned function does not have to be a completely accurate indication of latent variable predictivity, as long as the relative importance of states is preserved. We use Gaussian process regression (Rasmussen and Williams, 2005) to learn the selection function, though other regression techniques could also be applied.

The main advantage of GPs is that they have an analytic one-shot learning procedure, and that learning different functions based on the same inputs is computationally cheap, which makes adaptive learning of a changing target function efficient. We term this part of our approach *GP-select*. Our nonparametric generalization of ET may be applied as a black-box meta algorithm for accelerating inference in generative graphical models, with no expert knowledge required.

Our approach is the first to make ET a general purpose algorithm for discrete latents,whereas previously new versions of ET were required for every new latent variable model addressed. For instance, in Section 5.3 we will show that preselection is crucial for efficient inference in complex models. Although ET has already been successful in some models, this work shows that more complex models will crucially depend on an improved selection step and focuses on automating this step.

For empirical evaluation, we have applied GP-select in a number of experimental settings. First, we considered the case of sparse coding models (binary sparse coding, spike-and-slab, nonlinear spike-and-slab), where the relationship between the observed and latent variables is known to be complex and nonlinear.[1] We show that GP-select can produce results with equal performance to the respective manually-derived selection functions. Interestingly, we find it can be essential to use nonlinear GP regression in the spike-and-slab case, and that simple linear regression is not sufficiently flexible in modeling the posterior shape. Second, we illustrate GP-select on a simple Gaussian mixture model, where we can explicitly visualize the form of the learned regression

[1] Note that even when linear relations exist between the latents and outputs, a nonlinear regression may still be necessary in finding relevant variables, as a result of explaining away.

function. We find that even for a simple model, it can be be essential to learn a non-linear mapping. Finally, we present results for a recently published hierarchical model for translation invariant occlusive components analysis (Dai and Lücke, 2014). The performance of our inference algorithm matches that of the complex hand-engineered selection function of the previous work, while being straightforward to implement and having a far lower computational cost.

2 Related work

The general idea of aiding inference in graphical models by learning a function that maps from the observed data to a property of the latent variables is quite old. Early work includes the Helmholtz machine (Dayan et al., 1995) and its bottom-up connections trained using the wake-sleep algorithm (Hinton et al., 1995). More recently, the idea has surfaced in the context of learning variational distributions with neural networks (Kingma and Welling, 2014). A two-stage inference has been discussed in the context of computer vision (Yuille and Kersten, 2006) and neural inference (Körner et al., 1999). Recently, researchers (Mnih and Gregor, 2014) have generalized this idea to learning in arbitrary graphical models by training an "inference network" that efficiently implements sampling from the posterior distribution.

GPs have recently been widely used to "learn" the results of complicated models in order to accelerate inference and parameter selection. GP approximations have been used in lieu of solving complex partial differential equations (Sacks et al., 1989; Currin et al., 1991), to learn data-driven kernel functions for recommendation sys-

tems (Schwaighofer et al., 2004), and recently for quantum chemistry (Rupp et al., 2012). Other work has used GPs to simplify computations in approximate Bayesian computation (ABC) methods: namely to model the likelihood function for inference (Wilkinson, 2014), to aid in making Metropolis-Hastings (MH) decisions (Meeds and Welling, 2014), and to model the discrepancies between simulated/observed data parameter space simplification (Gutmann and Corander, 2015). Recently, instead of the typical choice of GPs for large scale Bayesian optimization, neural networks have been used to learn an adaptive set of basis functions for Bayesian linear regression (Snoek et al., 2015).

Our work follows the same high level philosophy in that we use GPs to approximate complex/intractable probabilistic models. None of the cited prior work address our problem setting, namely the selection of relevant latent variables by learning a non-parametric relevance function, for use in expectation truncation (ET).

3 Variable selection for accelerated inference

Notation. We denote the observed data by the $D \times N$ matrix $\mathbf{Y} = (\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(N)})$, where each vector $\mathbf{y}^{(n)} = (y_1^{(n)}, \ldots, y_D^{(n)})^{\mathrm{T}}$ is the nth observation in a D-dimensional space. Similarly we define corresponding binary latent variables by the matrix $\mathbf{S} = (\mathbf{s}^{(1)}, \ldots, \mathbf{s}^{(N)}) \in \{0,1\}^{H \times N}$ where each $\mathbf{s}^{(n)} = (s_1^{(n)} \ldots, s_H^{(n)})^{\mathrm{T}} \in \{0,1\}^H$ is the nth vector in the H-dimensional latent space, and for each individual hidden variable $h = 1, \ldots, H$, the vector $\mathbf{s}_h = (s_h^{(1)} \ldots, s_h^{(N)}) \in \{0,1\}^N$. Reduced latent spaces are denoted by H', where $H' \ll H$. Note that although we restrict ourselves to binary

latent variables here, the procedure could in principle be generalized to variables with higher cardinality (Exarchakis et al., 2012, e.g. see). We denote the prior distribution over the latent variables with $p(\mathbf{s}|\theta)$ and the likelihood of the data with $p(\mathbf{y}|\mathbf{s},\theta)$. Using these expressions, the posterior distribution over latent variables is

$$p(\mathbf{s}^{(n)}|\mathbf{y}^{(n)},\Theta) = \frac{p(\mathbf{s}|\Theta)\, p(\mathbf{y}^{(n)}|\mathbf{s}^{(n)},\Theta)}{\sum_{\mathbf{s}'^{(n)}} p(\mathbf{s}'|\Theta)\, p(\mathbf{y}|\mathbf{s}',\Theta)}. \tag{1}$$

3.1 Selection via Expectation Truncation in EM

Expectation Maximization (EM) is an iterative algorithm to optimize the model parameters of a given graphical model (see e.g. (Dempster et al., 1977; Neal and Hinton, 1998)). EM iteratively optimizes a lower bound on the data likelihood by inferring the posterior distribution over hidden variables given the current parameters (the E-step), and then adjusting the parameters to maximize the likelihood of the data averaged over this posterior (the M-step). When the number of latent states to consider is large (e.g. exponential in the number of latent variables), the computation of the posterior distribution in the E-step becomes intractable and approximations are required.

Expectation truncation (ET) is a meta algorithm, which improves convergence of the expectation maximization (EM) algorithm (Lücke and Eggert, 2010). The main idea underlying ET is that the posterior probability mass is concentrated in a small subspace of the full latent space. This is the case, for instance, if for a given data point $\mathbf{y}^{(n)}$ only a subset of the H latent variables $s_h^{(n)}$ are relevant. Note, however, that the posterior may still be concentrated even when all latents are relevant, since most of the probability mass may be concentrated on few of these.

A *selection function* can be used to identify a subset of salient variables, denoted by H' where $H' \ll H$, which in turn is used to define a subset, denoted $\mathcal{K}_n$, of the possible state configurations of the space per data point. State configurations not in this space (of variables deemed to be non-relevant) are fixed to 0 (assigned zero probability mass). The posterior distribution (1) can then be approximated by a truncated posterior distribution, computed on the reduced support,

$$\begin{aligned} & p(\mathbf{s}^{(n)}|\mathbf{y}^{(n)}, \Theta) \\ & \quad \approx q_n(\mathbf{s}^{(n)}; \Theta) = \frac{p(\mathbf{s}^{(n)}, \mathbf{y}^{(n)}|\,\Theta)\,\delta(\mathbf{s}^{(n)} \in \mathcal{K}_n)}{\sum\limits_{\mathbf{s}'^{(n)} \in \mathcal{K}_n} p(\mathbf{s}'^{(n)}, \mathbf{y}^{(n)}|\,\Theta)}, \end{aligned} \tag{2}$$

where $\mathcal{K}_n$ contains the latent states of the H' relevant variables for data point $\mathbf{y}^{(n)}$, and $\delta(\mathbf{s} \in \mathcal{K}_n) = 1$ if $\mathbf{s} \in \mathcal{K}_n$ and zero otherwise. In order words, Eq. (2) is proportional to Eq. (1) if $s \in \mathcal{K}_n$ (and zero otherwise). The set $\mathcal{K}_n$ contains only states for which $s_h = 0$ for all h that are not selected, i.e. all states where $s_h = 1$ for non-selected h are assigned zero probability. The sum over $\mathcal{K}_n$ is much more efficient than the sum for the full posterior, since it need only be computed over the reduced set of latent variable states deemed relevant: the state configurations of the irrelevant variables are fixed to be zero. The variable selection parameter H' is selected based on compute resources available: i.e. as large as resources allow in order to be closer to true EM, although empirically it's been shown that much smaller values suffice (see e.g. Sheikh et al., 2014, App. B on complexity-accuracy trade-offs).

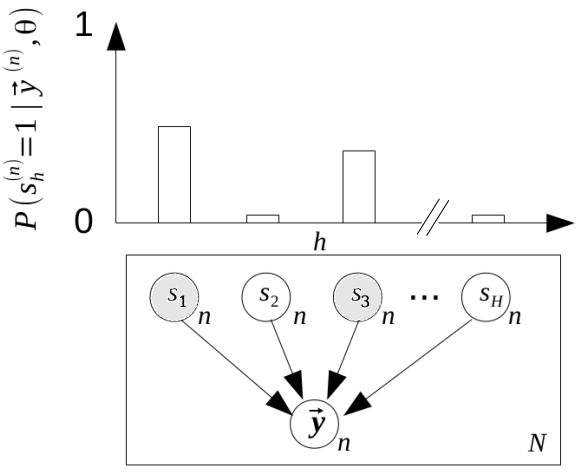

Figure 1: Illustration of the affinity function for selection. The affinity approximates the marginal posterior probability of each $h = 1, \ldots, H$ latent variable (top), which corresponds to the most relevant variables for a given data point $\mathbf{y}^{(n)}$ (bottom). Here, the affinity would return variables s_1 and s_3 as being relevant for $\mathbf{y}^{(n)}$.

3.2 ET with affinity

One way of constructing a selection function is by first ranking the latent variables according to an *affinity function* $f_h(\mathbf{y}^{(n)}) : \mathrm{R}^D \mapsto \mathrm{R}$ which directly reflects the relevance of latent variable s_h. A natural choice for such a function is the one that approximates the marginal posterior probability of each variable, e.g. we try to learn f as follows:

$$f_h(\mathbf{y}^{(n)}) = \hat{p}_h^{(n)} \approx p_h^{(n)} \equiv p(s_h^{(n)} = 1 | \mathbf{y}^{(n)}, \Theta), \tag{3}$$

meaning that, for the relevant variables, the marginal posterior probability p_h exceeds some threshold. See Figure 1 for a simplified illustration. When the latent variables $s_{h=1}^{(n)}, \ldots, s_H^{(n)}$ in the marginal posterior probability $\hat{\mathbf{p}}^{(n)} = \hat{p}_{h=1}^{(n)}, \ldots, \hat{p}_H^{(n)}$ are conditionally independent given a data point $\mathbf{y}^{(n)}$, this affinity function correctly isolates the most relevant variables in the posterior. Even when this strong assumption does not

hold in practice (which is often the case), however, the affinity can still correctly highlight relevant variables, and has been empirically shown to be quite effective even when dependencies exist (see e.g. source separation tasks in (Sheikh et al., 2014)).

Next, using all $\hat{p}_{h=1}^{(n)}, \ldots, \hat{p}_{H}^{(n)}$ from the affinity function $\mathbf{f}(\mathbf{y}^{(n)}) = (f_1(\mathbf{y}^{(n)}), \ldots, f_H(\mathbf{y}^{(n)}))$, we define $\gamma\left(\hat{\mathbf{p}}^{(n)}\right)$ to simultaneously sort the indices of the latent variables in descending order and *reduce* the sorted set to the H' highest (most relevant) variables' indices. To ensure that there is a non-zero probability of selecting each variable per EM iteration, 10% of the H' indices are uniformly chosen from H at random. This prevents the possible propagation of errors from $q_{(n)}$ continuously assigning small probabilities to a variable s_h in early EM iterations. $\gamma(\hat{\mathbf{p}}^{(n)})$ thus returns the H' selected variable indices I deemed by the affinity to be relevant to the nth data point. Finally, using the indices I from γ, we define $\mathcal{I}(I)$ to return an H'-dimensional subset of selected relevant latent states $\mathcal{K}_n$ for each data point $\mathbf{y}^{(n)}$. All 'non-relevant' variable states s_h for all variables $h \notin I$ are effectively set to 0 in Equation (2) by not being present in the state set $\mathcal{K}_n$.

Using $\mathbf{f}$, $\mathcal{I}$, and γ, we can define a *selection function* $\mathcal{S} : \mathrm{R}^D \mapsto 2^{\{1,\ldots,H\}}$ to select subsets $\mathcal{K}_n$ per data point $\mathbf{y}^{(n)}$. Again, the goal is for the states $\mathcal{K}_n$ to contain most of the probability mass $p(\mathbf{s} \,|\, \mathbf{y})$ and to be significantly smaller than the entire latent space. The *affinity based selection function* can be expressed as

$$\mathcal{S}(\mathbf{y}^{(n)}) \;=\; \mathcal{I}\left[\gamma\left[\mathbf{f}(\mathbf{y}^{(n)})\right]\right] \;=\; \mathcal{K}_n. \tag{4}$$

3.3 Inference in EM with selection

In each iteration of EM, the following occurs: prior to the E-step, the selection function $\mathcal{S}(\mathbf{y}^{(n)})$ in (4) is computed to select the most relevant states $\mathcal{K}_n$, which are then used to

compute the truncated posterior distribution $q_n(\mathbf{s})$ in (2). The truncated posterior can be computed using any standard inference method, such as exact inference or e.g. Gibbs sampling from $q(\mathbf{s})$ if inference is still intractable or further computational reduction is desired. The result of the E-step is then used to update the model parameters in the M-step.

4 GP-Select

In previous work, the selection function $\mathcal{S}(\mathbf{y}^{(n)})$ was a deterministic function derived individually for each model (see e.g. Shelton et al., 2011, 2012; Dai and Lücke, 2012a,b; Bornschein et al., 2013; Sheikh et al., 2014; Shelton et al., 2015). We now generalize the selection approach: instead of predefining the form of $\mathcal{S}$ for variable selection, we want to learn it in a black-box and model-free way based on the data. We learn $\mathcal{S}$ using Gaussian process (GP) regression (e.g. Rasmussen and Williams, 2005), which is a flexible nonparametric model and scales cubicly[2] with the number of data points N but linearly with the number of latent variables H. We define the affinity function f_h as being drawn from a Gaussian process model: $f_h(\mathbf{y}^{(n)}) \sim \mathrm{GP}\left(0,\, k(\cdot,\cdot)\right)$, where $k(\cdot,\cdot)$ is the covariance kernel, which can be flexibly parameterized to represent the relationship between variables. Again, we use f_h to approximate the marginal posterior probability p_h that $s_h^{(n)} = 1$. A nice property of Gaussian processes is that the kernel matrix K need only be computed once (until the kernel function hyperparameters are updated) to approximate $p_h^{(n)}$ for the entire $H \times N$ set of latent variables $\mathbf{S}$.

[2] If the scaling with N is still too expensive, an incomplete Cholesky approximation is used, with cost linear in N and quadratic in the rank Q of the approximation (see Section 5.3 for details).

Thus, prior to each E-step in each EM iteration, within each calculation of the selection function, we calculate the affinity using a GP to regress the expected values of the latent variables $\langle \mathbf{S} \rangle$ onto the observed data $\mathbf{Y}$. Specifically, we train on p_h from the previous EM iteration (where p_h is equal to $\langle s_h \rangle$), for training data of $\mathcal{D} = \{(\mathbf{y}^{(n)}, \langle \mathbf{s}^{(n)} \rangle_q) | n = 1, \ldots, N\}$, where we recall that $q_n(\mathbf{s}^{(n)})$ is the approximate posterior distribution for $\mathbf{s}^{(n)}$ in Eq. (2). In the first EM iteration, the expectations $\langle \mathbf{s}^{(n)} \rangle_q$ are initialized randomly; in each subsequent EM iteration, the expectations w.r.t. the $\mathcal{K}_n$-truncated posterior $q(\mathbf{s})$ are used. The EM algorithm is run for T iterations and the hyperparameters of the kernel are optimized by maximum likelihood every T^* EM iterations.

For each data point n and latent variable h we compute the predicted mean of the GP by leaving this data point out of the training set and considering all others, which is called leave-one-out (LOO) prediction. It can be shown that this can be implemented efficiently (see Section 5.4.2 in Rasmussen and Williams, 2005), and we use this result to update the predicted affinity as follows:

$$\hat{p}_h^{(n)} \leftarrow \langle s_h^{(n)} \rangle_{q_n} - \frac{[K^{-1} \langle \mathbf{s}_h \rangle_{q_n}]_{nn}}{[K^{-1}]_{nn}}. \tag{5}$$

Equation (5) can be efficiently implemented for all latent variables $h = 1, \ldots, H$ and all data points $n = 1, \ldots, N$ using matrix operations, thereby requiring only one kernel matrix inversion for the entire dataset.

Substituting Equation (5) for $\mathbf{f}$ in the affinity based selection function (4), we call the entire process *GP-select*. An outline is shown in Algorithm 1.

Algorithm 1 GP-Select to accelerate inference in Expectation Maximization

for EM iterations $t = 1, \ldots, T$ **do**

for data point $n = 1, \ldots, N$ **do**

compute affinity of all latent variables $\hat{\mathbf{p}}_t^{(n)}$: (5)

compute subset of relevant states $\mathcal{S}$: (4)

compute truncated posterior $q_{n,t}(\mathbf{s}^{(n)})$, E-step: (2)

update model parameters in M-step

store $\langle s_h^{(n)} \rangle_{q_{n,t}}$ for $\mathbf{p}^{(n)}$ in EM iteration $t+1$

end for

optimize kernel hyperparams every T^* EM iterations

end for

5 Experiments

We apply our GP-select inference approach to five different probabilistic generative models. First, we considered three sparse coding models (binary sparse coding, spike-and-slab, nonlinear spike-and-slab), where the relationship between the observed and latent variables is known to be complex and nonlinear. Second, we apply GP-select to a simple Gaussian mixture model, where we can explicitly visualize the form of the learned regression function. Finally, we apply our approach to a recent hierarchical model for translation invariant occlusive components analysis (Dai and Lücke, 2012a; Dai et al., 2013; Dai and Lücke, 2014).

5.1 Sparse coding models

Using hand-crafted functions to preselect latent variables, a variety of sparse coding models have been successfully scaled to high-dimensional latent spaces with use of selection (Henniges et al., 2010; Bornschein et al., 2013; Sheikh et al., 2014) and selection with Gibbs sampling (Shelton et al., 2011, 2012, 2015) inference approaches. In order to demonstrate our method, we consider three of these sparse generative models, and perform inference with our GP-select approach instead of a hand-crafted selection function. The models are:

A. *Binary sparse coding*:

$$\text{latents:} \quad \mathbf{s} \sim Bern(\mathbf{s}|\pi) = \textstyle\prod_{h=1}^{H} \pi^{s_h} \big(1-\pi\big)^{1-s_h},$$

$$\text{observations:} \quad \mathbf{y} \sim \mathcal{N}(\mathbf{y}; W\mathbf{s}, \sigma^2 I),$$

where $W \in \mathrm{R}^{D\times H}$ denotes the dictionary elements and π parameterizes the sparsity (see e.g. (Henniges et al., 2010)).

B. *Spike-and-slab sparse coding*:

$$\text{latents:} \quad \mathbf{s} = \mathbf{b} \odot \mathbf{z} \quad \text{where} \quad \mathbf{b} \sim Bern(\mathbf{b}|\pi) \text{ and } \mathbf{z} \sim \mathcal{N}(\mathbf{z};\, \mu, \Sigma_h),$$

$$\text{observations:} \quad \mathbf{y} \sim \mathcal{N}(\mathbf{y}; W\mathbf{s}, \sigma^2 I)$$

where the point-wise multiplication of the two latent vectors, i.e., $(\mathbf{s}\odot\mathbf{z})_h = s_h\, z_h$ generates a 'spike-and-slab' distributed variable ($\mathbf{s}\odot\mathbf{z}$), that has either continuous values or exact zero entries (e.g. (Titsias and Lázaro-Gredilla, 2011; Goodfellow et al., 2013; Sheikh et al., 2014)).

C. *Nonlinear Spike-and-slab sparse coding*:

$$\text{latents:} \quad \mathbf{s} = \mathbf{b} \odot \mathbf{z} \quad \text{where} \quad \mathbf{b} \sim Bern(\mathbf{b}|\pi) \text{ and } \mathbf{z} \sim \mathcal{N}(\mathbf{z};\, \mu, \Sigma_h),$$

$$\text{observations:} \quad \mathbf{y} \sim \mathcal{N}(\mathbf{y}; \max_h\{s_h \mathbf{W}_h\}, \sigma^2 I)$$

for which the mean of the Gaussian for each $\mathbf{y}^{(n)}$ is centered at $\max_h\{s_h \mathbf{W}_h\}$, where $\max_h$ is a nonlinearity that considers all H latent components and takes the h yielding the maximum value for $s_h \mathbf{W}_h$ (Lücke and Sahani, 2008; Shelton et al., 2012; Bornschein et al., 2013; Shelton et al., 2015), instead of centering the data at the linear combination of $\sum_h s_h \mathbf{W}_h = W\mathbf{s}$.

In the above models, inference with the truncated posterior of Equation (2) using manually-derived selection functions has yielded results as good as or more robust than exact inference (converging less frequently to local optima than exact inference; see earlier references for details). For models **A** and **C**, the selection function was the cosine similarity between the weights $\mathbf{W}_h$ (e.g. dictionary elements, components, etc.) associated with each latent variable s_h and each data point $\mathbf{y}^{(n)}$: $\mathcal{S}_h(\mathbf{y}^{(n)}) = (\mathbf{W}_h^{\mathrm{T}} \,/\, ||\mathbf{W}_h||)\, \mathbf{y}^{(n)}$. For model **B**, the selection function was the data likelihood given a singleton state: $\mathcal{S}_h(\mathbf{y}^{(n)}) = p(\mathbf{y}^{(n)}|\mathbf{s} = \mathbf{s}_h, \Theta)$, where $\mathbf{s}_h$ represents a singleton state in which only the entry h is non-zero.

We generate $N = 2,000$ data points consisting of $D = 5 \times 5 = 25$ observed dimensions and $H = 10$ latent components according to each of the models **A-C**: N images of randomly selected overlapping 'bars' with varying intensities for models **B** and **C**, and additive Gaussian noise parameterized by ground-truth $\sigma^2 = 2$ (e.g. following the spike-and-slab prior). On average, each data point contains 2 bars, and we choose $H' = 5$. With this choice, we can select sufficiently many latents for virtually

all data points.

For each of the models considered, we run 10 repetitions of each of the following set of experiments: (1) selection using the respective hand-crafted selection function, (2) GP-select using a linear covariance kernel, (3) GP-select using an RBF covariance kernel, and (4) GP-select using a kernel composed by adding the following kernels: RBF, linear, bias and white noise kernels, which we will term the *composition kernel*. As hyperparameters of kernels are learned, the composition kernel (4) can adapt itself to the data and only use the kernel components required. See Rasmussen and Williams (2005, Chapter 4, Secton 4.2.4) for a discussion on kernel adaptation. Kernel parameters were model-selected via maximum marginal likelihood every 10 EM iterations. For models **A** and **B**, inference was performed exactly using the truncated posterior (2), but as exact inference is analytically intractable in model **C**, inference was performed by drawing Gibbs samples from the truncated space (Shelton et al., 2011, 2012, 2015). We run all models until convergence.

Results are shown in Figure 2. In all experiments, the GP-select approach was able to infer ground-truth parameters as well as the hand-crafted function. For models where the cosine similarity was used (in **A** and **C**), GP regression with a linear kernel quickly learned the ground-truth parameters, and hence fewer iterations of EM were necessary. In other words, even without providing GP-select explicit weights W as required for the hand-crafted function, its affinity function using GP regression (5) learned a similar enough function to quickly yield identical results. Furthermore, in the model with a less straight-forward hand-crafted function (in the spike-and-slab model of **B**), only GP regression with an RBF kernel was able to recover ground-truth parameters. In this case

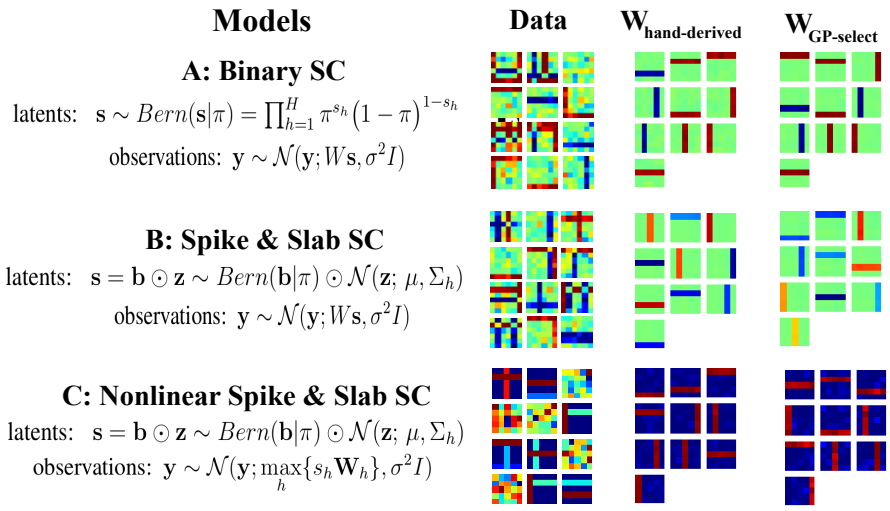

Figure 2: Sparse coding models results comparing GP-select with a successful hand-derived selection function. Results are shown on artificial ground-truth data with $H = 10$ latent variables and $H' = 5$ preselected variables for: **A** Binary sparse coding, **B** Spike-and-slab sparse coding, and **C** Nonlinear spike-and-slab sparse coding. First column: Example data points $\mathbf{y}^{(n)}$ generated by each of the models. Middle column: Converged dictionary elements W learned by the hand-crafted selection functions. Third column: Converged dictionary elements W learned by GP-select with $H' = 5$ using the kernel with best performance (matching that of inference with hand-crafted selection function). In all cases, the model using the GP-select function converged to the ground-truth solution, just as the hand-crafted selection functions did.

(model **B**), GP-select using an RBF kernel recovered the ground-truth 'bars' in 7 out of 10 repetitions, whereas the hand-crafted function recovered the bases in 8 instances. For the remaining models, GP-select converged to the ground-truth parameters with the same average frequency as the hand-crafted functions.

Finally, we have observed empirically that the composition kernel is flexible enough to subsume all other kernels: the variance of the irrelevant kernels dropped to zero in simulations. This suggests the composition kernel is a good choice for general use.

5.2 Gaussian mixture model

Next, we apply GP-select to a simple example, a Gaussian mixture model, where the flexibility of the approach can be easily and intuitively visualized. The model of the data likelihood is

$$p(\mathbf{y}^{(n)}|\mu_{\mathbf{c}}, \sigma_{\mathbf{c}}, \pi) = \sum_{c=1}^{C} \mathcal{N}(\mathbf{y}^{(n)}; \mu_{\mathbf{c}}, \sigma_{\mathbf{c}})\, \pi_c, \tag{6}$$

where C is the number of mixture components; the task is to assign each data point to its latent cluster.

The training data used for GP regression was $\mathcal{D} = \{(\mathbf{y}^{(n)}, \langle s_h^{(n)} \rangle_{q_n}) | n = 1, \ldots, N\}$, where the targets were the expected cluster responsibilities (posterior probability distribution for each cluster) for all data points, $\langle s_h \rangle_q$, and we use one-hot encoding for cluster identity. With this, we apply our GP-select approach to this model, computing the selection function according to Equation (4) with affinity f defined by GP regression (5) and following the approximate EM approach as in the previous experiments. In these experiments we consider two scenarios for EM learning of the data likelihood in Equation (6): GP-select with an RBF covariance kernel and a linear covariance kernel. We do not include the composition kernel suggested (based on experiments) in Section 4.1, as the goal of the current experiments is to show the effects of using the 'wrong' kernel. These effects would further support the use of the flexible composition kernel in general, as it can subsume both kernels considered in the current experiments (RBF

and linear).

To easily visualize the output, we generate 2-dimensional observed data ($\mathbf{y}^{(n)} \in \mathrm{R}^{D=2}$) from $C = 3$ clusters – first with randomly assigned cluster means, and second such that the means of the clusters lie roughly on a line. In the GP-select experiments, we select $C' = 2$ clusters from the full set, and run 40 EM iterations for both kernel choices (linear and RBF). Note that for mixture models, the notation of C' selected clusters of the C set is analogous to the H' selected latent variables from the H full set, as described in the non-mixture model setting, and the GP-select algorithm proceeds unchanged. We randomly initialize the variance of the clusters $\sigma_{\mathbf{c}}$ and initialize the cluster means $\mu_{\mathbf{c}}$ at randomly selected data points. Results are shown in Figure 3.

With cluster parameters initialized randomly on these data, the linear GP regression prediction cannot correctly assign the data to their clusters (as seen in Figure 3**B**), but the nonlinear approach successfully and easily finds the ground-truth clusters (Figure 3**A**). Furthermore, even when both approaches were initialized in the optimal solution, the cluster assignments from GP-select with a linear kernel quickly wandered away from the optimal solution and were identical to random initialization, converging to the same result shown in iteration 20 of Figure 3**B**). The RBF kernel cluster assignments remained at the optimal solution even with number of selected clusters set to $C' = 1$.

These experiments demonstrate that the selection function needs to be flexible even for very simple models, and that nonlinear selection functions are an essential tool even in such apparently straightforward cases.

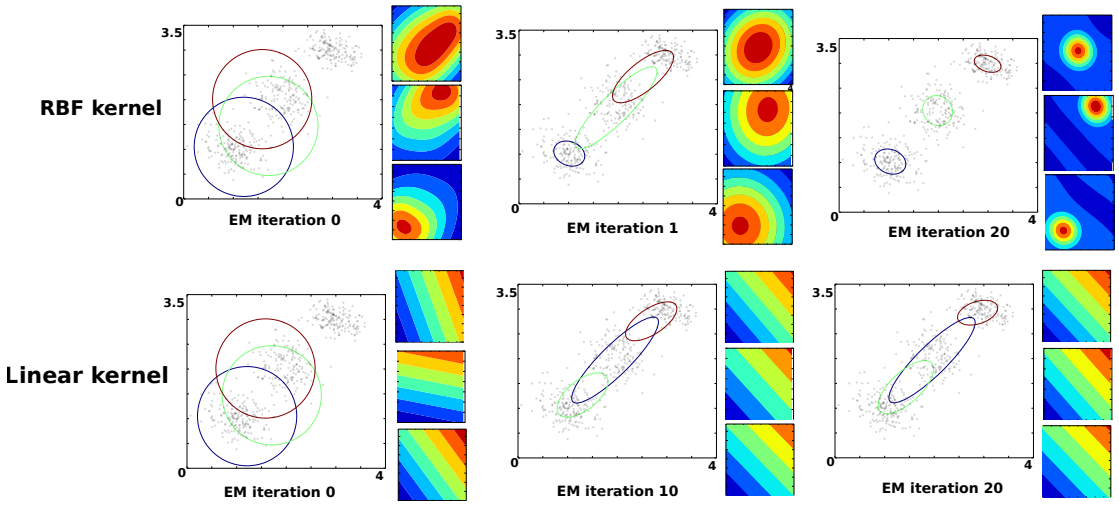

Figure 3: Gaussian mixture model results using GP-select (selection of $C' = 2$ in a $C = 3$ class scenario) for inference. Progress of the inference is shown using (row one) an RBF covariance kernel in the regression, and (row two) a linear covariance kernel. For each iteration shown, we see (1) the observed data and their inferred cluster assignments and (2) the C corresponding GP regression functions learned/used for GP-select in that iteration. Different iterations are pictured due to different convergence rates. As shown, inference with GP-select using a linear kernel is unable to assign the data points to the appropriate clusters, whereas GP-select with an RBF kernel succeeds.

5.3 Translation Invariant Occlusive models

Now that we have verified that GP-select can be applied to various generative graphical models and converge to ground-truth parameters, we consider a more challenging model that addresses a problem in computer vision: *translations of objects in a scene*.

Model. Translation invariant models are particularly difficult to optimize because they must consider a massive latent variable space: evaluating multiple objects and locations in a scene leads a *latent space complexity of the number of locations expo-*

nentiated by the number of objects. Inference in such a massive latent space heavily relies on the idea of variable selection to reduce the number of candidate objects and locations. In particular, hand-engineered selection functions that consider *translational invariance* have been successfully applied to this type of model (Dai and Lücke, 2012b, 2014; Dai et al., 2013). The selection function used so far for reducing latent space complexity in this model has been constructed as follows: first, the candidate locations of all the objects in the model are predicted, and then a subset of candidate objects that might appear in the image are selected according to the predicted locations. Next, the subset of states $\mathcal{K}_n$ is constructed according to the combinatorics of the different locations of all the candidate objects. The posterior distribution is then computed following Equation (2).

This selection system is very costly: the selection function has parameters which need to be hand-tuned, e.g., the number of representative features, and it needs to scan through the entire image, considering all possible locations, which becomes computationally demanding for large-scale experiments. In translation invariant models, instead of predicting the existence of a component, the selection function has to predict all possible locations a component could be. To maximally exploit the capabilities of GP-selection function, we directly use the GP regression model to *predict the possible locations* of a component *without introducing any knowledge of translation invariance* into the selection function. In this work, a GP regression model is fitted from the input image to marginal posterior probabilities of individual components appearing at all possible locations. Therefore, the input to the GP-selection function is the image to be inferred and the output is a score for each possible location of each component in the

Figure 4: COIL Dataset (Nene et al., 1996): A handful of data points used in experiments with the Translation Invariant Occlusive (InvECA) model, showing the occluding objects to be learned.

model. For example, when learning 10 components in a $D = 30 \times 30$ pixel image patch, the output dimensionality of GP-select is 9000. This task is computationally feasible, since GP models scale linearly with output dimensionality. The inference of components' locations with GP-select is significantly faster than the selection function in the original work, as it avoids explicitly scanning through the image.

Although there are additional computations necessary for an automatic selection function like GP-select, for instance due to the adjustment of its parameters, there are many options to reduce computational costs. First, we may approximate the full $N \times N$ Gram matrix by an incomplete Cholesky approximation (Fine and Scheinberg, 2001) resulting in a cost of $O(N \times Q)$, where $Q << N$ is the rank of the Cholesky approximation. Second, we may reduce the frequency of kernel hyperparameter updates to only every T^* EM iterations, where a $T^* > 1$ represents a corresponding computation reduction. The combination of the Cholesky approximation plus infrequent updates will have the following benefits: a factor of five speedup for infrequent updates, and a factor of $(N - Q)^2$ speedup from incomplete Cholesky, where Q is the rank of the Cholesky approximation and N is the number of original data points.

COIL Dataset. We apply our GP-selection function to the *Invariant Exclusive*

Component Analysis (InvECA) model (Dai and Lücke, 2012b; Dai et al., 2013). For our experiments, we consider an image dataset used in previous work: data were generated using objects from the COIL-100 image dataset (Nene et al., 1996), taking 16 different objects, downscaled to $D = 10 \times 10$ pixels and segmented out from the black background. A given image was generated by randomly selecting a subset of the 16 objects, where each object has a probability of 0.2 of appearing. The appearing objects were placed at random positions on a 30×30 black image. When the objects overlap, they occlude each other with a different random depth order for each image. In total, $N = 2000$ images were generated in the dataset (examples shown in Figure 4).

The task of the InvECA model is to discover the visual components (i.e. the images of 16 objects) from the image set without any label information. We compare the visual components learned by using *four different selection functions in the InvECA model*: the hand-crafted selection function used in the original work by Dai and Lücke (2012b), GP-select updated every iteration, GP-select updated every $T* = 5$ iterations, and GP-select with incomplete Cholesky decomposition updated every iteration, or $T* = 1$ (in this manner we isolate the improvements due to Cholesky from those due to infrequent updates). In these experiments, the parameters of GP-select are optimized at the end of each $T*$ EM iteration(s), using a maximum of 20 gradient updates. The number of objects to be learned is $H = 20$ and the algorithm pre-selects $H' = 5$ objects for each data point. The kernel used was the composition kernel, as suggested in Section 4.1, although after fitting the hyperparameters only the RBF kernel remained with large variance (i.e. a linear kernel alone would not have produced good variable selection, thus the flexible composition kernel was further shown to be a good choice).

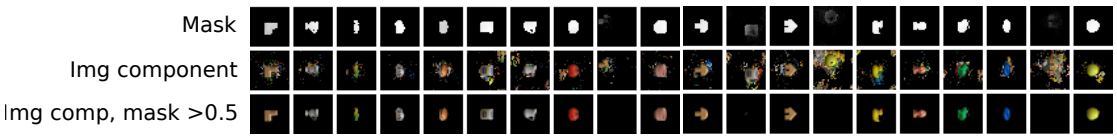

Figure 5: Image components and their masks learned by GP-select with the Translation Invariant model. GP-select learned all objects in the dataset. The first row shows the mask of each component, the second row shows the learned image components, and the third row shows only the area of the learned components that had a mask > 0.5.

Results. All four versions of the InvECA model using each of the selection functions considered successfully recover each individual objects in our modified COIL image set. The learned object representations with GP-select are shown in Figure 5. Four additional components developed into representations, however these all had very low mask values, allowing them to easily be distinguished from other true components.

Next, we compare the accuracy of the four selection functions. For this, we collected the object locations (pixels) indicated by each selection function after all EM iterations, applied the selection functions (for the GP selection functions, this was using the final function learned after all EM iterations) to the entire image dataset again, then compared these results with the ground-truth location of all of the objects in the dataset. The accuracy of the predicted locations was then computed by comparing the distance of all ground-truth object location to the location of the top candidate locations from each selection function. See Figure 6 for a histogram of these distances and the corresponding accuracy for all selection functions. Note that the percentages in the histogram are plotted in log scale. Also, as a baseline verification, we computed and compared the pseudo log likelihood (Dai et al., 2013) of the original selection function to the three

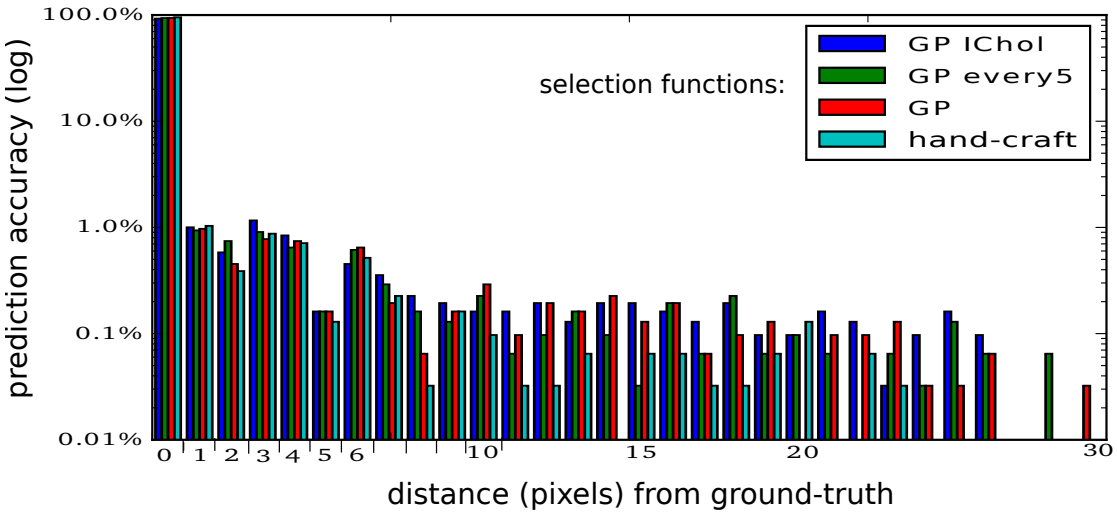

Figure 6: Prediction accuracy of the four selection functions in the InvECA model. Functions depicted in the figures: GP-select with no modifications (GP, red), the incomplete Cholesky decomposition (GP IChol, blue), with updated kernel hyperparameters every 5 EM iterations (GP every5, green), and with hand-crafted selection (hand-craft, cyan). Shown: the log-scale histogram of the prediction accuracy for the four selection functions, measured by the distance each function's predicted object location was to the ground-truth object location. All bars of the selection functions show very similar accuracy for the various distances.

GP-select based ones. The pseudo log likelihood for all selection functions is shown in Figure 7. Figures 6-7 show that all four selection functions can very accurately predict the locations of all the objects in the dataset – the GP-select selection functions yields no loss in inference performance in comparison to the original hand-engineered selection function. Even those using speed-considerate approximations (incomplete Cholesky decomposition of the kernel matrix (GP IChol) and updating kernel hyperparameters only every 5 EM iterations (GP every5)) have indistinguishable prediction accuracy on

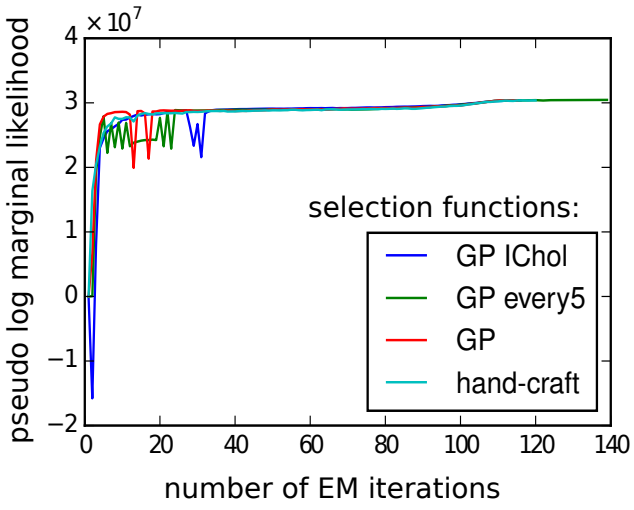

Figure 7: Baseline comparison of the four selection functions in the InvECA model. Functions depicted in the figures are identical to those in Figure 6. Shown: the convergence of the pseudo log marginal likelihood [of the model parameters learned at each EM iteration] for the four selection functions over all EM iterations. After about 40 EM iterations, all selection function versions of the algorithm converge to the same likelihood solution. Simultaneously, the GP-select approaches exhibit no loss of accuracy compared to the hand-crafted function, and 'GP IChol' represents a factor of 100 speedup vs. 'GP', and 'GP every5' represents a factor of 5 speedup.

the task.

An analysis of the benefits indicate that, as GP-select avoids explicitly scanning through the image, the time to infer the location of an object is significantly reduced compared to the hand-crafted function. GP-select requires 22.1 seconds on a single CPU core to infer the locations of objects across the whole image set, while the hand-crafted function requires 1830.9 seconds. In the original work, the selection function was implemented with GPU acceleration and parallelization. Although we must com-

pute the kernel hyperparameters for GP-select, it is important to note that the hyperparameters need not be fit perfectly each iteration – for the purposes of our approach, a decent approximation suffices for excellent variable selection. In this experiment, updating the parameters of GP-select with 10 gradient steps took about 390 seconds for the full-rank kernel matrix. When we compute the incomplete Cholesky decomposition while inverting the covariance matrix, compute time was reduced to 194 seconds (corresponding to the $(N - Q)^2$ speedup, where Q is the rank of the Cholesky approximation), with minimal loss in accuracy. Furthermore, when updating the GP-select hyperparameters only every 5 iterations, average compute time was reduced by another one fifth, again without loss in accuracy.

6 Discussion

We have proposed a means of achieving fast EM inference in Bayesian generative models, by learning an approximate selection function to determine relevant latent variables for each observed variable. The process of learning the relevance functions is interleaved with the EM steps, and these functions are used in obtaining an approximate posterior distribution in the subsequent EM iteration. The functions themselves are learned via Gaussian process regression, and do not require domain-specific engineering, unlike previous selection functions. In experiments on mixtures and sparse coding models with interpretable output, the learned selection functions behaved in accordance with our expectations for the posterior distribution over the latents.

The significant benefit we show empirically is that by learning the selection func-

tion in a general and flexible nonparametric way, we can avoid using expensive hand-engineered selection functions. Cost reduction is both in terms of required expertise in the problem domain, and computation time in identifying the relevant latent variables. Inference using our approach required 22.1 seconds on a single CPU core, versus 1830.9 seconds with the original hand-crafted function for the complex hierarchical model of (Dai et al., 2013).

A major area where further performance gains might be expected is in improving computational performance, since we expect the greatest advantages of GP-select to occur for complex models at large scale. For instance, kernel ridge regression may be parallelized (Zhang et al., 2014), or the problem may be solved in the primal via random Fourier features (Le et al., 2013). Furthermore, there are many recent developments regarding the scaling up of GP inference to large-scale problems, e.g., sparse GP approximation (Lawrence et al., 2002), stochastic variational inference (Hensman et al., 2013, 2012), using parallelization techniques and GPU acceleration (Dai et al., 2014), or in combination with stochastic gradient descent (Bottou and Bousquet, 2008). For instance, for very large datasets where the main model is typically trained with mini-batch learning, stochastic variational inference can be used for GP fitting as in (Hensman et al., 2013) and the kernel parameters can be efficiently updated each (or only every $T*$ few) iteration with respect to a mini-batch.

Acknowledgments

We acknowledge funding by the German Research Foundation (DFG) under grant LU 1196/4-2 (JS), by the Cluster of Excellence EXC 1077/1 "Hearing4all" (JL), and by the

RADIANT and WYSIWYD (EU FP7-ICT) projects (ZD).

References

Bornschein, J., Henniges, M., and Lücke, J. (2013). Are V1 simple cells optimized for visual occlusions? A comparative study. *PLoS Computational Biology*, 9(6):e1003062.

Bottou, L. and Bousquet, O. (2008). The tradeoffs of large scale learning. In Platt, J. C., Koller, D., Singer, Y., and Roweis, S. T., editors, *Advances in Neural Information Processing Systems 20 (NIPS), Vancouver, British Columbia, Canada*, pages 161–168. MIT Press.

Currin, C., Mitchell, T., Morris, M., and Ylvisaker, D. (1991). Bayesian prediction of deterministic functions, with applications to the design and analysis of computer experiments. *J. American Statistical Association*, 86(416):953–963.

Dai, Z., Damianou, A., Hensman, J., and Lawrence, N. (2014). Gaussian process models with parallelization and gpu acceleration. *NIPS 2014, Workshop on Modern nonparametrics: automating the learning pipeline*. In *NIPS Workshop on Modern nonparametrics: automating the learning pipeline*.

Dai, Z., Exarchakis, G., and Lücke, J. (2013). What are the invariant occlusive components of image patches? a probabilistic generative approach. In Burges, C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger., K., editors, *Advances in Neural Information Processing Systems*, pages 243–251.

Dai, Z. and Lücke, J. (2012a). Autonomous cleaning of corrupted scanned documents – a generative modeling approach. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3338–3345.

Dai, Z. and Lücke, J. (2012b). Unsupervised learning of translation invariant occlusive components. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2400–2407.

Dai, Z. and Lücke, J. (2014). Autonomous document cleaning – a generative approach to reconstruct strongly corrupted scanned texts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(10):1950–1962.

Dayan, P., Hinton, G. E., Neal, R. M., and Zemel, R. S. (1995). The helmholtz machine. *Neural Computation*, 7:889–904.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society B*, 39:1–38.

Exarchakis, G., Henniges, M., Eggert, J., and Lücke, J. (2012). Ternary sparse coding. In *LVA/ICA*, Lecture Notes in Computer Science, pages 204–212. Springer.

Fine, S. and Scheinberg, K. (2001). Efficient SVM training using low-rank kernel representations. *Journal of Machine Learning Research*, 2:243–264.

Goodfellow, I. J., Courville, A., and Bengio, Y. (2013). Scaling up spike-and-slab models for unsupervised feature learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1902–1914.

Gutmann, M. U. and Corander, J. (2015). Bayesian Optimization for Likelihood-Free Inference of Simulator-Based Statistical Models. Technical report, University of Helsinki. *http://arxiv.org/abs/1501.03291*.

Henniges, M., Puertas, G., Bornschein, J., Eggert, J., and Lücke, J. (2010). Binary Sparse Coding. In *Proceedings LVA/ICA*, LNCS 6365, pages 450–57. Springer.

Hensman, J., Fusi, N., and Lawrence, N. (2013). Gaussian processes for big data. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence, UAI 2013, Bellevue, WA, USA, August 11-15, 2013*.

Hensman, J., Rattray, M., and Lawrence, N. D. (2012). Fast Variational Inference in the Conjugate Exponential Family. In Bartlett, P. L., Pereira, F. C. N., Burges, C. J. C., Bottou, L., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 25 (NIPS), Lake Tahoe, Nevada, United States*, pages 2897–2905.

Hinton, G. E., Dayan, P., Frey, B. J., and Neal, R. M. (1995). The 'wake-sleep' algorithm for unsupervised neural networks. *Science*, 268:1158 – 1161.

Kingma, D. P. and Welling, M. (2014). Efficient gradient-based inference through transformations between bayes nets and neural nets. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, pages 1782–1790.

Körner, E., Gewaltig, M. O., Körner, U., Richter, A., and Rodemann, T. (1999). A model of computation in neocortical architecture. *Neural Networks*, 12:989 – 1005.

Lawrence, N., Seeger, M., and Herbrich, R. (2002). Fast sparse gaussian process methods: The informative vector machine. In *Advances in Neural Information Processing Systems 15 (NIPS), Vancouver, British Columbia, Canada*, pages 609–616. MIT Press.

Le, Q., Sarlos, T., and Smola, A. J. (2013). Fastfood — computing hilbert space expansions in loglinear time. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, pages 244–252.

Lücke, J. and Eggert, J. (2010). Expectation truncation and the benefits of preselection in training generative models. *Journal of Machine Learning Research*, 11:2855–900.

Lücke, J. and Sahani, M. (2008). Maximal causes for non-linear component extraction. *Journal of Machine Learning Research*, 9:1227–67.

Meeds, E. and Welling, M. (2014). GPS-ABC: gaussian process surrogate approximate bayesian computation. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence (UAI), Quebec City, Quebec, Canada, July 23-27, 2014*, pages 593–602.

Mnih, A. and Gregor, K. (2014). Neural variational inference and learning in belief networks. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, pages 1791–1799.

Neal, R. and Hinton, G. (1998). A view of the EM algorithm that justifies incremental, sparse, and other variants. In Jordan, M. I., editor, *Learning in Graphical Models*. Kluwer.

Nene, S. A., Nayar, S. K., and Murase, H. (1996). Columbia object image library (coil-100). Technical report, CUCS-006-96.

Rasmussen, C. E. and Williams, C. K. I. (2005). *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press.

Rupp, M., Tkatchenko, A., Müller, K.-R., and von Lilienfeld, O. A. (2012). Fast and accurate modeling of molecular atomization energies with machine learning. *Phys. Rev. Lett.*, 108:058301.

Sacks, J., Welch, W. J., Mitchell, T. J., and Wynn, H. P. (1989). Design and analysis of computer experiments. *Statist. Sci.*, 4(4):433–435.

Schwaighofer, A., Tresp, V., and Yu, K. (2004). Learning gaussian process kernels via hierarchical bayes. In *Advances in Neural Information Processing Systems 17 (NIPS), Vancouver, British Columbia, Canada*, pages 1209–1216. MIT Press.

Sheikh, A.-S., Shelton, J., and Lücke, J. (2014). A truncated EM approach for spike-and-slab sparse coding. *Journal of Machine Learning Research*, 15:2653–2687.

Shelton, J., Bornschein, J., Sheikh, A.-S., Berkes, P., and Lücke, J. (2011). Select and sample - A model of efficient neural inference and learning. In *Advances in Neural Information Processing Systems 24 (NIPS), Granada, Spain.*, pages 2618–2626.

Shelton, J., Sterne, P., Bornschein, J., Sheikh, A.-S., and Lücke, J. (2012). Why MCA? nonlinear sparse coding with spike-and-slab prior for neurally plausible image encoding. In *Advances in Neural Information Processing Systems 25 (NIPS), Lake Tahoe, Nevada, United States*, pages 2285–2293.

Shelton, J. A., Sheikh, A.-S., Bornschein, J., Sterne, P., and Lücke, J. (2015). Nonlinear spike-and-slab sparse coding for interpretable image encoding. *PLoS ONE*, 10(5):1–25.

Snoek, J., Rippel, O., Swersky, K., Kiros, R., Satish, N., Narayanan Sundaram, M., Ali, M., Patwary, P., and Adams, R. (2015). Scalable bayesian optimization using deep neural networks. Technical report, Harvard University. *http://arxiv.org/abs/1502.05700*.

Titsias, M. and Lázaro-Gredilla, M. (2011). Spike and slab variational inference for multi-task and multiple kernel learning. In *Advances in Neural Information Processing Systems 24 (NIPS), Granada, Spain*, pages 2510–2518.

Wilkinson, R. D. (2014). Accelerating ABC methods using gaussian processes. Technical report, University of Sheffield. *http://arxiv.org/abs/1401.1436*.

Yuille, A. and Kersten, D. (2006). Vision as Bayesian inference: analysis by synthesis? *Trends in Cognitive Sciences*, 10(7):301–308.

Zhang, Y., Duchi, J. C., and Wainwright, M. J. (2014). Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates. Technical report, University of California, Berkeley. *http://arxiv.org/abs/1305.5029*.