The Maximal Causes of Binary Data

Jörg Bornschein, Jacquelyn Shelton,
Abdul-Saboor Sheikh, and Jörg Lücke
Frankfurt Institute for Advanced Studies,
Goethe-University, Germany

Introduction

Neural activity encodes multiple-cause stimuli with discrete events. Neurons either spike or remain inactive. Many modeling approaches therefore rely on binary units for encoding. Prominent examples are, for instance, restricted Boltzmann machines [5] and, more recently, deep belief networks [6]. In this work we study a probabilistic generative model with binary units. We investigate the component extraction capabilities of a model with hidden and observed layer both encoding binary data through Bernoulli distributions. In this setting basis functions can not be combined using summation as in sparse coding models [3] but require non-linear combination rules.

BMCA: Binary Maximal Causes Analysis

In this work we explore a generative model with the combination of binary variables using a maximum combination (BMCA), previously used for continuous variables, ([1,2]), and optimize the model parameters using EM. As in standard approaches such as Sparse Coding [3], Independent Component Analysis, or Binary Sparse Coding (BSC), MCA assumes a sparse prior with independent hidden variables.

$$p(\vec{s}\,|\,\Theta) \;=\; \prod_{h=1}^{H} \pi_h^{s_h}(1-\pi_h)^{(1-s_h)},$$

$$p(\vec{y}\,|\,\vec{s},\Theta) \;=\; \prod_{d=1}^{D} W_d(\vec{s})^{y_d}\big(1-W_d(\vec{s})\big)^{(1-y_d)}$$

$$W_d(\vec{s}) \;=\; \max_h\{s_h\,W_{dh}\} \;=\; \lim_{\rho\to\infty}\left(\sum_h (W_{dh}s_h)^{\rho}\right)^{\frac{1}{\rho}}$$

$\vec{y} \in \{0,1\}^D$ observed variables $\quad \pi$ prior parameter
$\vec{s} \in \{0,1\}^H$ hidden variables
$W \in [0,1]^{D\times H}$ basis functions

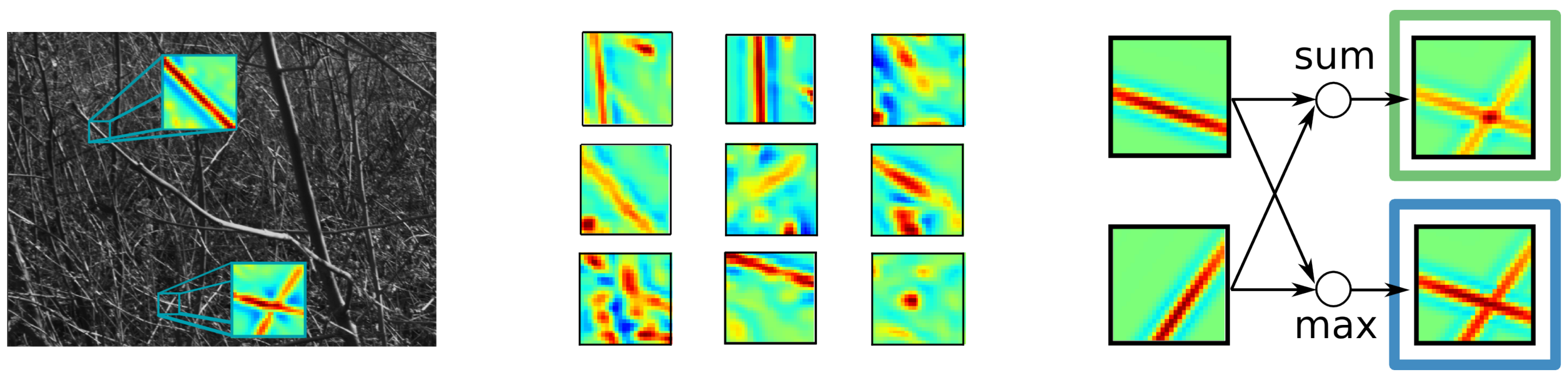
with M-step update equations for basis functions W:

$$W_{dh}^{\text{new}} = \frac{\sum_n \langle \mathcal{C}_{dh}\rangle\, y_d^{(n)}}{\sum_n \langle \mathcal{C}_{dh}\rangle}$$

$$\text{with}\quad \mathcal{C}_{dh} = \frac{s_h}{W_d(\vec{s}) - W_d(\vec{s})^2}\left(\frac{W_{dh}}{W_d(\vec{s})}\right)^{\rho-1}$$

In the place where standard sparse coding approaches, NMF, or ICA use the sum to combine basis functions, BMCA uses a (pixel-wise) maximum operation. To derive tractable approximations for parameter estimation we apply Expectation Truncation (ET; [4]) - a variational EM approach which reduces the hidden space to only those variables contributing most posterior mass. This allows one to infer all model parameters, namely the basis functions W and the degree of sparseness, π. The resulting algorithm is applicable to large-scale problems with hundreds of observed and hidden variables.

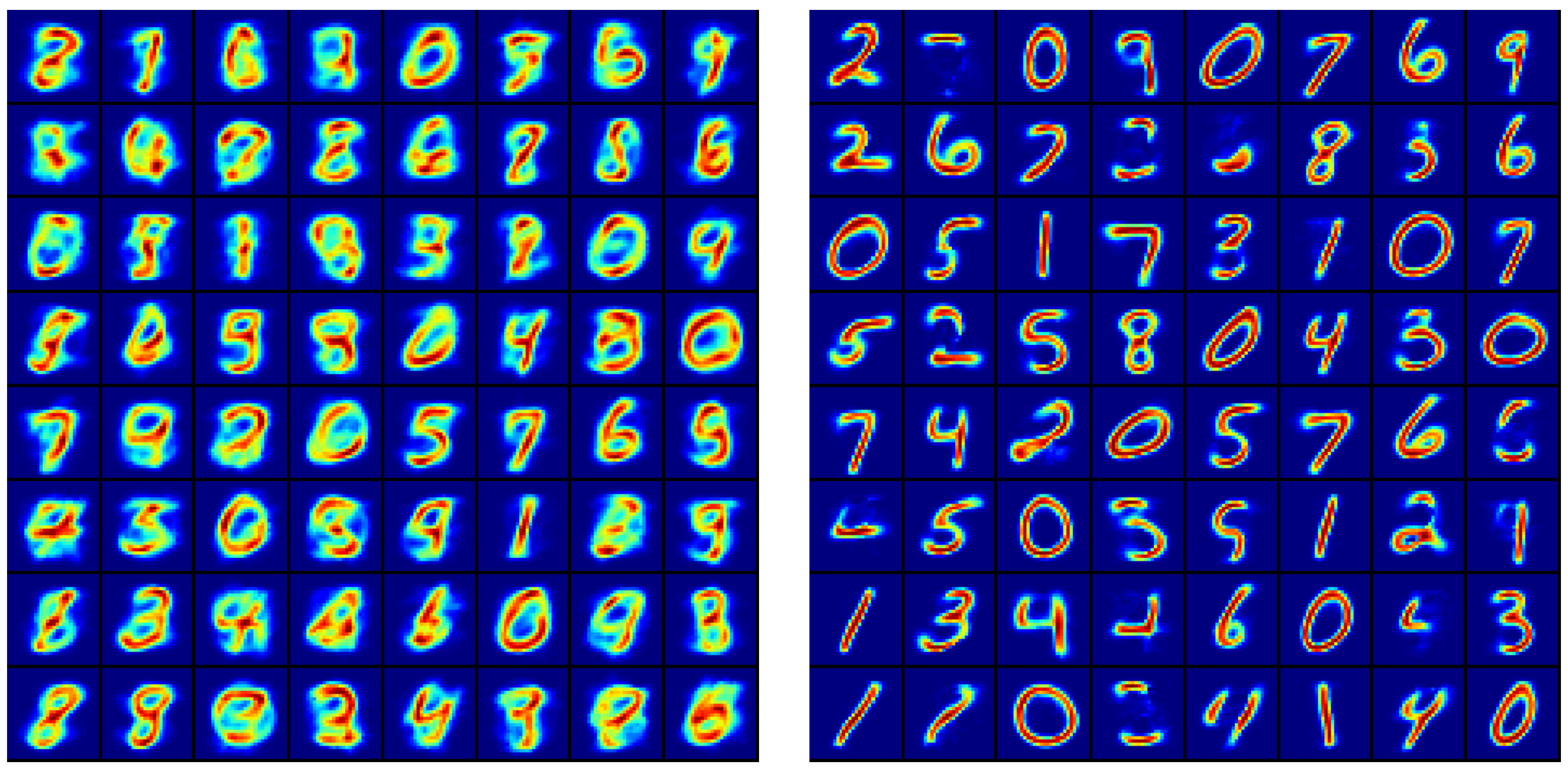
Application to MNIST digits database



The non-linearity of the BMCA is illustrated above using an extracted patch from a natural scene; shown is the contrast between the non-linear max of BMCA ($\max_h\{s_h W_{dh}\}$) and the sum of standard superposition ($\sum_h\{s_h W_{dh}\}$). This may represent a more plausible assumption for the superposition of hand-strokes in the construction of digits.

To study the implications of the non-linear superposition for visual data, the BMCA algorithm was applied to $N = 40\,000$ MNIST digits with $D = 28 \times 28 = 784$ pixels, $H = 64$ hidden units, and $H' = 8$.

Inferred basis functions (H=64):



Basis functions inferred at first EM iteration

Basis functions inferred at final (150th) EM iteration

Conclusions

- This work explored an approach that combines a fast preselection of relevant data components with a subsequent recurrent processing phase (compare [7]), and has recently been linked to neural processing.
- Variational training in the model infers the weights of the connections between the hidden and the observed units as well as the prior activation probabilities of the hidden units.
- In numerical experiments on artificial data and more realistic data, we show that components of mixtures in binary data can successfully be recovered and that such experiments can be scaled to high dimensional observed and hidden spaces.
- Future work will explore different non-linear combinations of basis functions with hidden units, like that used in RBMs.

References

[1] Lücke, J., and Sahani, M. (2008). Maximal Causes for Non-linear Component Extraction, J. Mach. Learn. Res., vol. 9, pp. 1227-1267.
[2] G. Puertas, J. Bornschein, J. Lücke. The Maximal Causes of Natural Scenes are Edge Filters. Proc. NIPS 23: 1939-1947, 2010.
[3] B. A. Olshausen, D. J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature 381:607 – 609, 1996.
[4] J. Lücke, J. Eggert. Expectation Truncation and the Benefits of Preselection in Training Generative Models. JMLR: 11:2855–2900, 2010.
[5] Hinton, G. E. (2002). Training products of experts by minimizing contrastive divergence. Neural Computation, 14(8):1711-1800.
[6] Hinton, G. E, Osindero, S., and Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. Neural Computation, 18:1527-1554.
[7] Supèr, H., Spekreijse, H., and Lamme, V. A. F. (2001). Two distinct modes of sensory processing observed in monkey primary visual cortex (V1). Nature Neuroscience 4, 304 - 310.

This project was supported by the German Federal Ministry of Education and Research (BMBF) within the Bernstein Focus: Neurotechnology Frankfurt through research grant 01GQ0840 and by the German Research Foundation (DFG) in the project LU 1196/4-1. Furthermore, we gratefully acknowledge support by the Frankfurt Center for Scientific Computing.