ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

An Analysis of Gaussian-Binary Restricted Boltzmann Machines for Natural Images

Nan Wang[1,2] Jan Melchior[1] Laurenz Wiskott[1,2]

1 - Institut für Neuroinformatik
Ruhr-Universität Bochum, 44780 Bochum, Germany

2 - International Graduate School of Neuroscience
Ruhr-Universität Bochum, 44780 Bochum, Germany

Abstract. A Gaussian-binary restricted Boltzmann machine is a widely used energy-based model for continuous data distributions, although many authors reported difficulties in training on natural images. To clarify the model's capabilities and limitations we derive a rewritten formula of the probability density function as a linear superposition of Gaussians. Based on this formula we show how Gaussian-binary RBMs learn natural image statistics. However the probability density function of the model is not a good representation of the data distribution.

1 Introduction

In this paper we present an analysis of Gaussian-binary restricted Boltzmann machines (GB-RBMs) from the density estimation perspective and from the particular perspective of modeling natural image statistics. We find that the marginal probability distribution of the visible units in GB-RBMs can be written as a linear superposition of Gaussians, which are positioned on the vertices of a projected parallelotope, i.e. a parallelepiped in high dimensions. In addition, our analysis suggests that the variance of the visible units in GB-RBMs plays an important role in modeling the input distribution.

GB-RBMs were first proposed by Welling et al. [1]. In practice, Lee et al. proposed to impose a sparse penalty term on the GB-RBMs [2]. However, Krizhevsky succeeded to use GB-RBMs only to extract features from tiny images [3]. Le Roux et al. quantitatively evaluated the model as a generative model [4] and demonstrated the defects of the model from the view of the image reconstruction. Cho et al. addressed the defects by some remedies for the training procedure [5]. Theis et al. further illustrate the defects based on the estimation of loglikelihood [6]. Our analysis and results suggest that GB-RBMs with simple Contrastive Divergence algorithm are capable to learn the independent components as well, even though the learned distribution is not a good representation of the data.

2 Gaussian-Binary RBMs

A GB-RBM is a bipartite graphical model with stochastic visible and hidden variables which are denoted as $\mathbf{X} := (X_1, \ldots, X_M)^T$ and $\mathbf{H} := (H_1, \ldots, H_N)^T$

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

respectively. And the joint probability distribution is defined as:

$$P(\mathbf{X}, \mathbf{H}) := \frac{1}{Z} \mathrm{e}^{-E(\mathbf{X},\mathbf{H})},$$

where $E(\mathbf{X}, \mathbf{H})$ denotes the energy function, which defines the dependence between $\mathbf{X}$ and $\mathbf{H}$. The partition function Z normalizes the probability distribution by summing over all possible states of $\mathbf{X}$ and $\mathbf{H}$ given by:

$$Z := \sum_{\mathbf{x}} \sum_{\mathbf{h}} \mathrm{e}^{-E(\mathbf{x},\mathbf{h})}.$$

And the energy function for GB-RBMs becomes:

$$E(\mathbf{X}, \mathbf{H}): \quad = \quad \frac{||\mathbf{X} - \mathbf{b}||^2}{2\sigma^2} - \mathbf{c}^T\mathbf{H} - \frac{\mathbf{X}^T\mathbf{W}\mathbf{H}}{\sigma^2},$$

where $||\mathbf{X}||^2$ denotes the second norm of the vector $\mathbf{X}$. $\mathbf{W}$ is the weight matrix between M visible units and N hidden units, $\mathbf{b}$ and $\mathbf{c}$ are the bias vectors for visible and hidden units respectively. In addition, the visible units are assumed to have variance, σ^2.

3 Analysis of Gaussian-Binary RBM

In general, we want the model's probability density function (pdf) $P(\mathbf{X})$ to become as close as possible to the data distribution. The pdf of GB-RBMs is usually formulated as a product of experts [7] and can be rewritten as follows:

$$P(\mathbf{X}) = \frac{1}{Z} \mathrm{e}^{-\frac{||\mathbf{X}-\mathbf{b}||^2}{2\sigma^2}} \prod_{j=1}^{N} \left(1 + \mathrm{e}^{c_j + \frac{\mathbf{X}^T\mathbf{w}_j}{\sigma^2}}\right) \quad (1)$$

$$= \eta_0 \mathcal{N}(\mathbf{X}|\mathbf{b}, \sigma) + \sum_{j=1}^{N} \eta_j \mathcal{N}(\mathbf{X}|\mathbf{b} + \mathbf{w}_j, \sigma)$$

$$+ \sum_{j=1}^{N} \sum_{k>j}^{N} \eta_{jk} \mathcal{N}(\mathbf{X}|\mathbf{b} + \mathbf{w}_j + \mathbf{w}_k, \sigma) + \ldots, \quad (2)$$

where

$$\eta_0 = \frac{\left(\sqrt{2\pi\sigma^2}\right)^M}{Z}$$

$$\eta_j = \eta_0 \, \mathrm{e}^{\frac{||\mathbf{b}+\mathbf{w}_j||^2 - ||\mathbf{b}||^2}{2\sigma^2} + c_j}, where\, 1 \leq j \leq N$$

$$\eta_{jk} = \eta_0 \, \mathrm{e}^{\frac{||\mathbf{b}+\mathbf{w}_j+\mathbf{w}_k||^2 - ||\mathbf{b}||^2}{2\sigma^2} + c_j + c_k}, where\, 1 \leq j < k \leq N$$

$$\ldots$$

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

In (1), every expert in GB-RBMs consists of two Gaussian distributions. One of the Gaussians is shifted by the visible bias, $\mathbf{b}$, from the origin. The other one is shifted from the first one by N times the weight vector, $N\mathbf{w}_j$. Both Gaussians share the same variance, $N\sigma^2$. Every expert presents a bimodal distribution. Therefore, their product will form the model distribution, a multimodal distribution with 2^N modes. Similar conclusion was stated independently by Theis et al. [6].

Each Gaussian distribution in (2) is called a *component* of the model distribution and has a *mixing coefficient*. Although all the components have their own means, they have an explicit regularity. The first component is shifted from the origin by the visible bias, $\mathbf{b}$, and named as *anchor component*. Then, there are N components shifted from the anchor one by a single weight vector, $\mathbf{w}_j$. We call them *first order components*. Following these are the j*th order components*, which are shifted from the anchor one by combinations of j weight vectors.

Notice that only the anchor and the first order components are independent, i.e. they can be placed freely in the data space. The positions of other components are just combinations of the $N + 1$ independent ones. Thus the 2^N components are constructed to lie on the vertices of a projected parallelotope. The contour for a GB-RBM with two hidden units are shown in Figure 1.

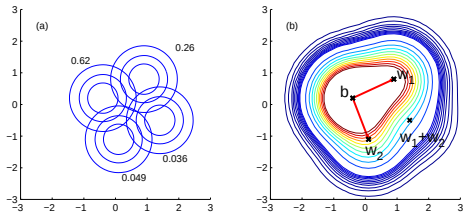

Fig. 1: Illustration of a GB-RBM with 2 hidden units in a two-dimensional space. (a) Contours of constant density for each of the components and coefficients of each components. (b) Contours of the marginal probability density $P(\mathbf{X})$ of the model. The centers of the components are marked with crosses.

The low order components play more important roles in the reconstruction of the data distribution, because they usually have large mixing coefficients. In order to fit to the data, a GB-RBM will firstly try to place its low order components correctly and further damp its high order components by scaling their coefficients down, if they are placed in the non-data area.

The variance, σ^2, in GB-RBMs is usually set to be the same as the variance of the data [2]. However, our analysis above indicates that the variance indeed plays an important role. Take the two dimensional case as an example, the components turn out to be bumps on the surface. With a small variance, the bumps will shrink and can be placed more freely. Conversely, a large variance will result in large bumps. Therefore the model will not have much space to move them within high-density regions. As a result, the model distribution would be

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

more like a monomodal distribution.

4 Experiment 1 - Artificial 2-Dimensional data

In this section, we consider the classic experiment of Independent Component Analysis (ICA). We sampled data from two independent Laplacian distributions. The data was mixed by a mixing matrix, which was generated stochastically. The mixture was whitened by Zero-Phase-Component Analysis (ZCA) so that the joint density had zero mean and unit variance, shown in Figure 2.

We trained a GB-RBM by maximizing the loglikelihood (LL) using Contrastive Divergence-k (CD-k) algorithm [7] with two visible units and two hidden ones. This simple setup allowed us to visualize the distribution of the model.

It is insteresting to see how the GB-RBMs ultilize the Gaussians to model the desired distribution. For different variances, the contours of the learned distributions are plotted in Figure 2. We noted that the distribution of the model depends on the setup of the variance. With small variances, the model places the four components equally scaled at the four corners, the independent components (ICs) of the data distribution. By increasing the variance, the anchor component is placed in the data's mean while two first order components are still located at the ICs but scaled down. The second order component is placed between these two and scaled down even more. While the variance becomes bigger the components will be scaled down further and move to the mean. Finally with variances bigger than one, all components are located at the mean and the pdf is simply the anchor component, all higher order components have been scaled to zero.

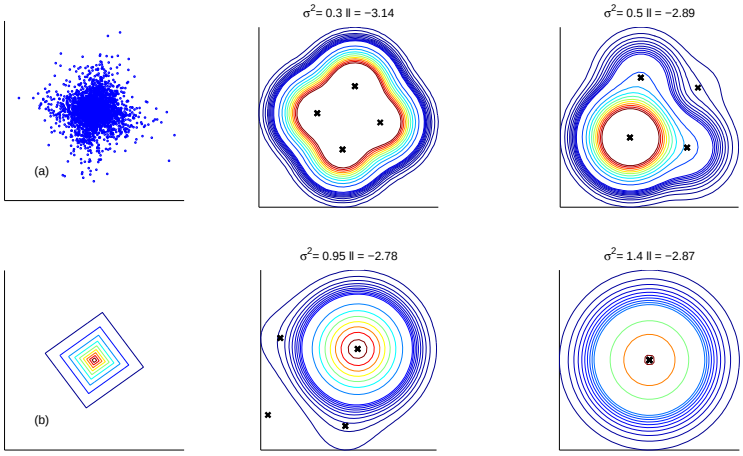

Fig. 2: (left) Plot of the whitened data in (a) and the corresponding pdf in (b). (right) Illustration of GB-RBMs with different variances, σ^2.

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

Likewise, the variance of the model will also affects the LL of the model. In other words, an improper choice of the variance will impair the performance of the model, as shown in Figure 3 (left).

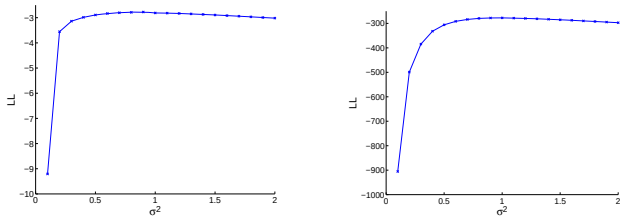

Fig. 3: Plot of LL with different variances in the 2D artificial data (left) and the natural images (right).

5 Experiment 2 - Natural images

To verify the theoretical result of the previous section we trained GB-RBMs with CD-k algorithm on natural image patches sampled from the van Hateren natural image Database [8]. The 70000 gray scale image patches of size of 14x14 pixel were whitened by ZCA. We used 16 hidden units to be able to calculate the partition function in (1) exactly.

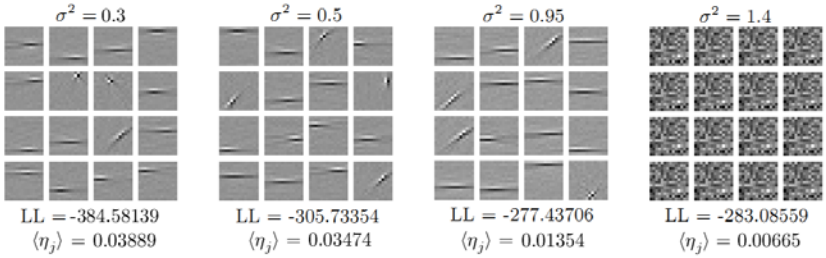

Fig. 4: Filters and corresponding LL of GB-RBMs with 16 hidden units and different variances σ^2 trained on natural image patches. $\langle \eta_j \rangle$ indicate the average relative mixing coefficients for the first order components compared to the anchor component. All four images are normalized to emphasize the filter structure.

Figure 4 shows the learned filters, which are the reshaped columns $\mathbf{w_i}$ of the weight matrix $\mathbf{W}$ and the corresponding LL for different variances. For variances less than one we get independent component (IC) filters comparable to the results shown in [8]. For a variance of one or slightly higher we still get IC filters but some of the units have converged to zero. Variances bigger than one will result in all filters close to zero and therefore noisy shape. The average relative mixing coefficients $\langle \eta_j \rangle$ for the different variances indicates that

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

the components with larger variances are damped more. We get the maximum likelihood at the variance of 0.95, as shown in Figure 3 (right). So the results are consistent with the 2D artificial data and the theoretical analysis.

6 Conclusion and future work

We have shown that a GB-RBM is capable of learning independent components and that the model's performance is highly dependent on the choice of σ^2. We also show that the model's restriction of the single components placed on the vertices of a projected parallelotope prevents the model of learning a good approximation of the true probability distribution. Our future work will be focused on modification of energy function to improve the flexibility of the model.

7 Acknowledgements

We acknowledged Asja Fischer and Oswin Krause for helpful discussions.

References

[1] Max Welling, Michal Rosen-Zvi, and Geoffrey E. Hinton. Exponential family harmoniums with an application to information retrieval. In *Proceedings of the 17th Conference on Neural Information Processing Systems*. MIT Press, December 2004.

[2] Chaitanya Ekanadham. *Sparse deep belief net models for visual area V2*. PhD thesis, Stanford University, 2007.

[3] Alex Krizhevsky. Learning multiple layers of features from tiny images. Master's thesis, University of Toronto, Toronto, April 2009.

[4] Nicolas Le Roux, Nicolas Heess, Jamie Shotton, and John Winn. Learning a generative model of images by factoring appearance and shape. *NEURAL COMPUT*, 23(3):593–650, December 2011.

[5] KyungHyun Cho, Alexander Ilin, and Tapani Raiko. Improved learning of gaussian-bernoulli restricted boltzmann machines. In *Proceedings of the International Conference on Artificial Neural Networks*, pages 10–17, 2011.

[6] Lucas Theis, Sebastian Gerwinn, Fabian Sinz, and Matthias Bethge. In all likelihood, deep belief is not enough. *J MACH LEAN RES*, 12:3071–3096, November 2011.

[7] Geoffrey E. Hinton. Training products of experts by minimizing contrastive divergence. *NEURAL COMPUT*, 14:1771–1800, August 2002.

[8] J. H. van Hateren and A. van der Schaaf. Independent component filters of natural images compared with simple cells in primary visual cortex. *P ROY SOC LOND B BIO*, 265(1394):359–366, 1998.