A two-layer ICA-like model estimated by Score Matching

Urs Köster* and Aapo Hyvärinen

University of Helsinki and Helsinki Institute for Information Technology

Abstract. Capturing regularities in high-dimensional data is an important problem in machine learning and signal processing. Here we present a statistical model that learns a nonlinear representation from the data that reflects abstract, invariant properties of the signal without making requirements about the kind of signal that can be processed. The model has a hierarchy of two layers, with the first layer broadly corresponding to Independent Component Analysis (ICA) and a second layer to represent higher order structure. We estimate the model using the mathematical framework of Score Matching (SM), a novel method for the estimation of non-normalized statistical models. The model incorporates a squaring nonlinearity, which we propose to be suitable for forming a higher-order code of invariances. Additionally the squaring can be viewed as modelling subspaces to capture residual dependencies, which linear models cannot capture.

1 Introduction

Unsupervised learning has the goal of discovering the underlying statistical structure of a stream of observed data. This is a difficult problem since most real world data has a complex structure which is hard to capture without prior knowledge. Typically, linear models like Independent Component Analysis (ICA) are utilized. Previous nonlinear extensions of ICA have incorporated prior knowledge on the data [1] [2], so they are not applicable to general data with unknown structure. Therefore we attempt to move towards more general models that can extract complex higher order structure rather than presupposing it. In addition, there is a strong incentive to develop algorithms for the efficient estimation of unsupervised statistical models since recent experiments show they can significantly improve the performance of supervised models [3].

Here we present a model that goes beyond the limitations of ICA without sacrificing generality. It has two layers of weights freely learned from the data, along with a nonlinearity forming a nonlinear representation of the input. The model is specified as a generalization of previous ICA-type models like Topographic ICA (TICA)[4] and Independent Subspace Analysis (ISA)[2]. Since both layers are learned from the data, no prior structure is imposed on the second layer.

* Urs Köster is supported by a scholarship from the *Alfried Krupp von Bohlen und Halbach*-foundation.

Learning in models like this can be done by maximizing the likelihood of the model distribution wrt. observed data. Here one often faces the problem that a model PDF (probability density function) cannot be normalized, and a straightforward estimation of the model is not possible. With Score Matching we present a novel approach to attack this problem. We recently showed [5] that a consistent estimation of the parameters maximizing the likelihood is possible without knowledge of the normalization constant. While other methods based on Monte Carlo methods or approximations have been successfully applied in the past, Score Matching has the advantage that it is a computationally efficient method guaranteeing statistical consistency.

The paper is organized as follows: In section 2, we present the two-layer probabilistic model in more detail, and we explain how it can be estimated using the Score Matching framework. In section 3 we first verify the estimation method by applying the model to artificial data with a known statistical structure. Following this, we present results on real-world data, image patches and natural sounds. The discussion, section 4, puts the new model in perspective with related methods. We highlight the important difference that our model gives rise to sparse connections in the second layer, which is not the case for related work on Contrastive Divergence [6] or modelling "Density Components" [7]. Finally in section 5 we conclude the paper with remarks about the scalability of the model and sketch some possible extensions to other types of data and more than two layers.

2 Model and Estimation

2.1 A Two-layer Model

While supervised learning methods have often used multiple representation layers, as in multi-layer Perceptrons trained with backpropagation, few unsupervised methods have used such a multi-layer representation. A major problem is that it is usually impossible to obtain the probability distribution of such a model in closed form. For this reason training such models often seems to require a lot of computational resources, because Markov Chain Monte Carlo or similar approximative methods have to be applied.

Still multi-layer models can provide a superior representation for a wide variety of data. We suggest that the lack of suitable estimation principle is a major reason for the poor performance of multilayer models in the past. Using the novel Score Matching approach we show that a very simple and general model can be demonstrated to perform well on a variety of tasks. We propose that our new approach provides a viable alternative to simpler models. Since we formulate it as a generalization of ICA, we find an intuitive way to interpret the results of the model in terms of generalized independent components.

The model that we present here is a bare-bones two layer network with two layers of weights and a scalar nonlinearity acting on the sum of the inputs to each unit.

$$y_i = \mathbf{V}_i g(\mathbf{W}\mathbf{x}) \tag{1}$$

The output of one top-level unit y_i is thus obtained from the data vector $\mathbf{x}$ given the weight matrix $\mathbf{W}$, the row of weights $\mathbf{V}_i$ as well as the nonlinearity $g(\mathbf{u})$. The size of $\mathbf{W}$ and $\mathbf{V}$ is chosen to be equal to the data dimensionality n for simplicity, but the estimation method we propose can also deal with overcompleteness in one or both layers. The weight matrix $\mathbf{V}$ is further constrained to have non-negative elements.

After the first layer of weights $\mathbf{W}$ has performed a linear transform of the data, the scalar nonlinearity $g(\mathbf{u})$ is applied to the outputs. This nonlinearity is the same for all units, and it is fixed in advance rather than learned from the data. We choose to focus on a squaring for the first nonlinearity, i.e. $g(\mathbf{u}) = \mathbf{u}^2$ where the nonlinearity is taken to be element-wise. The second layer $\mathbf{V}$ computes linear combinations of these squared outputs. There are several ways to interpret the squaring nonlinearity that we propose here. Firstly, we would like to point out the connection to our work on Independent Subspace Analysis (ISA) [2], where the components inside a subspace are squared to compute the L_2-norm of the projection onto a subspace. This provides a way to model dependencies of squares that cannot be removed by a simple linear transform. Modelling these dependencies explicitly allows a better fit to the data than linear models could achieve, since high correlations exist between the *activity* of similar features even if they are linearity uncorrelated. [8] The second way to describe the model is to in terms of invariant features. This can provide high selectivity to certain aspects of the data while ignoring aspects that are not relevant to describe the statistical structure of the input. From this point of view the outputs would be features highly invariant under a specific kind of transformation on the input data. A sum of squares, an operation that preserves amplitude but discards the phase of a signal, could perform such an invariant feature extraction. [9]

Finally there is an output nonlinearity acting on the second layer outputs. It has the purpose of shaping the overall model PDF to match the statistics of the data. In principle, this could be matched to the optimal distribution for the data under consideration e.g. by an iterative optimization. For simplicity however, we assume the data can be modeled in terms of sparse sources, so we choose an element-wise square root nonlinearity of the form $h(\mathbf{u}) = -\sqrt{\mathbf{u} + 1}$. Such a convex choice of h is related to supergaussianity of the PDF.

For learning, the outputs of the second nonlinearity are summed together to define a probability distribution q over the input data.

$$\log q(\mathbf{x}|W, V) = \sum_{i=1}^{n} h\left(\mathbf{V}_i g(\mathbf{W}\mathbf{x})\right) \tag{2}$$

Intuitively, this model can be thought of as a two layer neural network processing the incoming data vector and computing the probability that the data came from the distribution defined by the model. This immediately provides a means of training the model by adjusting the parameters to maximize the likelihood of the model given the observed training data.

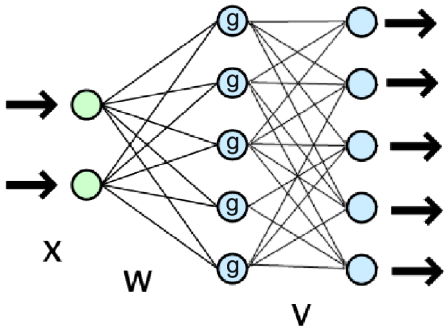


Fig. 1. Graphical representation of the two-layer model

For estimation, we usually need to compute the log-likelihood of the model

$$\log l(\mathbf{W}, \mathbf{V}|\mathbf{x}) = \sum_{i=1}^{n} h\left(\mathbf{V}_i g(\mathbf{W}\mathbf{x})\right) - \log\left(Z(W, V)\right) \tag{3}$$

where Z denotes the normalization constant of the distribution, which is obtained by integrating over all space. It is obvious that the normalization constant cannot be computed in closed form, which makes the estimation of the model impossible with standard methods. Therefore we apply the novel estimation method Score Matching which is described below.

2.2 Score Matching

As we stated above, the probability distribution of the data can in general only be obtained up to a multiplicative constant. This makes it impossible to compute the likelihood of the model, and standard optimization methods like gradient descent on the log-likelihood cannot be used. In the past, Monte Carlo methods such as Contrastive Divergence [10] have been applied to this problem, or approximations of the likelihood were used. Here we circumvent the problem by focusing on the *score function of the density*, $\boldsymbol{\Psi}(\boldsymbol{\eta}; \mathbf{W}, \mathbf{V})$ with respect to $\boldsymbol{\eta}$, where $\boldsymbol{\eta}$ is a variable which replaces the data vector $\mathbf{x}$ for notational unambiguity.

$$\boldsymbol{\Psi}(\boldsymbol{\eta}; \mathbf{W}, \mathbf{V}) = \nabla_{\boldsymbol{\eta}} \log p(\boldsymbol{\eta}; \mathbf{W}, \mathbf{V}) \tag{4}$$

Additionally we can define the *data score function* $\boldsymbol{\Psi}_x(.) = \nabla_\eta \log p_x(.)$ for the distribution of observed data. The model is optimized by matching the data and model score functions (hence the name Score Matching). We can achieve this by minimizing the squared distance

$$J(\mathbf{W}, \mathbf{V}) = \frac{1}{2} \int_{\boldsymbol{\eta}} \|\boldsymbol{\Psi}(\boldsymbol{\eta}; \mathbf{W}, \mathbf{V}) - \boldsymbol{\Psi}_{\mathbf{x}}(\boldsymbol{\eta})\|^2 d\boldsymbol{\eta} \tag{5}$$

This could painstakingly be computed using a nonparametric estimation of the density, but as shown in [5] the expression can be expressed in a much simpler form in terms of derivatives of the data score function:

$$\tilde{J}(\mathbf{W},\mathbf{V}) = \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{n}\left[\frac{\partial}{\partial\eta_i}\boldsymbol{\Psi}_i(\mathbf{x}(t);\mathbf{W},\mathbf{V}) + \frac{1}{2}\boldsymbol{\Psi}_i^2(\mathbf{x}(t);\mathbf{W},\mathbf{V})\right] + C \qquad (6)$$

Here the $\tilde{J}$ indicates a sampled version of the objective function, but in the limit of $T \to \infty$ and given the existence of a nondegenerate optimum, this estimator is statistically consistent. C is a constant that does not depend on any the parameters. Estimation of the parameters can easily be performed by following the gradient of this function wrt. the parameters.

3 Experiments

3.1 Methods

We performed experiments on a variety of data to show the power and adaptability of the model. The focus was on natural data, i.e. natural image patches and speech recordings, to demonstrate the particular suitability of our model to this very complex and rich kind of data that is poorly modeled by simpler methods. For the natural data we performed preprocessing in the form of whitening (decorrelation), Contrast Gain Control by dividing each data vector by its L_2-norm, and some dimensionality reduction by PCA.

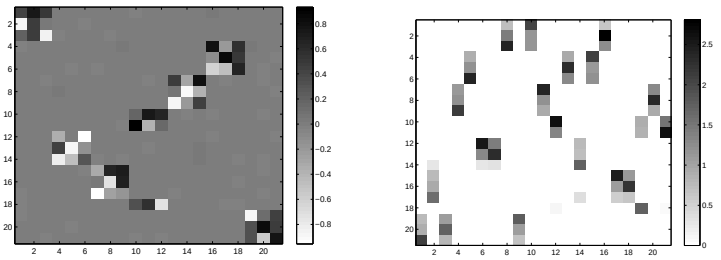
In general we start the optimization by learning the independent components of the data, which is achieved by clamping the second layer weights to the identity matrix. This serves to avoid local minima and speed up the convergence of the algorithm. After this, the second layer connections are learned. It is an important feature of the estimation method that learning for the first layer is not stopped; rather the first layer features start to move away from ICA features to adjust to the second layer as it forms more complex and invariant features.

An additional technical constraint was the use of L_2-normalization on the rows of $\mathbf{V}$, corresponding to the second layer output vectors. This prevents individual units from "dying" and also sets a bound on the maximum activity. We verified that it does not qualitatively change the structure of the outputs. $\mathbf{W}$ was constrained to be orthogonal as it is customary with ICA algorithms. For the optimization we used a stochastic gradient approach with mini batches consisting of 100 data samples. Not only does this significantly increase the speed of convergence, but we found that without stochasticity, local minima hindered the convergence of the second layer weights.

3.2 Artificial data

As a first test for the model and estimation method we generated data according to the ISA model[2]. This is supergaussian data with dependencies *within*, but

not *between* subspaces of the data variables. This data was then mixed with a random mixing matrix $\mathbf{A}$. We used 10,000 samples of 21-dimensional data generated with a subspace size of three.



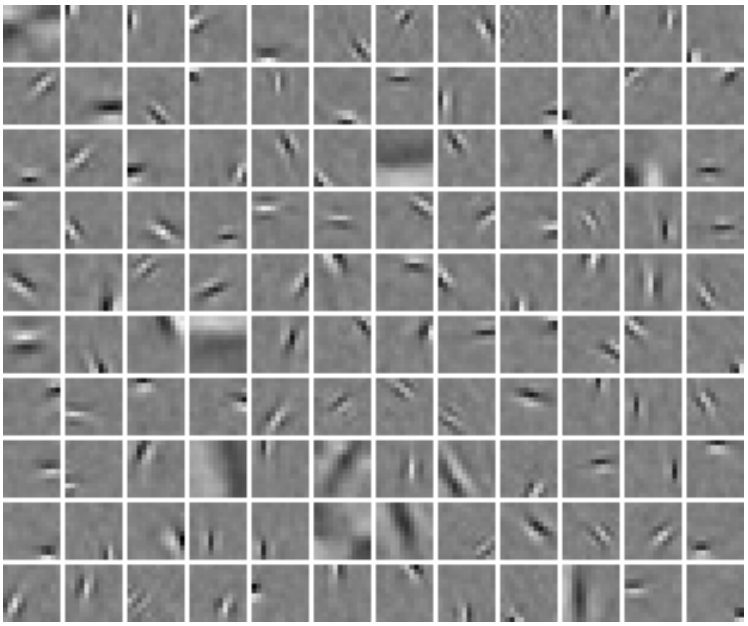
(a) (b)

Fig. 2. The model was tested with ISA data, convergence is fast and finds the global minimum. We show (a) the product of the estimated demixing and known mixing matrix, (b) the learned second layer weights. The rows of the matrices are sorted in ascending order on the columns of $\mathbf{V}$. This does not affect the result and is purely for easier visualization.

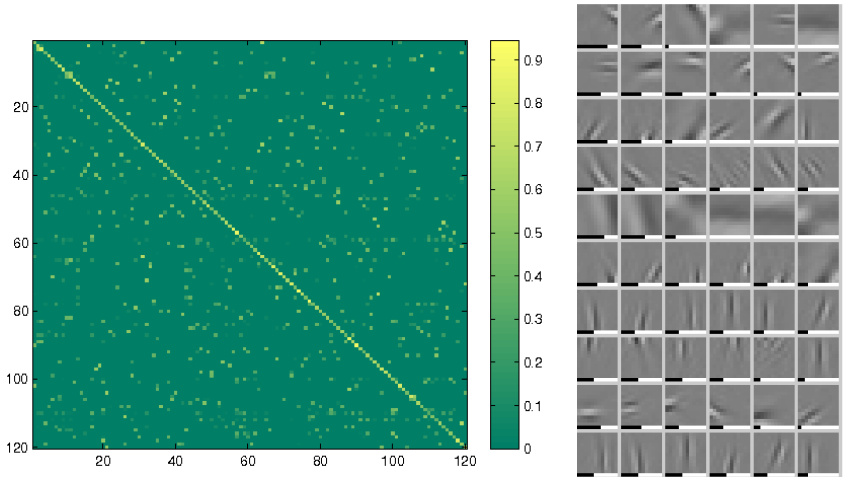
Figure 2 shows how the first layer weights $\mathbf{W}$ invert the mixing up to subspace membership, while $\mathbf{V}$ determines which variables belong together in one subspace. Since the dimensionality of $\mathbf{V}$ is 21×21, and there are only 7 subspaces, some rows of $\mathbf{V}$ go to zero and some are duplicated. Contrary to later experiments, both weight layers were initialized randomly and learned simultaneously, and no normalization on the rows $\mathbf{V}$ was performed.

3.3 Experiments on Natural Images

After confirming the identifiability of the method, we tested the model on natural images which have a particularly rich statistical structure with many higher order dependencies. We use 20,000 image patches of 12×12 pixels, whitened, performed Contrast Gain Control [11] and reduced the data dimensionality to 120 by PCA. We specified the dimensionality of both $\mathbf{W}$ and $\mathbf{V}$ to be 120×120. Optimizing $\mathbf{W}$ first gives familiar ICA features as shown in fig. 3a. In fact variants such as TICA and ISA can easily be performed by setting $\mathbf{V}$ appropriately. The second layer learns connections between similar first layer features (fig. 3b), giving rise to complex-cell like outputs which are invariant to the spacial phase of the data (fig. 3c). Continued learning on the first layer features increased the similarity of the position and size of filter feeding into the same second layer unit while keeping the phase difference. This result was also confirmed with an overcomplete model.



(a) First Layer

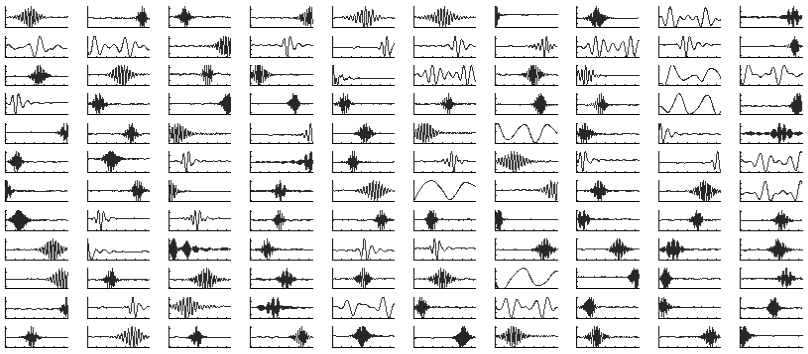


(b) Second Layer (c) Some Outputs

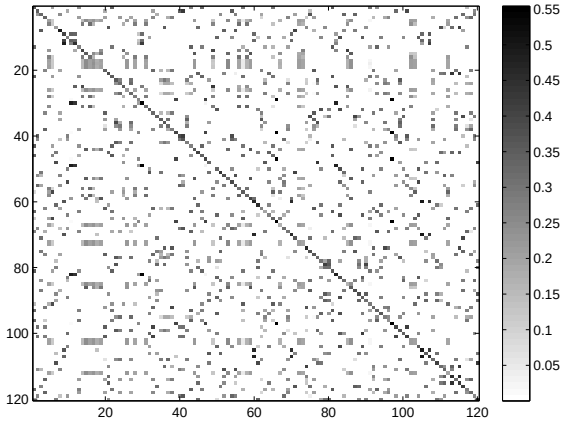
Fig. 3. a) First layer filters show the classical Simple-Cell type structure. b) Connection in the second layer are sparse, with connections between similar units. c) A random selection of outputs where each row shows the most active contributers to the response with the black bars indicating "synaptic strength", i.e. how strongly the filter contributes to the output.

3.4 Audio data

In order to demonstrate the general applicability of our model to a variety of data, we also tested it on speech data from the TIMIT database. We sampled random rectangular sound windows of 8ms length, and resampled them to 8kHz. We also applied our standard preprocessing consisting of removing the DC component, whitening and contrast gain control. Simultaneously we reduced the dimensionality from 64 to 60 which amounts to low-pass filtering and serves to eliminate artifacts from the windowing procedure. The results are presented in figure 4.



(a) First Layer



(b) Second Layer

Fig. 4. (a) The first layer gives outputs localized in both frequency and time. (b) The second layer gives connections between features with dependencies of squares.

4 Discussion

We have shown that an unsupervised model using two completely flexible layers of weights to learn the statistical structure of its input data can effectively be estimated using Score Matching. While including previous extensions of ICA as special cases, this is far more general that previous models. For example ISA forces the filters to group into subspaces of a constant size and with an equal contribution, and did not allow a single filter to be active in more than one higher order unit. These constraints have been lifted with the new model. Topographic ICA is also included in our model as a special case. If the second layer is fixed to an identity matrix convolved with a kernel (neighborhood function) that leaks activity to off-diagonal elements, a topographic ICA model can be estimated. A more complex topography can be obtained by allowing interactions other than along the main diagonal.

Two models have recently been proposed that have a similar hierarchical structure but are estimated differently. Most close related to our work is the work by Osindero et al. [6]. Instead of using the traditional "independent component" point of view, the model is defined as a "product of experts" model following Student-t distributions. The estimation is performed using contrastive divergence (CD), which was recently shown [12] to be equivalent to Score Matching. The key difference between the models is in the results obtained on natural data. While Osindero et al. report sparse *activation* of second layer units, we also see sparse *connectivity*, which has interesting implications not only because of the striking similarity to biological networks, but also for efficient signal processing.

The second work that we would like to mention is that of Karklin and Lewicki [7]. They present a generative two layer model that performs ICA on the data followed by a variance-modelling stage as in TICA[4]. Contrary to the PoT model of Osindero et al. and our SM model, both layers are estimated separately using the *maximum a posteriori* estimate. The authors observe that in contrast to our model, the first layer units do not change significantly depending on the "density components" modelling the variance of the first layer outputs. Applied to natural stimulus data, this model gives rise to broadly tuned features in the second layer that describe global properties of the data. Again this is in contrast to the sparse connectivity obtained from our model.

5 Conclusion

We have presented a two layer model that that can be used to learn the statistical structure of various kinds of data. By using the novel estimation principle Score Matching, unsupervised learning in this type of model is made faster and more straightforward than with alternatives such as Monte Carlo methods. Contrary to previous linear models, higher order dependencies in the data can be captured to give better models of real world data. Compared to similar models [6] [7], we report the emergence of sparse connectivity in the second layer. Furthermore our model is very general, so it can be overcomplete, and it can be extended to incorporate a third or more layers.

References

1. Jean-Franois Cardoso. Multidimensional independent component analysis. In *Proc. ICASSP'98. Seattle.*, 1998.
2. A. Hyvärinen and P. Hoyer. Emergence of phase and shift invariant features by decomposition of natural images into independent feature subspaces. *Neural Computation, 12(7):1705-1720.*, 2000.
3. G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, July 2006.
4. A. Hyvärinen, P. Hoyer, and M. Inki. Topographic independent component analysis. *Neural Computation.*, 2001.
5. A. Hyvärinen. Estimation of non-normalized statistical models using score matching. *Journal of Machine Learning Research*, 6:695–709, 2005.
6. S. Osindero, M. Welling, and G. E. Hinton. Topographic product models applied to natural scene statistics. *Neural Computation*, 18, 2006.
7. Y. Karklin and M. S. Lewicki. A hierarchical bayesian model for learning nonlinear statistical regularities in non-stationary natural signals. *Neural Computation*, 17(2):397–423, 2005.
8. E. Simoncelli and E. Adelson. Noise removal via bayesian wavelet coding. *Intl Conf. on Image Processing. 379-382*, 1996.
9. A. Hyvärinen, J. Hurri, and J. Väyrynen. Bubbles: a unifying framework for low-level statistical properties of natural image sequences. *Journal of the Optical Society of America A, 20(7):1237-1252*, 2003.
10. Geoffrey E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Comput.*, 14(8):1771–1800, 2002.
11. U. Köster A. Hyvärinen. Complex cell pooling and the statistics of natural images. *Network: Computation in Neural Systems, in print*, available online: cs.helsinki.fi/u/koster/koster05.pdf, 2005.
12. A. Hyvärinen. Connections between score matching, contrastive divergence, and pseudolikelihood for continuous-valued variables. *IEEE Transactions on Neural Networks*, in press.