Unsupervised Models of Images by Spike-and-Slab RBMs

Aaron Courville AARON.COURVILLE@UMONTREAL.CA
James Bergstra JAMES.BERGSTRA@UMONTREAL.CA
Yoshua Bengio YOSHUA.BENGIO@UMONTREAL.CA
DIRO, Université de Montréal, Montréal, Québec, Canada

Abstract

The *spike-and-slab* Restricted Boltzmann Machine (RBM) is defined by having both a real valued "slab" variable and a binary "spike" variable associated with each unit in the hidden layer. In this paper we generalize and extend the spike-and-slab RBM to include non-zero means of the conditional distribution over the observed variables given the binary spike variables. We also introduce a term, quadratic in the observed data that we exploit to guarantee that all conditionals associated with the model are well defined – a guarantee that was absent in the original spike-and-slab RBM. The inclusion of these generalizations improves the performance of the spike-and-slab RBM as a feature learner and achieves competitive performance on the CIFAR-10 image classification task. The spike-and-slab model, when trained in a convolutional configuration, can generate sensible samples that demonstrate that the model has captured the broad statistical structure of natural images.

1. Introduction

Recently, there has been considerable interest in the problem of unsupervised learning of features for supervised tasks in natural image domains. Approaches based on unsupervised *pretraining* followed by either whole-model *finetuning* or simply the linear classification of features dominate in benchmark tasks such as CIFAR-10 (Krizhevsky, 2009). One of most popular energy-based modelling paradigms for unsupervised feature learning is the Restricted Boltzmann Machine (RBM). An RBM is a Markov random field with a bipartite graph structure consisting of a visible layer and a hidden layer. The bipartite structure excludes connections between the variables within each layer so that the latent variables are conditionally independent given the visible variables and vice versa. The factorial nature of these conditional distributions enables efficient Gibbs sampling which forms the basis of the most widespread RBM learning algorithms such as contrastive divergence (Hinton, 2002) and stochastic maximum likelihood (Tieleman, 2008).

The spike-and-slab RBM (ssRBM) (Courville et al., 2011) departs from other similar RBM-based models in the way the hidden layer latent units are defined. They are modelled as the element-wise product of a real valued vector with a binary vector, i.e., each hidden unit is associated with a binary *spike* variable and a real-valued *slab* variable. The name *spike and slab* is inspired from terminology in the statistics literature (Mitchell & Beauchamp, 1988), where the term refers to a prior consisting of a mixture between two components: the spike, a discrete probability mass at zero; and the slab, a density (typically uniformly distributed) over a continuous domain.

In this paper, we introduce a generalization of the ssRBM model, which we refer to as the μ-ssRBM. Relative to the original ssRBM, the μ-ssRBM includes additional terms in the energy function which give extra modelling capacity. One of the additional terms allows the model to form a conditional distribution of the spike variables (by marginalizing out the slab variables, given an observation) that is similar to the corresponding conditional of the recently introduced mean covariance RBM (mcRBM) (Ranzato & Hinton, 2010) and mPoT model (Ranzato et al., 2010). Conditional on both the observed and spike variables, the μ-ssRBM slab variables and input are jointly Gaussian with diagonal covariance matrix; conditional on both the spike and slab variables, the observations are Gaussian with diagonal covariance. Thus, unlike the mcRBM or the more recent mPoT model, the μ-

Appearing in *Proceedings of the 28th International Conference on Machine Learning*, Bellevue, WA, USA, 2011.

ssRBM is amenable to simple and efficient Gibbs sampling. This property of the ssRBM makes the model an excellent candidate as a building block for the development of more sophisticated models such as Deep Boltzmann Machines (Salakhutdinov & Hinton, 2009).

One potential drawback of the ssRBM is the lack of a guarantee that the resulting model constitutes a valid density over the whole real-valued data space. In this paper, we develop several strategies that guarantee all conditionals are well defined by adding energy terms to the μ-ssRBM. However, experimentally we find that loosening the constraint yields better models.

2. The μ-Spike-and-Slab RBM

The μ-ssRBM describes the interaction between three random vectors: the visible vector v representing the observed data, the binary "spike" variables h and the real-valued "slab" variables s. The ith hidden unit is associated both with an element h_i of the binary vector and with an element s_i of the real-valued variable. Suppose there are N hidden units: $h \in [0,1]^N$, $s \in \mathbb{R}^N$ and a visible vector of dimension D: $v \in \mathbb{R}^D$. The μ-ssRBM model is defined via the energy function:

$$E(v,s,h) = -\sum_{i=1}^{N} v^T W_i s_i h_i \; + \frac{1}{2} v^T \left(\Lambda + \sum_{i=1}^{N} \Phi_i h_i \right) v + \frac{1}{2}\sum_{i=1}^{N} \alpha_i s_i^2 \; - \sum_{i=1}^{N} \alpha_i \mu_i s_i h_i \; - \sum_{i=1}^{N} b_i h_i + \sum_{i=1}^{N} \alpha_i \mu_i^2 h_i, \tag{1}$$

in which W_i denotes the ith weight vector ($W_i \in \mathbb{R}^D$), each b_i is a scalar bias associated with h_i, each α_i is a scalar that penalizes large values of s_i^2, and Λ is a diagonal matrix that penalizes large values of $\|v\|_2^2$. In comparison with the original ssRBM (Courville et al., 2011), the μ-ssRBM energy function includes three additional terms. First, the $1/2\ v^T \left(\sum_{i=1}^N \Phi_i h_i\right) v$ term, with non-negative diagonal matrices Φ_i, $i \in [1,N]$, establishes an h-dependent quadratic penalty on v. Second, associated with each slab variable is a mean parameter μ – from which the μ-ssRBM takes its name. Finally, the $\sum_{i=1}^N \alpha_i \mu_i^2 h_i$ term acts as an additional bias term for the h_i, which we include to simplify the parametrization of the conditionals. In addition to offering additional flexibility to model the statistics of natural images, the inclusion of the parameters $\mu = [\mu_1,\dots,\mu_N]$ and $\Phi = [\Phi_1,\dots,\Phi_N]$ also allows us to derive constraints on the model that ensure that the model remains well-behaved over the entire data domain of $\mathbb{R}^D$.

The joint probability distribution over v, $s = [s_1,\dots,s_N]$ and $h = [h_1,\dots,h_N]$ is defined as:

$$p(v,s,h) = \frac{1}{Z}\exp\{-E(v,s,h)\} \tag{2}$$

where Z is the normalizing partition function. We can think of the distribution presented in Eqns. 1 and 2 as associated with the standard RBM bipartite graph structure with the distinction that the hidden layer is composed of an element-wise product of the random vectors s and h.

To gain insight into the μ-ssRBM model, we will look at the conditional distributions $p(v \mid s,h)$, $p(s \mid v,h)$, $P(h \mid v)$ and $p(v \mid h)$. First, we consider the conditional $p(v \mid s,h)$:

$$\begin{aligned} p(v \mid s,h) &= \frac{1}{p(s,h)}\frac{1}{Z}\exp\{-E(v,s,h)\} \\ &= \mathcal{N}\left(C_{v|s,h}\sum_{i=1}^{N} W_i s_i h_i \;,\; C_{v|s,h} \right) \end{aligned} \tag{3}$$

where $C_{v|s,h} = \left(\Lambda + \sum_{i=1}^N \Phi_i h_i\right)^{-1}$. The conditional distribution of v given both s and h is a Gaussian with mean $C_{v|s,h}\sum_{i=1}^N W_i s_i h_i$ and covariance $C_{v|s,h}$. Since Λ and Φ_i are diagonal ($\forall\ i \in [1,N]$), the covariance matrix of $p(v \mid s,h)$ is also diagonal. Eqn. 3 shows the role played by the Φ_i in augmenting the precision with the activation of h_i. Indeed, hidden unit i contributes a component not only to the mean proportional to $W_i s_i$, but also to the global scaling of the conditional mean.

The conditional over the slab variables s given the spike variables h and the visible units v is given by:

$$p(s \mid v,h) = \prod_{i=1}^{N} \mathcal{N}\left(\left(\alpha_i^{-1} v^T W_i + \mu_i\right) h_i \;,\; \alpha_i^{-1} \right). \tag{4}$$

As was the case with the conditional $p(v \mid s,h)$, deriving the conditional $p(s \mid v,h)$ from the joint distribution in Eqn. 2 reveals a Gaussian distribution with diagonal covariance. Eqn. 4 also shows how the mean of the slab variable s_i, given $h_i = 1$, is linearly dependent on v, and as the precision $\alpha_i \to \infty$, s_i converges in probability to μ_i.

Marginalizing out the slab variables s yields the traditional RBM conditionals $p(v \mid h)$ and $p(h \mid v)$. The conditional $p(v \mid h)$ is also Gaussian,

$$\begin{aligned} p(v \mid h) &= \frac{1}{P(h)}\frac{1}{Z}\int \exp\{-E(v,s,h)\}\ ds \\ &= \mathcal{N}\left(C_{v|h}\sum_{i=1}^{N} W_i \mu_i h_i \;,\; C_{v|h} \right) \end{aligned} \tag{5}$$

where $C_{v|h} = \left(\Lambda + \sum_{i=1}^N \Phi_i h_i - \sum_{i=1}^N \alpha_i^{-1} h_i W_i W_i^T\right)^{-1}$, the last equality holds only if the covariance matrix

$C_{v|h}$ is positive definite. Note that the covariance is not obviously parametrized to guarantee that it is positive definite. In Section 3, we will discuss strategies to ensure that $C_{v|h}$ be positive definite via constraints on Λ and Φ.

In marginalizing over s, the visible vector v remains Gaussian-distributed, but the parametrization has changed. While the distribution $p(v \mid s, h)$ uses h with s to parametrize the conditional mean with a corresponding diagonal covariance, the conditional $p(v \mid h)$ uses h to parametrize a covariance matrix that is non-diagonal due to the $\sum_{i=1}^{N} \alpha_i^{-1} h_i W_i W_i^T$ term, and a conditional mean mediated by μ.

A closer look at Eqn. 5 reveals an important aspect of the inductive bias of the model. The conditional mean of v given h and principal axis of conditional covariance are generally in a similar direction, and if $\left(\Lambda + \sum_{i=1}^{N} \Phi_i h_i\right)$ is a scalar matrix (equivalent to scalar × Identity) the two vectors are in exactly the same direction. Having the principal component of the conditional covariance in the same direction as the mean has the property of $p(v \mid h)$ being maximally invariant to changes in the norm $\|v\|_2$. This is a desirable property for a model of natural images where the norm is particularly sensitive to illumination conditions and image contrast levels – factors that are often irrelevant to tasks of interest such as object classification.

The final conditional that we will consider is the distribution over the latent spike variables h given the visible vector, $P(h \mid v) = \prod_i^N P(h_i \mid v)$ and

$$\begin{aligned} P(h_i = 1 \mid v) &= \frac{1}{p(v)} \frac{1}{Z_i} \int \exp\left\{-E(v, s, h)\right\} \; ds \\ &= \sigma\left(\frac{1}{2}\alpha_i^{-1}(v^T W_i)^2 + v^T W_i \mu_i - \frac{1}{2} v^T \Phi_i v + b_i\right), \quad (6) \end{aligned}$$

where σ represents a logistic sigmoid. As with the conditionals $p(v \mid s, h)$ and $p(s \mid v, h)$, the distribution of h given v factorizes over the elements of h. Eqn. 6 shows the interaction between three data-dependent terms. The first term, $\frac{1}{2}\alpha_i^{-1}(v^T W_i)^2$, is the contribution due to the variance in s about its mean (note the scaling with α_i^{-1}) and appears in the sigmoid as a result of marginalizing out s. This term is always non-negative, so it always acts to increase $P(h_i \mid v)$. Countering this tendency to activate h_i is the other term quadratic in v, $-\frac{1}{2} v^T \Phi_i v$, that is always a non-positive contribution to the sigmoid argument. In addition to these two quadratic terms, there is the term $v^T W_i \mu_i$ whose behaviour mimics the data-dependent term in the analogous GRBM version of the conditional distribution over h: $P_{\mathrm{G}}(h_i \mid v) = \sigma\left(v^T W_i + b_i\right)$.

Another perspective on the behaviour of $p(h_i \mid v)$ as a function of v is gained by considering an alternative arrangement of the terms:

$$P(h_i = 1 \mid v) = \sigma(\hat{b}_i - \frac{1}{2}(v - \xi_{v|h_i})^T C_{v|h_i}^{-1}(v - \xi_{v|h_i})) \quad (7)$$

in which $C_{v|h_i} = \left(\Phi_i - \alpha_i^{-1} W_i W_i^T\right)^{-1}$, $\xi_{v|h_i} = C_{v|h_i} W_i \mu_i$ and $\hat{b}_i$ is a re-parameterization of the bias that incorporates the remainder of the completion of the square. It is evident from this form of $P(h_i = 1 \mid v)$ that, in the event that the matrix $C_{v|h_i}$ is positive definite, then $P(h_i \mid v)$ reaches its maximum when $v = C_{v|h_i} W_i \mu_i$. We can easily confirm that the tail behaviour, as v departs from its maximum, is Gaussian:

$$\sigma(-x^2) = \frac{\exp(-x^2)}{1 + \exp(-x^2)} \underset{x \to \infty}{\rightarrow} \exp(-x^2)$$

We will look to take advantage of the μ-ssRBM relationship between Φ_i and $\alpha_i^{-1} W_i W_i^T$ when we consider ways to constrain the covariance of $p(v \mid h)$ to be positive definite in Section 3.

To complete the exposition of the basic μ-ssRBM model, we present the free energy $f(v)$ of the visible vector.

$$\begin{aligned} f(v) &= -\log \sum_h \int \exp\left\{-E(v, s, h)\right\} \; ds \\ &= \frac{1}{2} v^T \Lambda v - \frac{1}{2} \sum_{i=1}^{N} \log\left(2\pi\alpha_i^{-1}\right) \\ &- \sum_{i=1}^{N} \log\left[1 + \exp\left\{-\frac{1}{2} v^T C_{v|h_i}^{-1} v + v^T W_i \mu_i + b_i\right\}\right] \end{aligned}$$

3. Positive Definite Parameterizations of the μ-ssRBM

Equation 5 reveals an important property of the μ-ssRBM model. The conditional $p(v \mid h)$ is only a well-defined Gaussian distribution if the covariance matrix $C_{v|h}$ is positive definite (PD). However, the covariance matrix is not parametrized to guarantee that this condition is met. If there exists a vector x such that $x^T C_{v|h} x \leq 0$, then the covariance matrix is not positive definite. In the original presentation of the ssRBM in Courville et al. (2011), the possibility of the non-positive-definiteness of the conditional covariance of v given h was dealt with by limiting the support over the domain of v (i.e., $\mathbb{R}^D$) to a large but finite ball that encompasses all the training data. Such an approach is feasible but unsatisfying as the model would naturally tend to build-up probability density close to this arbitrary boundary – a region of the input space that should have low density.

Here, in the context of the μ-ssRBM model, we turn to the question of how we can constrain the model parameters to guarantee that the model remains well-behaved (i.e., all conditionals are well-defined probability densities). The problem we face is ensuring that the covariance or equivalently the precision matrix of $p(v \mid h)$ is positive definite, i.e., we wish to satisfy the constraint:

$$x^T C_{v|h}^{-1} x > 0 \quad \forall x \neq \mathbf{0}$$

To satisfy this constraint, we need to ensure that $\Lambda + \sum_{i=1}^N \Phi_i h_i$ is large enough to offset $\sum_{i=1}^N \alpha_i^{-1} W_i W_i^T h_i$. We consider two basic strategies: (1) define Λ to be large enough to offset a worst-case setting of the h; and (2) define the Φ_i to ensure that the contribution of each active h_i is itself PD.

3.1. Constraining Λ

One option to ensure that $C_{v|h}$ remains PD for all patterns of h activation is to constrain Λ to be large enough. In setting a constraint on Λ, we will ignore the contribution of the Φ_i terms (which leads to non-tightness of the constraint). Since the contribution of every $\alpha_i^{-1} W_i W_i^T h_i$ term is negative semi-definite, the worst case setting of the h would be to have $h_i = 1$ for all $i \in [1, N]$. This implies that Λ must be constrained such that:

$$x^T \left(\Lambda - \sum_{i=1}^N \alpha_i^{-1} W_i W_i^T \right) x > 0 \quad \forall x \neq \mathbf{0}. \tag{8}$$

If we take Λ to be a scalar matrix: $\Lambda = \lambda I$, then the problem of enforcing a PD precision matrix reduces to ensuring that λ is greater than the maximum eigenvalue ρ of $\sum_{i=1}^N \alpha_i^{-1} W_i W_i^T$. In practice we can use the power iteration method to quickly estimate an upper bound on the maximum eigenvalue, and then constrain $\lambda > \rho$ throughout training.

3.2. Constraining Φ

Another option to ensure that $C_{v|h}$ remains PD for all patterns of h activation is to constrain Φ_i to be large enough to ensure that the contribution of each h_i is PD. Let W_{ij} be the jth element of the filter W_i (or equivalently, the ijth element of the weight matrix W) and let Φ_{ij} denote the jjth element of the diagonal Φ_i matrix. We want to choose Φ_i such that:

$$J(x, \Phi_i) = \sum_j x_j^2 \Phi_{ij} - (\sum_j w_{ij} x_j)^2 > 0 \quad \forall x \in \mathbb{R}^D, x \neq \mathbf{0}.$$

This condition will be satisfied for all x if we can satisfy it for the $x = u$ of norm 1 that minimizes $J(x, \Phi_i)$ (u is the eigenvector of the smallest eigenvalue of the matrix $\left(\Phi_i - \alpha_i^{-1} W_i W_i^T\right)$). To find u, we define the Lagrangian

$$L(x, \Phi_i) = J(x, \Phi_i) + \eta(1 - \sum_j x_j^2)$$

to enforce the constraint $\sum_j x_j^2 = 1$. Setting $\frac{\partial L}{\partial x_j} = 0$ we recover the set of constraints:

$$\begin{aligned} \sum_j \frac{W_{ij}^2}{\Phi_{ij} - \eta} &= 1 \qquad (9) \\ \sum_j \frac{W_i^2 \Phi_{ij}}{(\Phi_{ij} - \eta)^2} &> 1 \\ \eta &> 0 \end{aligned}$$

Consider the following general parametrization of Φ_i:

$$\Phi_{ij} = \eta + q\alpha^{-1} \frac{W_{ij}^2}{\beta_{ij}}.$$

This particular form is chosen so that our constraint set gives us $q = \sum_j \beta_{ij}$ and $\beta_{ij} > 0$. So with Φ_{ij} parametrized as

$$\Phi_{ij} = \zeta_{ij} + \alpha^{-1} \frac{W_{ij}^2}{\beta_{ij} / \sum_j \beta_{ij}}$$

with $\zeta_{ij} > 0$, the covariance matrix of $p(v \mid h)$ is guaranteed to be PD. The parameter ζ_{ij} is an extra degree of freedom to Φ_{ij} to be estimated through maximum likelihood learning.

We are free to choose the parametrization of the β_{ij} provided $\beta_{ij} > 0$. For example, with the choice $\beta_{ij} = W_{ij}^2$, Φ_i simplifies to

$$\Phi_{ij} = \zeta_{ij} + \alpha^{-1} \sum_j W_{ij}^2 I \tag{10}$$

where Φ_{ij} takes the form of a scalar matrix. Alternatively, we could choose $\beta_{ij} = 1$ with the result that

$$\Phi_{ij} = \zeta_{ij} + \alpha^{-1} D W_{ij}^2 \tag{11}$$

where the jth elements on the diagonal of Φ_i is scaled with W_{ij}^2.

While we are free to chose $\beta_{ij} > 0$ as we would like, the decision affects the inductive bias of the model. In the case of the Φ_{ij} parametrization given in Eqn. 10, the presence of the $\sum_i^N \Phi_i h_i$ as a scaling on the mean of the conditional $p(v \mid s, h)$ (Eqn. 3) implies that the activation of any h_i will have an effect on the scaling of the mean across the entire visible vector (or layer) irrespective of how localized is the corresponding filter

W_i. Unsurprisingly, use of this parametrization tends to encourage both sparsely active h_i and W_i having relatively large receptive fields.

The Φ_i parametrization via Eqn. 11 has the property that the Φ_i receptive fields are steered in the direction of W_i. Where W_i is near zero, Φ_i has little effect (unless it is mediated by ζ_i). This is an appealing property for modeling images or other data that give rise to sparse receptive fields W_i.

3.3. Comparing strategies

The two general strategies to guarantee that the covariance matrix of $p(v \mid h)$ is positive definite are in some sense complementary. In the sparse operating regime of the μ-ssRBM (most h_i are inactive over most of the dataset), the Λ worst case assumption that $\forall i : h_i = 1$, becomes increasingly inaccurate and, as a result, the constraint on Λ becomes increasingly conservative. Therefore in the sparse regime, the Φ constraints would seem more appropriate. On the other hand, in a highly non-sparse regime, the individual contributions of the Φ to the global precision matrix can combine to form a more conservative PD precision matrix than would result from a constrained Λ.

It is also possible to distribute responsibility for ensuring the constraint is satisfied jointly to Λ and Φ. This would constitute a mixed strategy, apportioning responsibility for compensating for the negative definite $-\sum_{i=1}^{N} \alpha_i^{-1} W_i W_i^T h_i$ term to both Λ and Φ.

4. μ-ssRBM Learning and Inference

Learning and inference in the μ-ssRBM proceeds analogously to the original ssRBM and is rooted in the ability to efficiently draw samples from the model via Gibbs sampling. As with the original ssRBM, we seek a set of conditionals that will enable simple and efficient Gibbs sampling. Since sampling from the conditional $p(v \mid h)$ would involve the computationally prohibitive step of inverting a non-diagonal covariance matrix, we pursue a strategy of alternating sampling from the conditionals $P(h \mid v)$, $p(s \mid v, h)$ and $p(v \mid s, h)$. Each of these conditionals has the property that the distribution factors over the elements of the random vector, allowing us to efficiently samples.

In training the μ-ssRBM, we use the stochastic maximum likelihood algorithm (SML, also known as persistent contrastive divergence) (Tieleman, 2008), where only one or a few Markov Chain (Gibbs) simulations are performed between each parameter update. These samples are then used to approximate the expectations over the model distribution $p(v, s, h)$.

The data log likelihood gradient, $\frac{\partial}{\partial \theta_i}\left(\sum_{t=1}^{T} \log p(v_t)\right)$, is:

$$-\sum_{t=1}^{T} \left\langle \frac{\partial}{\partial \theta_i} E(v_t, s, h) \right\rangle_{p(s,h|v_t)} + T \left\langle \frac{\partial}{\partial \theta_i} E(v, s, h) \right\rangle_{p(v,s,h)}$$

The log likelihood gradient takes the form of a difference between two expectations, with the expectations over $p(s, h \mid v_t)$ in the "clamped" condition, and the expectation over $p(v, s, h)$ in the "unclamped" condition. As with the standard RBM, the expectations over $p(s, h \mid v_t)$ are amenable to analytic evaluation.

5. Comparison to Previous Work

There is now a significant body of work on modelling natural images with RBM-based models. The closest connection to this work is obviously to the original ssRBM (Courville et al., 2011) which we recover by setting the μ-ssRMB parameters $\mu_i = 0$ and $\Phi_i = \mathbf{0}$ for all $i \in [1, N]$. A slightly less obvious limiting case of the μ-ssRMB is the Gaussian RBM (GRBM). Setting Φ_i to be proportional to α_i^{-1} (as discussed in Section 3) and taking $\alpha_i = \alpha \to \infty$, we define a Dirac in s about μ, In this limit, the conditionals $p(v \mid h)$ and $P(h_i = 1 \mid v)$ (Eqns. 5 and 6 respectively) are given by:

$$\begin{aligned}
\lim_{\alpha \to \infty} p(v \mid h) &= \mathcal{N}\left(\Lambda^{-1} \sum_{i=1}^{N} W_i \mu_i h_i \quad \Lambda^{-1}\right) \\
\lim_{\alpha \to \infty} P(h_i = 1 \mid v) &= \sigma\left(v^T W_i \mu_i + b_i\right)
\end{aligned}$$

If we fix $\mu = 1$ and $\Lambda = I$, we recover the Gaussian RBM conditionals. Note that the connection between the μ-ssRBM and the Gaussian RBM is mediated entirely by the μ parameter, the original ssRBM has no such connection with the GRBM.

Beyond the ssRBM, the closest models to the μ-ssRBM are the mean and covariance RBM (mcRBM) (Ranzato & Hinton, 2010) and the mean Product of t-distributions model (mPoT) (Ranzato et al., 2010). Like the μ-ssRBM, both of these are energy-based models where the conditional distribution over the visible units conditioned on the hidden variables is a multivariate Gaussian with nonzero mean and a non-diagonal covariance matrix. However the μ-ssRBM differs from these models in the way the conditional means and covariance interact. In both the mcRBM and the mPoT model, the means are modelled by the introduction of additional GRBM hidden units, whereas in the μ-ssRBM, each hidden unit can potentially contribute to both the conditional mean and covariance. For the ith hidden unit, the contribution to each is controlled by the relative values of the μ_i and α_i. With large $|\mu_i|$ and small α_i, the unit predominantly contributes to the conditional mean. Conversely, with large α_i and small $|\mu_i|$, the unit mostly contributes to the conditional covariance. The advantage of the μ-ssRBM approach is that we are able to save a hyper-parameter by letting maximum likelihood induction optimize the trade-off between mean and covariance modeling. Empirically, we find that while some hidden units learn to focus exclusively on the conditional covariance (with $\mu_i \approx 0$); most units take advantage of the flexibility offered in the μ-ssRBM framework and contribute to both the conditional mean and covariance.

The mPoT and mcRBM also differ from the μ-ssRBM in how they parametrize the visible covariance. While the μ-ssRBM uses h activations to pinch the precision matrix along the direction specified by the corresponding weight vector, both the mcRBM and the mPoT models use their latent variable activations to maintain constraints, decreasing in value to allow variance in the direction of the corresponding weight vector.

Interestingly, the μ-ssRBM term involving Φ_i is very similar to the covariance term in the mcRBM energy function. The difference is that our restriction to a diagonal Φ_i significantly limits what we can model with it. However, this restricted structure allows us to sample efficiently from the model using Gibbs sampling which is not available to either the mcRBM or the mPoT model. In addition, the roles of these covariance terms are also quite different. In the mcRBM, the covariance term is associated with the model's feature vectors. In the μ-ssRBM, we use the Φ_i both to help constrain the model's conditionals to be well-defined, and in conjunction with W, to help maximize the likelihood of the training data.

Finally, the covariance structure of the μ-ssRBM conditional $p(v \mid h)$ (Eqn. 5) is very similar to the product of probabilistic principal components analysis (PoPPCA) model (Williams & Agakov, 2002) with components corresponding to the μ-ssRBM weight vectors associated with the active hidden units ($h_i = 1$).

6. Experiments

We demonstrate the utility of the μ-ssRBM on the CIFAR-10 dataset by classifying images and by sampling from the model. In particular, our experiments are directed toward exploring the properties of the different elements of the model, including the roles of μ and Φ and the effects of the various Λ and Φ PD constraints.

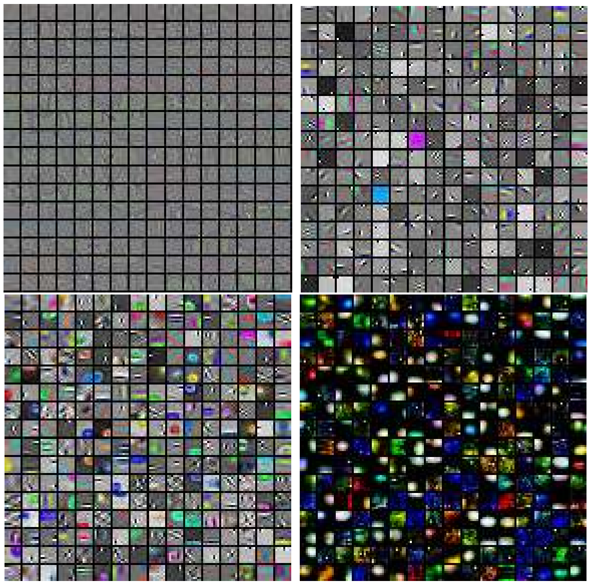

Figure 1. (Top left) ZCA-whitened data used for patchwise training. (Top right) Filters W learnt when μ and Φ were fixed at zero. These filters produce edges similar to many other models, and neatly separate black-and-white edges from colour ones. Filters W (bottom left) and Φ (bottom right) learnt when μ and Φ are fit to the data. The combination of W and Φ gives individual units more flexibility, and gives rise to a richer variety of features.

Our experiments are based on the CIFAR-10 image classification dataset consisting of 40 000 training images, 10 000 validation images, and 10 000 test images. The images are 32-by-32 pixel RGB images. Each image is labelled with one of ten object categories (aeroplane, automobile, bird, cat, deer, dog, frog, horse, ship, truck) according to the most prominent object in the image.

Classification: We evaluate the μ-ssRBM as a feature-extraction algorithm by plugging it into the classification pipeline developed by Coates et al. (2011). In broad strokes, the μ-ssRBM is fit to (192-dimensional) 8x8 RGB image patches, and then applied convolutionally to the 32x32 images. The image patches (starting from pixels between 0 and 255) on which the μ-ssRBM was trained were centered, and then normalized by dividing by the square root of their variance plus a noise-cancelling constant (10). The normalized patches were whitened by ZCA (Hyvärinen & Oja, 2000) with a small positive constant (0.1) added to all eigenvalues. The resulting patches (Figure 1, top left) are mostly grey with high spatial frequencies amplified, and lower spatial frequencies attenuated. Our models were trained from the 16 non-overlapping 8x8 patches from each of the first 10 000

Table 1. The performance of μ-ssRBM variants with 256 hidden units in CIFAR-10 image classification (± 95% confidence intervals). "no PD" = without PD constraint.

Model		Accuracy (%)
no PD, μ free, Φ free		73.1 ± 0.9
no PD, μ free, $\Phi = \mathbf{0}$		71.43 ± 0.9
no PD, $\mu = \mathbf{0}$, Φ free		71.19 ± 0.9
no PD, $\mu = \mathbf{0}$, $\Phi = \mathbf{0}$		68.92 ± 0.9
PD by Diag. W	(Eqn. 11)	69.1 ± 0.9
PD by Λ	(Eqn. 8)	68.3 ± 0.9
PD by scal. mat.	(Eqn. 10)	67.1 ± 0.9

training set images in CIFAR-10 (for a total of 160 000 training examples).

Models were trained for one hundred thousand mini-batches of 100 patches. On an NVIDIA GTX 285 GPU this training took on the order of 15 minutes for most models. We used SML training (Tieleman, 2008). Classification was done with an ℓ_2-regularized SVM. The SVM was applied to the conditional mean value of latent spike (h) variables, extracted from every 8x8 image patch in the 32x32 CIFAR-10 image. Prior to classification, our conditional h values were spatially pooled into 9 regions, analogous to the 4 quadrants employed in Coates et al. (2011). For a model with N hidden units, the classifier operated on a feature vector of $9N$ elements.

Table 1 lists the performance of several variants on the μ-ssRBM model. For this comparison all variants were trained with a small amount of sparsity aimed at maintaining 15% activity, and were configured with 256 hidden units. The lines labelled PD correspond to models that were constrained to have positive definite covariance of $p(v \mid h)$ while the lines labelled no PD are not. If $\mu = \mathbf{0}$ appears in a line the corresponding model was trained with $\mu = 0$ or equivalently with the μ terms removed from the energy function. The nomenclature for Φ is analogous. This implies for instance that the original ssRBM model would correspond to the no PD, $\mu = \mathbf{0}$, $\Phi = \mathbf{0}$ condition.

Table 1 reveals that it is possible to constrain Φ to enforce that $C_{x|h}$ is PD and achieve classification results that match that of the original ssRBM. However, if we take the same μ-ssRBM form and loosen the PD constraint, the model can perform much better. Also of note is that both the μ and the Φ terms seem to contribute approximately equally to improving the classification accuracy.

Table 2 situates the performance of the μ-ssRBM in the literature of results on CIFAR-10. The μ-ssRBM performs better than the most closely-related models - the GRBM, cRBM, and mcRBM. Recent work by Coates et al. (2011) has shown that a feature-extractor based on K-means actually out-performs these energy-based approaches to feature extraction on CIFAR-10, in the limit of very large hidden unit counts. Future work will look at more effective training strategies for energy models in this regime.

Table 2. The performance of μ-ssRBM relative to other models in the literature for CIFAR-10 (± 95% confidence interval). The k-means results are taken from Coates et al. (2011), the "conv. trained DBN" result is the convolutionally trained two-layer Deep Belief Network (DBN) with rectified linear units, reported in Krizhevsky (2010), and the GRBM, cRBM, mcRBM results are taken from Ranzato & Hinton (2010)

Model	Accuracy (%)
k-means (4000 units)	79.6 ± 0.9
conv. trained DBN	78.9 ± 0.9
μ-ssRBM (4096 units)	76.7 ± 0.9
k-means (1200 units)	76.2 ± 0.9
μ-ssRBM (1024 units)	76.2 ± 0.9
k-means (800 units)	75.3 ± 0.9
μ-ssRBM (512 units)	74.1 ± 0.9
μ-ssRBM (256 units)	73.1 ± 0.9
k-means (400 units)	72.7 ± 0.9
k-means (200 units)	70.1 ± 0.9
mcRBM (225 factors)	68.2 ± 0.9
cRBM (900 factors)	64.7 ± 0.9
cRBM (225 factors)	63.6 ± 0.9
GRBM	59.7 ± 1.0

Model Samples: To draw samples from the model, we trained it convolutionally, similarly to Krizhevsky (2010). Our convolutional implementation of the μ-ssRBM included 1000 fully-connected units to capture global structure, and 64 hidden units for every position of an 9x9 RGB filter. The model was trained on the CIFAR dataset, centered and globally contrast normalized. Filters W and Φ were shared across the image, though independent scalar-parameters μ_i, α_i, and hidden unit bias b_i were allocated for each individual hidden unit. Figure 2 illustrates some samples drawn from the model. The samples are taken from the negative phase near the end of training (with the learning rate annealed to near zero). These samples exhibit global coherence, and sharp region boundaries. Qualitatively, these samples are more compelling than samples from similar energy-based models, such as those featured in Ranzato et al. (2010).

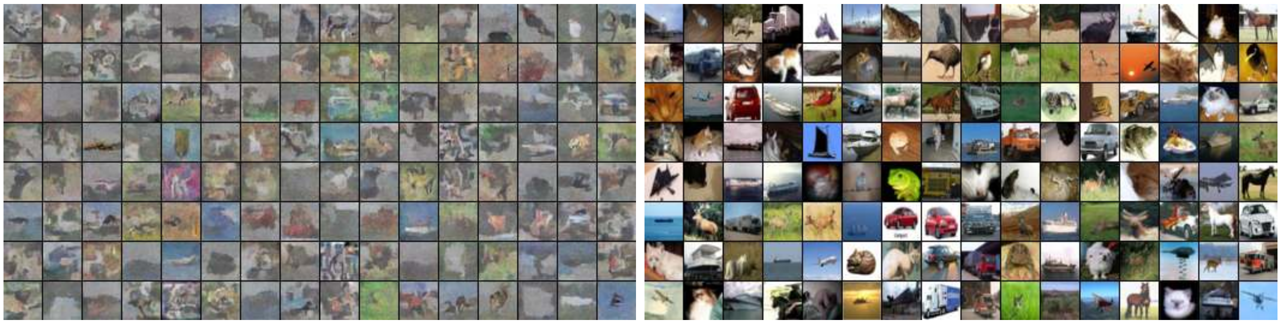

Figure 2. (Left) Samples from a convolutionally trained μ-ssRBM exhibit global coherence, sharp region boundaries, a range of colours, and natural-looking shading. (Right) The images in the CIFAR-10 training set closest (L2 distance with contrast normalized training images) to the corresponding model samples. The model does not appear to be capturing the natural image statistical structure by overfitting particular examples from the dataset.

7. Discussion

In this paper we have introduced the μ-ssRBM, a generalization of the ssRBM that includes extra terms in the energy function. One of these terms permits the model to capture a non-zero mean in the Gaussian conditional $p(v \mid h)$, bringing the ssRBM framework in line with the recent work of Ranzato & Hinton (2010) and Ranzato et al. (2010) which also modelled the conditional of the observed data given the latent variable value to be a general multivariate Gaussian with non-zero mean and full covariance. Unlike these other approaches, instantiating the slab vector s renders the μ-ssRBM amenable to efficient block Gibbs sampling.

The other functional term included in the μ-ssRBM energy function adds a positive definite diagonal contribution to the covariances associated with the Gaussian conditions over the observations. This term was used to define variants of the μ-ssRBM that were constrained to have well-defined conditional.

Still, our techniques for constraining the μ-ssRBM to have PD conditionals are based on loose worst-case scenarios, and potentially leave room for improvement. Our classification experiments indicate that the μ-ssRBM was able to use the extra capacity offered by the addition of these elements to the energy function to improve the classification accuracy. They also show that the addition of the PD constraint comes at the cost of classification performance.

Acknowledgements

We acknowledge NSERC, FQRNT, RQCHP and Compute Canada for their financial and computational support. We also thank Chris Williams for insightful comments on the model and pointing us to the PoPPCA model.

References

Coates, A., Lee, H., and Ng, A. Y. An analysis of single-layer networks in unsupervised feature learning. In *AISTATS'2011*, 2011.

Courville, A., Bergstra, J., and Bengio, Y. The spike and slab restricted Boltzmann machine. In *AISTATS'2011*, 2011.

Hinton, G. E. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14:1771–1800, 2002.

Hyvärinen, A. and Oja, E. Independent component analysis: Algorithms and applications. *Neural Networks*, 13 (4–5):411–430, 2000.

Krizhevsky, A. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Krizhevsky, A. Convolutional deep belief networks on CIFAR-10. Technical report, University of Toronto, 2010.

Mitchell, T. J. and Beauchamp, J. J. Bayesian variable selection in linear regression. *J. Amer. Statistical Assoc.*, 83(404):1023–1032, 1988.

Ranzato, M. and Hinton, G. E. Modeling pixel means and covariance using factorized third-order Boltzmann machines. In *CVPR'10*. IEEE Press, 2010.

Ranzato, M., Mnih, V., and Hinton, G. Generating more realistic images using gated MRF's. In *NIPS'10*, pp. 2002–2010. MIT Press, 2010.

Salakhutdinov, R. and Hinton, G. E. Deep Boltzmann machines. In *AISTATS'2009*, pp. 448–455, 2009.

Tieleman, T. Training restricted Boltzmann machines using approximations to the likelihood gradient. In *ICML 2008*, pp. 1064–1071, 2008.

Williams, C. K. I. and Agakov, F. V. Products of Gaussians and Probabilistic Minor Component Analysis. *Neural Computation*, 14(5):1169–1182, 2002.