Supplementary Material

A Variational EM Basis of the Learning Algorithm

The parameter update rules for the MCA generative model were derived based on approximation (6). Here we show that this approximation can be recovered as a form of a variational EM approach. The derivation will be technically detailed but the final result, which includes a general procedure to derive M-step equations, will be very compact (see Tab. A.1). It was used to derive the parameter update rules Eqns. 9 to 14 for the MCA model. To study the result, the reader may directly be referred to Sec. A.4. More details about the approximation method itself will be published in the paper "Expectation Truncation and the Benefits of Preselection in Training Generative Models", Lücke and Eggert, *Journal of Machine Learning Research*, 2010, in press.

A.1 Variational Approximation for Data Classes

Let us consider a generative model with a set of hidden variables denoted by $\vec{s}$, a set of observed variables denoted by $\vec{y}$, and a set of parameters denoted by Θ. Let us denote the prior distribution of the model by the (not further specified) function $p(\vec{s}\,|\,\Theta)$, and the noise distribution by the (not further specified) function $p(\vec{y}\,|\,\vec{s},\,\Theta)$. To distinguish this generative model from models introduced later, it will from now on be referred to as the *original* generative model.

Let $\mathcal{K}$ be a subset of the space of all possible values of $\vec{s}$. Given such a set, let us define two new generative models by introducing two new prior distributions that are based on the original prior:

$$p(\vec{s}\,|\,c=1,\Theta) \;=\; \begin{cases} \frac{1}{\tilde{\kappa}}\,p(\vec{s}\,|\,\Theta) & \text{if } \vec{s}\in\mathcal{K} \\ 0 & \text{if } \vec{s}\notin\mathcal{K} \end{cases} \quad\text{and}\quad p(\vec{s}\,|\,c=0,\Theta) \;=\; \begin{cases} 0 & \text{if } \vec{s}\in\mathcal{K} \\ \frac{1}{1-\tilde{\kappa}}\,p(\vec{s}\,|\,\Theta) & \text{if } \vec{s}\notin\mathcal{K} \end{cases} \tag{16}$$

where $\tilde{\kappa} \;=\; \sum_{\vec{s}\in\mathcal{K}} p(\vec{s}\,|\,\Theta)$. We take the noise distribution of the new models to be identical to the original noise distribution. We will refer to the new generative models as *truncated* models because their prior distributions are truncated to be zero outside of specific subsets. Note that the generation of data according to the truncated model with $c = 1$ corresponds to generating data according to the original model while only accepting data points generated by $\vec{s}\in\mathcal{K}$. Analogously, generating data according to the truncated model with $c = 0$ is equivalent to generating data according to the original model while accepting only data points generated by $\vec{s}\notin\mathcal{K}$.

Let us now mix these two truncated generative models by introducing $c\in\{0,1\}$ as additional hidden variable and by drawing $c = 1$ with probability κ. The prior distribution of this mixed model is thus given by:

$$p(c\,|\,\kappa) \;=\; \kappa^{c}\,(1-\kappa)^{1-c}\,, \tag{17}$$

$$p(\vec{s}\,|\,c,\Theta) \;=\; \Big(\frac{c}{\tilde{\kappa}}\;\delta(\vec{s}\in\mathcal{K}) \;+\; \frac{1-c}{1-\tilde{\kappa}}\;\delta(\vec{s}\notin\mathcal{K})\Big)\,p(\vec{s}\,|\,\Theta)\,. \tag{18}$$

where we have introduced $\delta(\vec{s}\in\mathcal{K}) = 1$ if $\vec{s}\in\mathcal{K}$ and zero otherwise, and $\delta(\vec{s}\notin\mathcal{K}) = 1$ if $\vec{s}\notin\mathcal{K}$ and zero otherwise. We will refer to this model as the *mixed* generative model. Note that the mixed model is identical to the original generative model if we choose $\kappa = \tilde{\kappa} \;=\; \sum_{\vec{s}\in\mathcal{K}} p(\vec{s}\,|\,\Theta)$ as mixing proportion. The mixed model thus contains the original model as a special case.

Now, consider a set of N data points $\{\vec{y}^{\,(n)}\}_{n=1,\ldots,N}$ generated according to the original generative model. Let us maximize the likelihood of the data under the mixed model (17) and (18). If we use EM for optimization, we obtain the free-energy

$$\begin{aligned}\tilde{\mathcal{F}}(q,\Theta,\kappa) \;\;=\;\; & \sum_{n}\sum_{c} q^{(n)}(c\,;\Theta')\,\log\big(p(\vec{y}^{\,(n)}\,|\,c,\Theta)\big) \\ & +\log(\kappa)\sum_{n} q^{(n)}(c=1;\Theta') + \log(1-\kappa)\sum_{n} q^{(n)}(c=0;\Theta') + H(q)\,,\end{aligned} \tag{19}$$

where $H(q)$ is the entropy w.r.t. $q^{(n)}(c;\Theta')$ (summed over all n and c). The free-energy (19) can be optimized iteratively by maximizing q in the E-step and (Θ,κ) in the M-step. For the E-step, choosing the exact posterior, $q^{(n)}(c\,;\,\Theta') = p(c\,|\,\vec{y}^{\,(n)},\Theta')$, represents the optimal choice. Unfortunately,

it is computationally intractable in general because

$$p(c=1\,|\,\vec{y}^{\,(n)},\Theta') = \frac{\sum_{\vec{s}\in\mathcal{K}} p(\vec{y}^{\,(n)},\vec{s}\,|\,\Theta)}{\sum_{\vec{s}} p(\vec{y}^{\,(n)},\vec{s}\,|\,\Theta)}\,, \tag{20}$$

requires a summation over the entire state space (similarly for $c=0$). We thus choose a variational approximation to the true posterior by setting $q^{(n)}(c\,|\,\Theta')$ to zero or one. This approximation reduces the free-energy (19) to:

$$\begin{aligned}\tilde{\mathcal{F}}(q,\Theta) \;\;=\;\; & \overbrace{\sum_{n\in\mathcal{M}} \log\left(p(\vec{y}^{\,(n)}\,|\,c=1,\Theta)\right)}^{\approx L_1(\Theta)} + \overbrace{\sum_{n\notin\mathcal{M}} \log\left(p(\vec{y}^{\,(n)}\,|\,c=0,\Theta)\right)}^{\approx L_0(\Theta)} \\ & + \log(\kappa)|\mathcal{M}| + \log(1-\kappa)(N-|\mathcal{M}|) + H(q)\,. \end{aligned} \tag{21}$$

where $\mathcal{M} = \{n\,|\,q^{(n)}(c=1\,;\,\Theta')=1\}$. Note that κ can be optimized independently of Θ because the first two summands in (21) only depend on Θ. As we also know its final optimal value ($\kappa=\tilde{\kappa}$), we will treat the mixing proportion as implicitly known. κ can thus be omitted as a parameter of the free-energy (21) and will not play a role for our further considerations.

For the data set $\mathcal{M}$ note that the best choice for $q^{(n)}(c\,;\,\Theta')$ under the constraint $q^{(n)}(c\,;\,\Theta') \in \{0,1\}$ is given by the assignment $q^{(n)}(c=1\,;\,\Theta')=1$ if $\vec{y}^{\,(n)}$ was generated by class $c=1$ (and zero otherwise). This would amount to setting $\mathcal{M}$ to

$$\mathcal{M}^{\mathrm{opt}} = \{n\,|\,\vec{y}^{\,(n)} \text{ generated by class } c=1\}\,. \tag{22}$$

Note that in general this best choice can not be computed exactly. Later in Sec. A.3 we will see, however, that tractable approximations to $\mathcal{M}^{\mathrm{opt}}$ can be derived.

A.2 Necessary Conditions for Global Likelihood Maxima

In the previous section we have seen that $\tilde{\mathcal{F}}(q,\Theta)$ in (21) is a lower bound of the data likelihood $L(\Theta)$. If the used variational distributions $q^{(n)}(c\,;\,\Theta)$ are good approximations to the exact posteriors $p(c\,|\,\vec{y}^{\,(n)},\Theta)$ in (20), then $L(\Theta) \approx \tilde{\mathcal{F}}(q,\Theta)$ after each E-step. Because of the variational approximation in Sec. A.1 the equality $\tilde{\mathcal{F}}(q,\Theta) = L(\Theta)$ holds if $\mathcal{M}$ is equal to $\mathcal{M}^{\mathrm{opt}}$ (22) and if the true posterior values in (20) are equal to zero or one for all n. Although the latter condition is fulfilled only in boundary cases, we will, in the following, assume the equality to hold (while keeping in mind that it is almost always an approximation). If the equality holds, $\tilde{\mathcal{F}}(q,\Theta)$ in (21) is in its global maximum equal to the likelihood $L(\Theta)$.

Let us assume that there exist parameters Θ^* such that the original generative model reproduces the underlying distribution of the data points, $p(\vec{y}) = p(\vec{y}\,|\,\Theta^*)$. From Sec. A.1 we then know that the mixed model with prior (17) and (18) and $\kappa=\tilde{\kappa}$ also reproduces the original distribution for these parameters. Using the mixed model, the data points $\{\vec{y}^{\,(n)}\}_{n=1,\ldots,N}$ can thus be taken to have been generated by the truncated generative models. That is, the data set can be subdivided into the two disjoint sets $\{\vec{y}^{\,(n)}\}_{n\in\mathcal{M}^{\mathrm{opt}}}$ and $\{\vec{y}^{\,(n)}\}_{n\notin\mathcal{M}^{\mathrm{opt}}}$ (compare Fig. A.1). If $p(\vec{y}\,|\,\Theta^*)$ is the underlying distribution of the whole data set, then $p(\vec{y}\,|\,c=1,\Theta^*)$ and $p(\vec{y}\,|\,c=0,\Theta^*)$ are the underlying distributions of the two disjoint parts.

We can approximately recover the distribution $p(\vec{y}\,|\,\Theta^*)$ by (globally) maximizing the data likelihood under the mixed generative model on $\{\vec{y}^{\,(n)}\}_{n=1,\ldots,N}$. Furthermore, we can recover the distributions $p(\vec{y}\,|\,c=1,\Theta^*)$ and $p(\vec{y}\,|\,c=0,\Theta^*)$ by (globally) maximizing the data likelihoods of the truncated generative models on $\{\vec{y}^{\,(n)}\}_{n\in\mathcal{M}^{\mathrm{opt}}}$ and $\{\vec{y}^{\,(n)}\}_{n\notin\mathcal{M}^{\mathrm{opt}}}$, respectively. Let us denote the parameters recovered by maximizing $L(\Theta)$ by $\Theta^\dagger$, and the parameters recovered by maximizing $L_1(\Theta)$ and $L_0(\Theta)$ by $\Theta^{\dagger 1}$ and $\Theta^{\dagger 0}$, respectively. In general, $\Theta^\dagger$, $\Theta^{\dagger 1}$, and $\Theta^{\dagger 0}$ are different. If the variational approximation $\mathcal{M}=\mathcal{M}^{\mathrm{opt}}$ is exact, we know, however, that in the limit of infinitely many data points (and by still assuming $p(\vec{y}) = p(\vec{y}\,|\,\Theta^*)$) applies:

$$p(\vec{y}\,|\,\Theta^*) = p(\vec{y}\,|\,\Theta^\dagger),\; p(\vec{y}\,|\,c{=}1,\Theta^*) = p(\vec{y}\,|\,c{=}1,\Theta^{\dagger 1}),\; p(\vec{y}\,|\,c{=}0,\Theta^*) = p(\vec{y}\,|\,c{=}0,\Theta^{\dagger 0}). \tag{23}$$

The equalities hold because for $N\to\infty$ and $p(\vec{y}) = p(\vec{y}\,|\,\Theta^*)$ it follows from $L(\Theta^*) = L(\Theta^\dagger)$ that $D_{KL}(p(\vec{y}\,|\,\Theta^*), p(\vec{y}\,|\,\Theta^\dagger)) = 0$. As the Kullback-Leibler divergence between two distributions

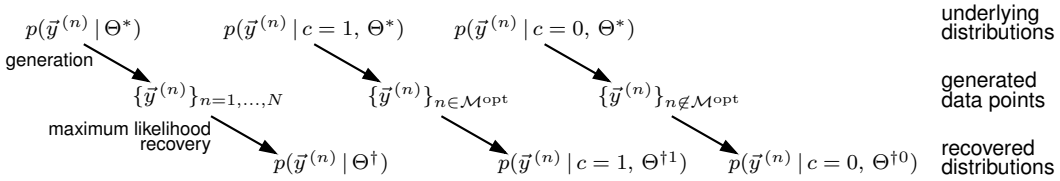

Figure A.1: Recovery of the generating distributions through the original and the truncated generative models. The original distributions can be recovered from the data sets if the corresponding likelihoods are maximized.

q and p is zero if and only if the distributions are identical, (23) has to hold. Note, however, that the recovered parameters can still be different from Θ^*. For instance, if there exist transformations $\mathcal{T}$ of Θ^* that do not change the distribution, then any Θ obtained from Θ^* through such a transformation, $\Theta = \mathcal{T}(\Theta^*)$, is a likelihood maximum as well. Such multiple global maxima are the norm rather than the exception. Global maxima of models such as Sparse Coding [1] or Independent Component Analysis [e.g. 2] remain global maxima under the exchange of any two basis functions or the negation of any of them.

Let all transformations $\mathcal{T}$ that map the global maxima of $L(\Theta)$ onto itself define a set that we will refer to as the *transformation set*. We say that the set of global maxima is *invariant* under the transformation set. For any two global maxima Θ^* and $\Theta^\dagger$ of $L(\Theta)$, there now exists a member $\mathcal{T}$ of the transformation set such that $\Theta^* = \mathcal{T}(\Theta^\dagger)$. Let us now demand that all global maxima of $L_1(\Theta)$ and $L_0(\Theta)$ are also invariant under the transformation set. Although this will usually be the case, e.g., for the exchange of any two basis functions, it is important to state this requirement explicitly as it is not fulfilled in general. If this property is fulfilled, however, we can infer (under the made assumptions):

$$
\begin{aligned}
&\Theta^\dagger \text{ is maximum likelihood solution on } L(\Theta)\\
\Rightarrow\ &\text{There exists } \mathcal{T} \text{ such that } \Theta^\dagger = \mathcal{T}(\Theta^*) \text{ with } \Theta^* \text{ being the generating parameters.}\\
\Rightarrow\ &p(\vec{y}\,|\,c=1,\Theta^*) \text{ is the actual generating distribution of } \{\vec{y}^{\,(n)}\}_{n\in\mathcal{M}^{\mathrm{opt}}}\\
\Rightarrow\ &\Theta^* \text{ is maximum likelihood solution of } L_1(\Theta)\\
\Rightarrow\ &p(\vec{y}\,|\,c=1,\Theta^*) = p(\vec{y}\,|\,c=1,\mathcal{T}(\Theta^*)) = p(\vec{y}\,|\,c=1,\Theta^\dagger)\\
\Rightarrow\ &\Theta^\dagger \text{ is maximum likelihood solution of } L_1(\Theta) \qquad (24)
\end{aligned}
$$

Analogously, $\Theta^\dagger$ is also a maximum likelihood solution of $L_0(\Theta)$ if it is a maximum likelihood solution of $L(\Theta)$. For the free-energy (21) this means that at a global maximum of $L(\Theta)$ both $L_1(\Theta)$ and $L_0(\Theta)$ also have a global maximum (under the made assumptions). A global maximum, e.g., in $L_1(\Theta)$ is thus a *necessary* condition for a global maximum in $L(\Theta)$. We have *not* shown that a maximum in $L_1(\Theta)$ is a sufficient condition for a maximum in $L(\Theta)$. Theoretically, $L_1(\Theta)$ might, for instance, not depend on all parameters, or it might have additional global maxima. Finally, note again that the necessary condition only holds under the introduced assumptions. While, e.g., the assumption on invariance under transformations $\mathcal{T}$ can exactly be fulfilled (depending on the generative model), the assumptions that the true data distribution can exactly be matched or that the variational approximation in Sec. A.1 is exact are in practice almost never fulfilled. The same applies for the assumption of infinitely many data points. All these assumptions can, however, be fulfilled approximately. By (globally) maximizing $L_1(\Theta)$ we can thus expect to recover parameters that maximize $L(\Theta)$ approximately.

Note that, intuitively, it makes sense that the maximization of the likelihood $L_1(\Theta)$ also approximately maximizes $L(\Theta)$. To see this consider the example of Fig. A.2 of a generative model with bar-like basis functions (or generative fields). If sufficiently many data points are available, likelihood maximization under the truncated generative model on $\{\vec{y}^{\,(n)}\}_{n\in\mathcal{M}}$ will recover basis functions that are also approximately the basis functions of the original model. Thus, also the likelihood of the original model based on the entire data set $\{\vec{y}^{\,(n)}\}_{n=1,\ldots,N}$ will be maximized approximately. Note that this example also demonstrates that a given input vector only has to be evaluated by a more limited number of possible states. The truncated model in Fig. A.2 only requires the evaluation of 56 instead of $2^{10} = 1024$ potential interpretations.

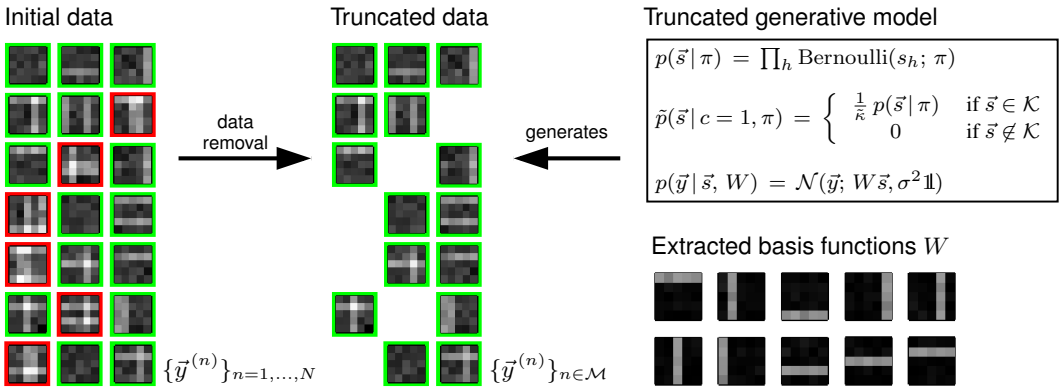

Figure A.2: Illustration of Expectation Truncation (without preselection) for a concrete generative model. The model generates data by linearly superimposing basis functions in the form of horizontal and vertical bars. Data generated by the original model contains up to ten bars chosen with a Bernoulli prior (example data points are shown on the left-hand-side). Data generated by the truncated generative model contains data with up to two bars (we set $\mathcal{K} = \{\vec{s}\,|\,\sum_j s_j \leq \gamma\}$ with $\gamma = 2$). If we train the truncated generative model on data from which data points with $\sum_j s_j > 2$ were removed (see truncated data), we can expect to approximately recover the true generating basis functions of the original model.

A.3 Preselection as Variational Approximation

To maximize $L(\Theta)$ we use the necessary condition (24) and maximize $L_1(\Theta)$ instead. However, we do not maximize $L_1(\Theta)$ directly but optimize the lower bound $\mathcal{F}_1(q,\Theta)$ given by:

$$\mathcal{F}_1(q,\Theta) = \overbrace{\sum_{n\in\mathcal{M}}\sum_{\vec{s}\in\mathcal{K}} q^{(n)}(\vec{s};\Theta')\,\log\left(p(\vec{y}^{\,(n)}\,|\,\vec{s},\Theta)\,\frac{1}{\tilde{\kappa}}\,p(\vec{s}\,|\,\Theta)\right)}^{Q_1(q,\Theta)} + H(q)\,, \tag{25}$$

with $\sum_{\vec{s}\in\mathcal{K}} q^{(n)}(\vec{s};\Theta') = 1$. $\mathcal{F}_1(q,\Theta)$ is derived by a variational approximation, this time w.r.t. the hidden variables $\vec{s}$. The free-energy equals the likelihood $L_1(\Theta)$ after each E-step if the distributions $q^{(n)}(\vec{s};\Theta')$ are given by:

$$q^{(n)}(\vec{s};\Theta') \;=\; p(\vec{s}\,|\,\vec{y}^{\,(n)},c=1,\Theta') \;=\; \frac{p(\vec{s}\,|\,\vec{y}^{\,(n)},\Theta')}{\sum_{\vec{s}'\in\mathcal{K}} p(\vec{s}'\,|\,\vec{y}^{\,(n)},\Theta')}\delta(\vec{s}\in\mathcal{K}). \tag{26}$$

M-step rules can be derived by setting the derivatives of $\mathcal{F}_1(q,\Theta)$ w.r.t. all parameters to zero. As the entropy term in (25) is independent of Θ if q is held fixed, we obtain

$$\frac{d}{d\Theta}\mathcal{F}_1(q,\Theta) = \frac{d}{d\Theta}Q_1(q,\Theta) = 0 \tag{27}$$

as necessary condition. The derivative $\frac{d}{d\Theta}$ hereby stands for derivatives w.r.t. all the individual parameters.

Based on condition (27) we can now introduce candidate preselection as a variational approximation. As briefly described in Sec. 2, preselection amounts to selecting, for a given $\vec{y}^{\,(n)}$, a subset $\mathcal{K}_n$ of the state space. For MCA we use $\mathcal{K} = \{\vec{s}\,|\,\sum_j s_j \leq \gamma\}$ and $\mathcal{K}_n$ given by Eqn. 7 (note that $\mathcal{K}_n \subseteq \mathcal{K}$). For MCA the set $\mathcal{K}_n$ is constructed using the selection function (8). More generally, we require from the sets $\mathcal{K}_n$ that for all data points generated by $\vec{s}\in\mathcal{K}$, they finally contain most of the posterior mass in $\mathcal{K}$. If this applies, we obtain an approximation to the posterior $q^{(n)}$ in (26) given by:

$$\tilde{q}^{(n)}(\vec{s};\Theta') \;=\; \frac{p(\vec{s}\,|\,\vec{y}^{\,(n)},\Theta')}{\sum_{\vec{s}\in\mathcal{K}_n} p(\vec{s}\,|\,\vec{y}^{\,(n)},\Theta')}\;\delta(\vec{s}\in\mathcal{K}_n) \;=\; \frac{p(\vec{s},\,\vec{y}^{\,(n)}\,|\,\Theta')}{\sum_{\vec{s}\in\mathcal{K}_n} p(\vec{s},\,\vec{y}^{\,(n)}\,|\,\Theta')}\;\delta(\vec{s}\in\mathcal{K}_n) \tag{28}$$

Note that $\tilde{q}^{(n)}(\vec{s};\Theta')$ sums to one in $\mathcal{K}$ as $\mathcal{K}_n \subseteq \mathcal{K}$ and thus fulfils the condition on $q^{(n)}$ required for (25). If preselection finds, at least finally, appropriate sets $\mathcal{K}_n$, we obtain with (28):

$\frac{d}{d\Theta}Q_1(\tilde{q},\Theta) \approx \frac{d}{d\Theta}Q_1(q_1,\Theta) \approx 0$. Parameter update rules derived from condition $\frac{d}{d\Theta}Q_1(\tilde{q},\Theta) = 0$ can therefore be expected to (at least approximately) optimize the free-energy (21) and thus $L_1(\Theta)$. The update rules derived will contain expectation values (the sufficient statistics) of the form $\langle g(\vec{s})\rangle_{\tilde{q}^{(n)}}$. If we use (28) for these expectations we obtain:

$$\langle g(\vec{s})\rangle_{\tilde{q}^{(n)}} \;=\; \sum_{\vec{s}} \tilde{q}^{(n)}(\vec{s};\Theta')\, g(\vec{s}) \;=\; \frac{\sum\limits_{\vec{s}\in\mathcal{K}_n} p(\vec{s},\vec{y}^{\,(n)}\,|\,\Theta')\; g(\vec{s})}{\sum\limits_{\tilde{\vec{s}}\in\mathcal{K}_n} p(\,\tilde{\vec{s}}\,,\vec{y}^{\,(n)}\,|\,\Theta')}\,, \tag{29}$$

i.e., precisely expression (6). The approximation is thus equivalent to a variational approximation. Importantly, this approximation is tractable if $\mathcal{K}_n$ is small. The computational gain of preselection compared to an approximation without preselection is reflected by the reduced size of $\mathcal{K}_n$ compared to $\mathcal{K}$. This large reduction of hidden states that have to be evaluated is the crucial property that can be exploited for large-scale applications.

The remaining intractability of the approximation scheme is the intractability in computing the best data point set $\mathcal{M}$ for the free-energy (21). As stated earlier, the best choice for $\mathcal{M}$ would be to choose $\mathcal{M}$ equal to $\mathcal{M}^{\rm opt}$ in Eqn. 22. Without ground-truth information, $\mathcal{M}^{\rm opt}$ can not be computed exactly. We can, however, try to approximate $\mathcal{M}^{\rm opt}$. To do so, first note that we can compute an expectation value for the size of $\mathcal{M}^{\rm opt}$. It is given by $N(\mathcal{K}) = N\,\sum_{\vec{s}\in\mathcal{K}} p(\vec{s}\,|\,\Theta)$. We can now find an approximation to $\mathcal{M}^{\rm opt}$ by computing the values $q^{(n)}(c=1\,;\,\Theta')$ for all data points, sort them, and take the data points with the $N(\mathcal{K})$ highest values. This would represent a good approximation to $\mathcal{M}^{\rm opt}$ but it seems that we have gained very little, since we still have to compute the intractable posteriors $q^{(n)}(c=1\,;\,\Theta')$ for all n. Note, however, that with this procedure, the absolute values of $q^{(n)}(c=1\,;\,\Theta')$ are not used for the approximation anymore. All that is required is a pairwise comparison of the data points based on the their values $q^{(n)}(c=1\,;\,\Theta')$.

To derive a tractable approximation of the pairwise comparison, consider two data points that are neighbors after a sorting according to $q^{(n)}(c=1\,;\,\Theta')$. For arbitrarily many data points and for non-zero noise, the differences between the two data points become arbitrarily small. In particular, it applies for neighboring data points that the difference between the denominators of $q^{(n)}(c=1\,;\,\Theta')$ become arbitrarily small: $\sum_{\vec{s}} p(\vec{y}^{\,(n)},\vec{s}\,|\,\Theta) \approx \sum_{\vec{s}} p(\vec{y}^{\,(n')},\vec{s}\,|\,\Theta)$. The same applies for differences between the numerators. However, as the numerators contain just small sums over $\vec{s}$, their values for neighboring data points can be expected to vary more strongly than those of the denominators. We can thus replace the comparison between $q^{(n)}(c=1\,;\,\Theta')$ by a comparison of their numerators $\sum_{\vec{s}\in\mathcal{K}} p(\vec{y}^{\,(n)},\vec{s}\,|\,\Theta)$. This is an approximation to the pairwise comparison required for exact sorting. In the limit of infinitely many data points this procedure can be expected to result in sets $\mathcal{M}$ that represent good approximations to $\mathcal{M}^{\rm opt}$.

If we now take preselection into account, the comparison for sorting can be reduced further. For this note that the posterior in (20) is approximated by $q^{(n)}(c=1\,;\,\Theta') \approx \frac{\sum_{\vec{s}\in\mathcal{K}_n} p(\vec{y}^{\,(n)},\vec{s}\,|\,\Theta)}{\sum_{\vec{s}} p(\vec{y}^{\,(n)},\vec{s}\,|\,\Theta)}$. Following the same arguments as above, an approximation of the sorting by comparing the values $q^{(n)}(c=1\,;\,\Theta')$ is given by sorting based on the values $\sum_{\vec{s}\in\mathcal{K}_n} p(\vec{y}^{\,(n)},\vec{s}\,|\,\Theta)$ for all n. This shows that the selection of $N^{\rm cut}$ data points (compare Eqn. 12) corresponds to defining the set $\mathcal{M}$ as an approximation to $\mathcal{M}^{\rm opt}$. Choosing $\mathcal{M}$ is, on the other hand, equivalent to choosing a $q^{(n)}(c\,;\,\Theta')$ with binary function values (as an approximation to Eqn. 20). The selection of $N^{\rm cut}$ data points is thus equivalent to a variational E-step. Combined with preselection, this E-step is tractable if $\mathcal{K}_n$ is sufficiently small.

A.4 Summary and Application to the MCA Model

Independent of a specific form of the generative model, we have seen that approximation (6) can be derived as a variational EM approach. This approach consists of two variational approximation steps: First, an approximation that assigns the data points to two classes (Sec. A.1). Second, a variational step that approximates the true posterior (26) by an approximate posterior (28) defined through preselection (Sec. A.3). Although the derivation of the approximation as a variational approach requires in parts rather technical steps, the final result is intuitive (Fig. A.2) and can be stated

Tab. A.1: **Expectation Truncation**

Preselection:	select a state space volume $\mathcal{K}_n$ for each data point $\vec{y}^{\,(n)}$
Data classification:	find a data set $\mathcal{M}$ that approximates $\mathcal{M}^{\mathrm{opt}}$ in Eqn. 22
E-step:	compute $\tilde{q}^{(n)}(\vec{s}\,;\Theta') \;=\; \dfrac{p(\vec{s},\,\vec{y}^{\,(n)}\,|\,\Theta')}{\sum_{\vec{s}\in\mathcal{K}_n} p(\vec{s},\,\vec{y}^{\,(n)}\,|\,\Theta')}$ for all $\vec{y}^{\,(n)}$ and $\vec{s}\in\mathcal{K}_n$
M-step:	find parameters Θ such that $\dfrac{d}{d\Theta}\sum_{n\in\mathcal{M}}\sum_{\vec{s}\in\mathcal{K}_n}\tilde{q}^{(n)}(\vec{s}\,;\Theta')\,\log\left(p(\vec{y}^{\,(n)}\,|\,\vec{s},\Theta)\,\dfrac{p(\vec{s}\,|\,\Theta)}{\sum_{\vec{s}'\in\mathcal{K}}p(\vec{s}\,'\,|\,\Theta)}\right)=0$

very compactly. Tab. A.1 summarizes all required steps of the approximation scheme. Note that also with preselection, the approximation still requires a summation over $\mathcal{K}$, namely $\sum_{\vec{s}\in\mathcal{K}} p(\vec{s}\,|\,\Theta)$. This sum has to be computed to determine N^{cut}, $N^{\mathrm{cut}} = N\sum_{\vec{s}\in\mathcal{K}} p(\vec{s}\,|\,\Theta)$, and it appears in the M-step equation (Tab. A.1). Because of symmetries in the usual priors for generative models (e.g., for the prior used in MCA), this sum can, however, be computed without summing over all $\vec{s}$ explicitly (compare Eqn. 12).

The update equations for MCA are derived based on Tab. A.1 using $\mathcal{K} = \{\vec{s}\,|\,\sum_j s_j \leq \gamma\}$ and $\mathcal{K}_n$ as given in (7). The derivation for W follows the same lines as the derivation in [19]. Note, however, that the sum over all data points is replaced by a restricted sum over $\mathcal{M}$. Furthermore, our derivation includes a general γ and the same truncation of numerator and denominator in (6). Also the derivation of the update rule for σ is relatively straight-forward because the prior distribution is independent of σ (as well as of W). The formula for the M-step in Tab. A.1 thus reduces to the usual form (except for the summation over $\mathcal{M}$). To derive the update rule for π, the prior and its marginal over $\mathcal{K}$ have to be taken into account. Using

$$\frac{d}{d\pi}\log(A(\pi)) \;=\; \frac{B(\pi)}{\pi\,(1-\pi)\,A(\pi)} - \frac{H}{1-\pi}\,, \tag{30}$$

with $A(\pi) = \sum_{\vec{s}\in\mathcal{K}} p(\vec{s}\,|\,\Theta)$ and $B(\pi)$ as in Eqn. 14. We obtain by taking the derivative w.r.t. π in the M-step of Tab. A.1:

$$\pi \;=\; \frac{A(\pi)\,\pi}{B(\pi)}\,\frac{1}{|\mathcal{M}|}\sum_{n\in\mathcal{M}}\langle|\vec{s}|\rangle_{q_n} \;\text{ with }\; |\vec{s}| = \sum_{h=1}^{H} s_h\,. \tag{31}$$

Applying this equation in the fix-point sense (compare Eqn. 13), results in a convergence to values π that represent solutions of Eqn. 31. Note that (31) reduces to the exact update rule if $\mathcal{K}$ and $\mathcal{K}_n$ are chosen to contain all data points.

B Upper Bound Property of the Selection Function

Note first that the natural starting point to estimate candidate hidden variables for a data point $\vec{y}^{\,(n)}$ is the evaluation of probabilities $p(s_h = 1 \,|\, \vec{y}^{\,(n)}, \Theta) = \frac{p(s_h=1, \vec{y}^{\,(n)} \,|\, \Theta)}{p(\vec{y}^{\,(n)} \,|\, \Theta)}$. Here we will show that $\frac{\mathcal{S}_h(\vec{y}^{\,(n)})}{p(\vec{y}^{\,(n)} \,|\, \Theta)}$ (with $\mathcal{S}_h(\vec{y}^{\,(n)})$ given in Eqn. 8) is an upper bound of $p(s_h = 1 \,|\, \vec{y}^{\,(n)}, \Theta)$. This implies that $p(s_h = 1 \,|\, \vec{y}^{\,(n)}, \Theta)$ is small if the upper bound is small. As $p(\vec{y}^{\,(n)} \,|\, \Theta)$ is independent of h, selecting candidates based on this upper bound is equivalent to selecting candidates based on the function $\mathcal{S}_h(\vec{y}^{\,(n)})$.

We show that $p(s_h = 1, \vec{y}^{\,(n)} \,|\, \Theta)$ is bounded by $\mathcal{S}_h(\vec{y}^{\,(n)})$ from above. Let us first define $\overline{W}_d(\vec{s}, W) = \max_h\{s_h W_{dh}\}$. Now, consider the set $\delta_h := \{d \in \{1, \ldots, D\} \,|\, y_d^{(n)} < W_{dh}\}$ and note that if $y_d^{(n)} < W_{dh}$ and $s_h = 1$ then $p(y_d^{(n)} \,|\, W_{dh}) < p(y_d^{(n)} \,|\, \overline{W}_d(\vec{s}, W))$. This is because $\overline{W}_d(\vec{s}, W)$ can only be larger than W_{dh} and therefore the mono-modal Gaussian distribution $p(y_d^{(n)} \,|\, \overline{W}_d(\vec{s}, W))$ is further away from the maximum value. $p(y_d^{(n)} \,|\, w)$ with $w = y_d^{(n)}$ is on the other hand larger or equal to $p(y_d^{(n)} \,|\, w')$ for any other value w'. It follows:

$$
\begin{aligned}
p(s_h = 1, \vec{y}^{\,(n)} \,|\, \Theta) \;&=\; \sum_{\substack{\vec{s} \\ s_h \,=\, 1}} p(\vec{y}^{\,(n)} \,|\, \vec{s}, \Theta)\, p(\vec{s} \,|\, \pi) \\
&=\; \sum_{\substack{\vec{s} \\ s_h \,=\, 1}} \big(\prod_{d=1}^{D} p(y_d^{(n)} \,|\, \overline{W}_d(\vec{s}, W), \sigma) \big)\, p(\vec{s} \,|\, \pi) \\
&=\; \sum_{\substack{\vec{s} \\ s_h \,=\, 1}} \big(\prod_{d \in \delta_h} p(y_d^{(n)} \,|\, \overline{W}_d(\vec{s}, W), \sigma) \big) \big(\prod_{d \notin \delta_h} p(y_d^{(n)} \,|\, \overline{W}_d(\vec{s}, W), \sigma) \big)\, p(\vec{s} \,|\, \pi) \\
&\leq\; \big(\prod_{d \in \delta_h} p(y_d^{(n)} \,|\, W_{dh}, \sigma) \big) \big(\prod_{d \notin \delta_h} p(y_d^{(n)} \,|\, y_d^{(n)}, \sigma) \big) \sum_{\substack{\vec{s} \\ s_h \,=\, 1}} p(\vec{s} \,|\, \pi) \\
&=\; \big(\prod_{d \in \delta_h} p(y_d^{(n)} \,|\, W_{dh}^{\mathrm{eff}}, \sigma) \big) \big(\prod_{d \notin \delta_h} p(y_d^{(n)} \,|\, W_{dh}^{\mathrm{eff}}, \sigma) \big) \pi \\
&=\; \pi\, p(\vec{y}^{\,(n)} \,|\, \vec{W}_h^{\mathrm{eff}}, \sigma) \;=:\; \mathcal{S}_h(\vec{y}^{\,(n)})\,,
\end{aligned}
$$

where $W_{dh}^{\mathrm{eff}} = \max\{y_d^{(n)}, W_{dh}\}$ as in Eqn. 8.

C Details of Numerical Experiments

We briefly describe some technical details and additional results about the application of the algorithm to natural image patches.

C.1 DoG Preprocessing

The algorithm is applied to natural image patches that were DoG preprocessed. Before comparing the obtained generative fields with receptive fields measured *in vivo*, the preprocessing has to be taken into account. In the case of DoG preprocessing, the measured receptive fields should (as a first order approximation) be compared to generative fields convoluted with the same DoG filter (see [27] for details). The type of preprocessing can influence the obtained generative fields. For the experiments in this work, the distributions of spatial frequencies of the convoluted generative fields follow the bandwidth property of the used DoG filter. The distribution of convoluted fields in n_x/n_y-space can be expected to be influenced just very weakly if the mean bandwidth frequency of the DoG filter is changed. This is because the learned σ_x and σ_y can be expected to change in the same way as the fitted spatial period length T (note that $n_x \sim \frac{\sigma_x}{T}$, similar n_y). Indirect effects might be possible, however.

C.2 Additional Results on Image Patches

Fig. C.1 shows all generative fields obtained from the run displayed in Fig. 4. Note the relatively high percentage of globular fields (compare [33]) which is not observed for standard linear approaches. In linear approaches with continuous hidden variables, globular fields can be obtained by linearly superimposing two orthogonal Gabor functions (compare Fig. 1). This can explain why globular fields are not or only rarely obtained by linear approaches such as sparse coding or ICA.

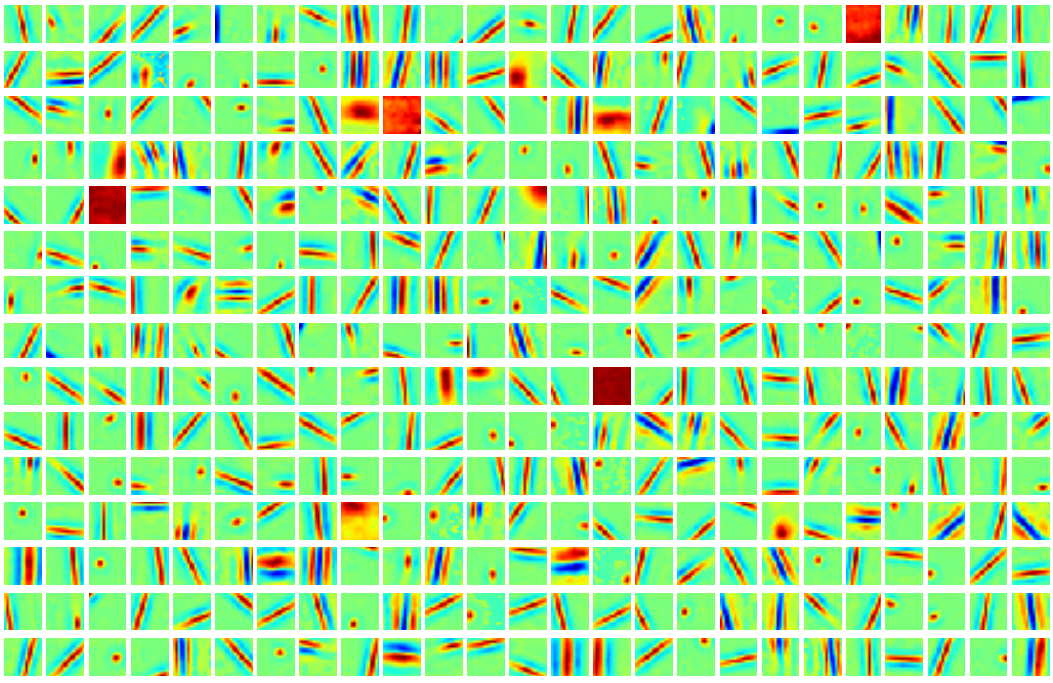

Figure C.1: Complete set of $H = 400$ basis functions obtained after learning (same run as in Fig. 4). For visualization purposes, the common DC component of all fields was subtracted before the positive and negative parts were combined (the same applies to all basis function visualizations in this work). The DC component accounts for less than 10% of the fields' total amplitude.

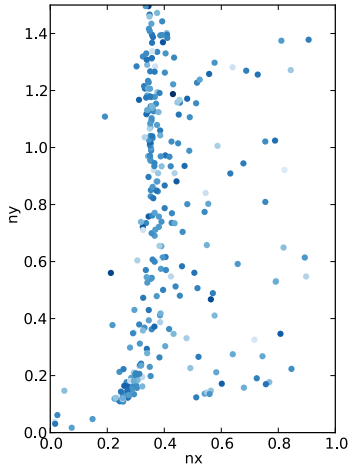

Figure C.2: n_x/n_y distribution if the basis functions of the run in Fig. 4 are directly matched by Gabors (without convoluting them before).

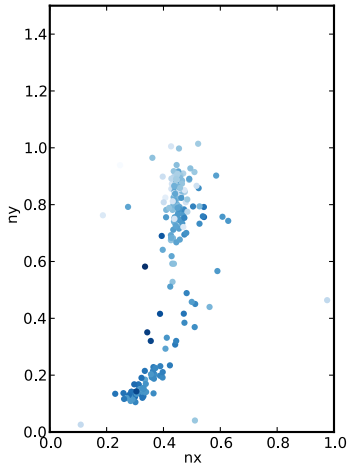

Figure C.3: n_x/n_y distribution of the (convoluted) basis functions for a run with $H = 200$. All other parameters were set to the same values as described in Sec. 3.