Modeling receptive fields with non-negative sparse coding

Patrik O. Hoyer

Neural Networks Research Centre

Helsinki University of Technology

P.O. Box 9800, FIN-02015 HUT, Finland

patrik.hoyer@hut.fi

Abstract

An important approach in visual neuroscience considers how the processing of the early visual system is dependent on the statistics of the natural environment. A particularly influential model in this respect has been sparse coding. In this paper we argue for a non-negative variant of the model. This is based partly on neurophysiological grounds and partly on the intuitive understanding of parts-based representations. We discuss the logic behind our reasoning and show experiments on natural images demonstrating the usefulness of the new model.

Key words: natural image statistics, sparse coding, non-negative representations

1 Introduction

A fundamental problem in visual neuroscience concerns understanding how the statistics of the environment is reflected in the processing of the early visual system. Research in this area originated almost half a century ago [1] and has

recently attracted significant attention as the computational power and the theoretical tools needed to investigate these issues have become available. For a review of the field, see [14].

A large share of this research has concerned the mammalian primary visual cortex, also known as V1. This is largely due to the fact that, ever since the pioneering work of Hubel and Wiesel [6], much is known about the processing performed there. Neurons in V1 can be divided into simple- and complex-cells. Of these, the responses of simple-cells can be reasonably described as a linear filtering of the visual input. The spatial properties of these linear filters (called the receptive fields) can be characterized as being localized, oriented, and bandpass [6,3]. The question of why this particular type of filtering is performed has undergone considerable debate during the years.

In a highly influential paper, Olshausen and Field [11] showed that linear sparse coding (SC) of natural images yielded features qualitatively very similar to receptive fields of simple-cells in V1. Subsequently, the very closely related model of independent component analysis (ICA) [7] was shown to give similar results [2]. This framework gave the observed receptive fields an *information-theoretic explanation* which had been lacking in previous work, demonstrating how sensory coding is related to the statistics of the environment.

The basic idea in sparse coding is relatively simple: The observed multidimensional data (random vector) $\mathbf{x}$ is modeled as a linear superposition of some features (called basis vectors) $\mathbf{a}_i$, as in $\mathbf{x} \approx \sum_{i=1}^{n} \mathbf{a}_i s_i$. In the basic image data case, each input vector $\mathbf{x}$ consists of the pixel intensities of a small image patch, and the $\mathbf{a}_i$ likewise contain the pixel intensities of basis patches. The latent variables s_i that give the mixing proportions are stochastic and differ for each

observed input $\mathbf{x}$. The crucial assumption in the sparse coding framework is that these hidden variables exhibit sparseness. Sparseness is a property independent of scale (variance), and implies that the s_i have probability densities which are highly peaked at zero and have heavy tails. Essentially, the idea is that any single typical input vector $\mathbf{x}$ can be accurately described using only a few active (significantly non-zero) units s_i. However, all of the basis vectors $\mathbf{a}_i$ are needed to represent the data because the set of active units changes from input to input.

The model, as typically used, is completely symmetric around zero. The data is normalized to zero mean, and the sources s_i are assumed to have even-symmetric probability densities (i.e. $p(s_i) = f(|s_i|)$) with a high peak at zero and heavy tails. This also implies that the observed data is completely symmetric with respect to the origin. When this model is learned from natural image data, the learned basis patches have the principal properties of the spatial receptive fields of simple-cells in V1. The neural interpretation of the model is thus that simple-cells in V1 perform sparse coding on the visual input they receive, with the receptive fields closely related to the sparse coding basis vectors and the firing rates of the neurons representing the latent variables s_i.

Could the sparse coding model be used more generally to understand how cortical circuits represent their inputs? The fact that areas of the cerebral cortex analyzing quite different inputs nevertheless anatomically are relatively similar [13] suggests the existence of common computational strategies for cortical sensory coding. Could sparse coding perhaps explain further properties of the visual system, or possibly properties of other sensory modalities? We believe that to tackle such questions, the model must first be slightly modified,

as explained in the next section.

2 Non-negative sparse coding

There are at least two obvious ways in which the standard sparse coding image model is unrealistic as a model of V1 simple-cell behavior. Perhaps the most glaring discrepancy is the fact that in the model each unit s_i can, in addition to being effectively silent (close to zero), be either positively or negatively active. This basically means that every feature contributes to representing stimuli of opposing polarity; for example, the same unit that codes for a dark bar on a bright background also codes for a bright bar on a dark background. This is in clear contrast to the behavior of simple-cells in V1: these neurons tend to have quite low background firing rates and, as firing rates cannot go negative, thus can only represent one half of the output distribution of a signed feature s_i.

Another major difference between the model and neural reality is that the input data in the model is double-sided (signed), whereas V1 receives the visual data from the lateral geniculate nucleus (LGN) in the form of separated ON- and OFF-channels. Of course, as an abstract model of visual coding, the input data should indeed be (signed) image contrast. But if we on the other hand are interested in how V1 recodes its input signals, we must consider separate ON- and OFF-channel input.

Thus, if we would like to transform sparse coding from a relatively abstract model of image representation in V1 to a concrete model of simple-cell recoding of LGN inputs, the model must be changed. First, our input data should

consist of hypothetical activities of ON- and OFF-channels in response to natural image patches. Second, all coefficients s_i should be restricted to non-negative values. As both the sources s_i and the data $\mathbf{x}$ thus have non-negative values only, it is logical to assume the same of the features $\mathbf{a}_i$. This is due to the fact that if the features contained negative values, this would inevitably be true of the observed data $\mathbf{x}$ as well.

Although we so far considered non-negativity constraints with the objective of making the model better fit known neurophysiology, there are also other equally important arguments for non-negativity. In particular, it has been argued that non-negativity can be important for learning parts-based representations [8]: In the standard sparse coding model, the data is described as a combination of elementary features involving both additive and subtractive interactions. The fact that features can 'cancel each other out' using subtraction is contrary to the intuitive notion of combining parts to form a whole. Non-negativity ensures that elementary object features combine additively.

Representations involving only additive combinations of object parts seem especially attractive for modeling higher-level features of images: In [5] it was shown how contour-coding could be learned from the statistics of complex-cell-like responses to natural images. This was possible by representing the model complex-cell responses by a non-negative, sparse, code.

Arranging all observed data vectors $\mathbf{x}$ into the columns of a matrix $\mathbf{X}$, the corresponding unknown sources into the matrix $\mathbf{S}$, and the unknown features $\mathbf{a}_i$ into the columns of $\mathbf{A}$, the linear model is given by $\mathbf{X} \approx \mathbf{AS}$. The non-negativity constraints require that all elements of $\mathbf{A}$, and $\mathbf{S}$ are zero or positive. The same is assumed of the observed data $\mathbf{X}$.

We suggest to use the following objective function for *non-negative sparse coding* (NNSC) [4,5]:

$$C(\mathbf{A}, \mathbf{S}) = \frac{1}{2}\|\mathbf{X} - \mathbf{AS}\|^2 + \lambda \sum_{ij} S_{ij} \tag{1}$$

With λ equal to zero this objective reduces to the squared error version of non-negative matrix factorization (NMF) [12,8,9]. However, for positive λ this objective emphasizes decompositions where $\mathbf{S}$ is sparse.

Minimizing the objective with respect to both $\mathbf{A}$ and $\mathbf{S}$ without any constraints other than non-negativity yields the NMF solution. This is because scaling up $\mathbf{A}$ while correspondingly scaling down $\mathbf{S}$ eventually drives the effective value of λ to zero. Thus, some constraint is needed on the scales of $\mathbf{A}$ and $\mathbf{S}$. We suggest to fix the norm of each column of $\mathbf{A}$ to unity, i.e. $\|\mathbf{a}_i\| = 1, \ \forall i$. An efficient algorithm for minimizing the objective subject to the suggested constraints is given in [4].

3 Learning receptive fields from ON/OFF-channels

Here we show how features resembling simple-cell receptive fields can be learned from natural image data preprocessed into separate ON- and OFF-channels. We used a set of natural images kindly provided by Bruno Olshausen. These images can be found as part of the *sparsenet* software package [10]. The images had been preprocessed by a spatial filter that approximately whitened the data, for details see [11]. We sampled 12×12 -pixel patches from the images, and then separated positive and negative values into separate channels. Each image patch was thus represented by a $2 \times 12 \times 12 = 288$ -dimensional vector, each element of which mimics the activity of an ON- or OFF-center

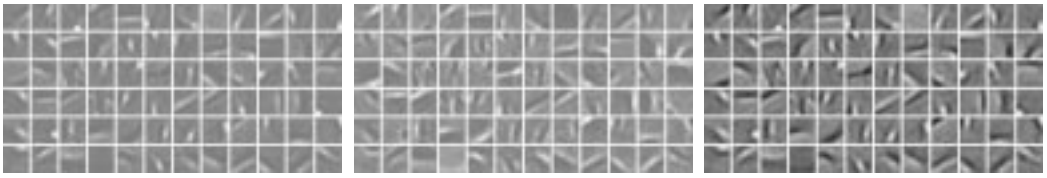

Fig. 1. Basis vectors from non-negative sparse coding of ON/OFF-channel image data. Left: Generative weights for the ON-channel. Each patch represents the part of one basis vector $\mathbf{a}_i$ corresponding to the ON-channel. Gray pixels denotes zero weight, brighter pixels are positive weights. Middle: Corresponding weights for the OFF-channel. Right: Weights for ON minus weights for OFF.

neuron to the input patch. These vectors made up the columns of $\mathbf{X}$.

We then minimized the objective given in Equation 1 under the stated constraints, with 72 sources, λ set to 5 and with each input channel scaled to have unit average squared activation. The complete code for performing the experiments can be found at `http://www.cis.hut.fi/phoyer/code/`. The resulting basis patterns (columns of $\mathbf{A}$) are shown in Figure 1. Note how the basis vectors represent Gabor-like image input by elongated ON- and OFF-subregions. This is most clearly seen in the rightmost panel, where the weights from the OFF-channel have been subtracted from those from the ON-channel. These learned features are at least qualitatively very similar to the ones found with the standard sparse coding model applied to symmetric image data.

To gain some insight into the roles of the sparseness objective and the non-negativity constraints in these results we can compare the learned decomposition to those given by other methods. For example, in Figure 2 we show the results of non-negative matrix factorization (NMF) on our data. This is actually a special case of our model, when $\lambda = 0$. Note that all features are unoriented; this held for all tested dimensionalities. Thus it seems that

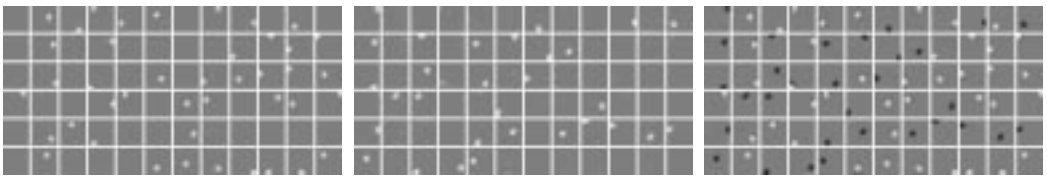

Fig. 2. Basis vectors learned by non-negative matrix factorization. Left: Weights for ON-channel. Middle: Weights for OFF-channel. Right: Weights for ON minus weights for OFF.

sparseness is important for learning oriented, localized features. The table below summarizes our findings on the behaviour of different decompositions on the data.

	NNSC	NMF	PCA	ICA/SC
localized	**yes**	**yes**	**no**	**yes**
oriented	**yes**	**no**	**yes**	**yes**
bandpass	**yes**	**no**	**yes**	**yes**
non-negative	**yes**	**yes**	**no**	**no**

4 Discussion and conclusions

Although we here showed only the learning of simple cell-like receptive fields, we believe that this modified sparse coding model can be used to explain higher-order visual receptive fields as well. For example, in [5] it was shown how contour coding and end-stopped receptive fields could be learned by non-negative sparse coding on the simulated responses of complex cells to natural images. It remains to be seen if the proposed model could also be useful in predicting or understanding properties of neurons analyzing other sensory

modalities.

Finally, we would like to point out one common misconception about models with non-negativity constraints: A frequent objection to non-negativity constraints is the fact that inhibition is widespread in the cortex. However, it is important to note that the constraints are on the activities and the *generative* representation. In fact, inhibition is a vital part of our model: it is required during *inference* as an important part of the competition between units to represent the stimuli as sparsely as possible.

Acknowledgements

I wish to thank Aapo Hyvärinen and Jarmo Hurri for useful discussions and helpful comments on an earlier version of the manuscript.

References

[1] H. B. Barlow. Possible principles underlying the transformations of sensory messages. In W. A. Rosenblith, editor, *Sensory Communication*, pages 217–234. MIT Press, 1961.

[2] A. J. Bell and T. J. Sejnowski. The 'independent components' of natural scenes are edge filters. *Vision Research*, 37:3327–3338, 1997.

[3] R. L. DeValois, D. G. Albrecht, and L. G. Thorell. Spatial frequency selectivity of cells in macaque visual cortex. *Vision Research*, 22:545–559, 1982.

[4] P. O. Hoyer. Non-negative sparse coding. Submitted. Available online, see `http://www.cis.hut.fi/phoyer/papers/`.

[5] P. O. Hoyer and A. Hyvärinen. A multi-layer sparse coding network learns contour coding from natural images. *Vision Research*. In press. Available online, see `http://www.cis.hut.fi/phoyer/papers/`.

[6] D. Hubel and T. Wiesel. Receptive fields and functional architecture of monkey striate cortex. *Journal of Physiology*, 195:215–243, 1968.

[7] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. Wiley Interscience, 2001.

[8] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[9] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing 13 (Proc. NIPS*2000)*. MIT Press, 2001.

[10] B. A. Olshausen. The Sparsenet software package for MATLAB. Available at `http://redwood.ucdavis.edu/bruno/sparsenet.html`.

[11] B. A. Olshausen and D. J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609, 1996.

[12] P. Paatero and U. Tapper. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5:111–126, 1994.

[13] A. J. Rockel, R. W. Hiorns, and T. P. S. Powell. The basic uniformity in structure of the neocortex. *Brain*, 103:221–244, 1980.

[14] E. P. Simoncelli and B. A. Olshausen. Natural image statistics and neural representation. *Annual Review of Neuroscience*, 24:1193–1215, 2001.