EXAMINING THE DEPENDENCIES BETWEEN ICA FEATURES OF IMAGE DATA

Mika Inki

Neural Networks Research Centre
Helsinki University of Technology
P.O. Box 5400, FIN-02015 HUT, Finland

ABSTRACT

In this paper we study the dependencies of features found by independent component analysis (ICA) in image data by examining how the activation of one feature changes certain statistics of the data. We look at how the PCA components are affected when we know a certain ICA feature is highly active, and also study the ICA components in this situation. This can be thought of as a simple form of two-layer ICA. We show that the activation level of features with similar properties is elevated, and the activation level of features with distinctly different properties is decreased.

1. INTRODUCTION

Independent component analysis can be considered to be a generative model for low-level features of many types of natural data, e.g. natural image data. In ICA the observed data is expressed as a linear transformation of latent variables that are nongaussian and mutually independent. We may express the model as

$$\mathbf{x} = \mathbf{A}\mathbf{s} = \sum_i \mathbf{a}_i s_i, \tag{1}$$

where $\mathbf{x} = (x_1, x_2, \ldots, x_m)$ is the vector of observed random variables, $\mathbf{s} = (s_1, s_2, \ldots, s_n)$ is the vector of the latent variables called the independent components or source signals, and $\mathbf{A}$ is an unknown constant matrix, called the mixing matrix. The columns of $\mathbf{A}$ are often called features or basis vectors. Exact conditions for the identifiability of the model were given in [1], and several methods for estimation of the classic ICA model have been proposed in the literature, see [3] for a review. We consider here a simple extension of ICA, which can also be thought of as an overcomplete case where $n > m$, but we only need algorithms developed for the complete case $n = m$.

The assumption of independence is fundamental in ICA. It is, however, intuitively clear that for example image data does not have such independent (linear) features that ICA attempts to find. It is important to know about the data structures not captured by the ICA model. We look here at what can be thought of as a second layer of ICA. We look at the ICA components of the data subset where we know that a certain ICA feature of the whole data set is highly active. So, we normally want to find the most independent components, i.e. a $\mathbf{W}$ such that the likelihood of the following model is maximal:

$$P_{\mathbf{x}}(\mathbf{x}) = P_{\mathbf{s}}(\mathbf{W}^T\mathbf{x})\det(\mathbf{W}) \approx (\prod_i P_{s_i}(\mathbf{w}_i^T\mathbf{x}))\det(\mathbf{W}). \tag{2}$$

Here the basis vectors $\mathbf{a}_i$ can be obtained from the separation vectors $\mathbf{w}_i$ by multiplying with the covariance matrix. But now we are interested in the best fitting model when the activation of component l is high. One may think that in these cases feature l is a good description for some object or higher level feature in the patches. Now the problem is to find $\mathbf{W}^l$ such that the likelihood of the following model is maximized:

$$P_{\mathbf{x}}(\mathbf{x} \mid \|\mathbf{w}_l\mathbf{x}\| > \alpha) \approx (\prod_i P_{s_i^l}((\mathbf{w}_i^l)^T\mathbf{x} \mid \|\mathbf{w}_l\mathbf{x}\| > \alpha))\det(\mathbf{W}^l). \tag{3}$$

If the ICA model were to hold, these new features and the corresponding independent components would not be any different from those of regular ICA, as the activation of one component would tell nothing of the activation of the others. Therefore all the apparent additional structure would have to be due to random variation. However, we will show that the activity levels of the features change both qualitatively and quantitatively in a way that is not expected if the ICA model is assumed. We would surmise that the features themselves do not change much when one feature is active, essentially only their activity levels change.

It is possible to think of think of this two-level ICA description as an overcomplete basis for the data. But

the model it suggests is one of concurrent activity of signals and not a simple overcomplete model, something akin to Topographic ICA [2], but we don't formulate an explicit generative model here. Somewhat similar models for the dependencies between the variances of wavelet features have also been presented in the literature, see [6].

The basic ICA model can be used with varying degrees of success in image processing and analysis, and with good success in modeling simple cell receptive fields in the primary visual cortex (V1) [5, 2]. However, in order to be truly successful in image processing or modeling higher brain areas, one has to build more complicated models and obtain better understanding of higher order features of image data. The study in this paper can be understood as a small step in that direction. It should be noted that we still do not impose any restrictions on the shapes of the features.

2. PREPROCESSING

As data we use 13 images of landscapes, plants and animals. These images can be found at the web address **http://www.cis.hut.fi/projects/ica/data/images/**. We sample 200000 12×12 patches of the data, and from each patch we subtract the mean.

Data normalization is of utmost importance here. As the activation of one feature in an image patch correlates strongly with the activation of all the other features, i.e. the energies of the features are positively correlated [2], it is important that we normalize the patch variances in order to emphasize the more interesting dependencies. This energy correlation is partly due to the fact that in images there are 'empty' areas (which Mumford refers to as the 'blue sky' axiom [4]) where no or only a few components are active (have large values) but also 'complex' areas with numerous active components. Another reason is that the contrast variations can be large even inside images. Also, as we attempt to find out how the activation of one feature changes the activity level of the other features, it is important that the patches are whitened. It is then simple to find those directions whose activation level changes as this corresponds to variance deviating from unity.

We first normalize the patch variances, which essentially projects all the patches on the surface of an m-dimensional hypersphere. Note that of the image patches, just over half a percent consist of a single intensity value, which means that they are originally in the origin. These patches are left in the origin.

Next we whiten the data. Whitening the data changes the variances of the patches, and therefore we have to again normalize the patch variances. After the initial patch variance normalization, we whiten the data and normalize the patch variances (alternately) a total of five times. Before the first whitening, the largest eigenvalue of the covariance matrix is more than 1400 times the smallest eigenvalue. Before the second whitening this difference is 4.6, before the third 1.58, before the fourth 1.13, before the fifth 1.034, and after the process is finished 1.0088, i.e. the largest eigenvalue is 0.88% larger than smallest. Note that the whole process can be described as multiplying the whole data by a matrix and each patch with its own scalar.

3. RESULTS

We then use FastICA [3] to perform ICA on the data. The basis we obtained thus is in figure 1. A good portion of the features appear to be rather small line detectors with only a few active pixels. Data normalization may may be responsible for the emergence of these features, a lot less of them appear when patch variances aren't normalized. If a patch essentially contains a single line on a 'flat' background, this line is much emphasized by the patch variance normalization. Also quantization used in converting the images into digital format may affect some of these features. As mentioned earlier, half a percent of the patches consisted of only one intensity value, and additionally 1.6 percent of the patches have between 2 and 10 intensity levels inside them (on average, the patches have about 60 intensity levels). This means that there are clear (aliased) boundaries between intensity value areas, which may promote 'jumps' in the intensity values between neighboring pixels in the features.

For each of these ICA features we select the patches for whom the activation (the absolute value of the IC) exceeds a certain threshold. The higher a threshold we use, the better we observe the differences to the normal case (zero threshold), but at the same time, the less patches we have. We selected a value of two, i.e. twice the standard deviation of the feature(s). This left between 8200 and 11700 samples, i.e. about 5% of the data. If the ICA model were to hold, the expected activity level of the other components would not change. As we have normalized the patch variances (thus the ICA model does not hold by construction), the activity level of all the other features could be expected to fall slightly. However, for a large number of components (features), the average activity level actually rises.

There is, of course, random variation in the activity levels due to a finite sample size. In figure 2 we have plotted the eigenvalues of the covariance matrix one could expect if the ICA model were to hold, and the actual observed eigenvalues. To estimate the values for

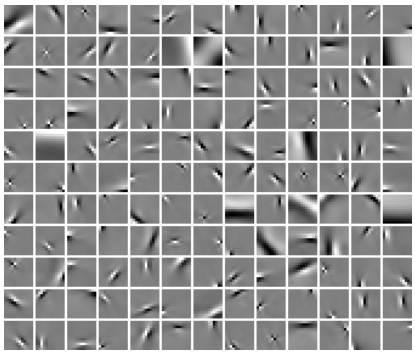

Figure 1: The ICA basis obtained with normalized data.

the pure ICA model, we kept the values for the ICA feature used in the selection, and randomly selected new values for the other components from their values in the original 200000 sample data set, and normalized patch variances. This provides some indication of what kind of eigenvalues one could expect for the covariance matrix. As one can see, the two curves differ noticeably.

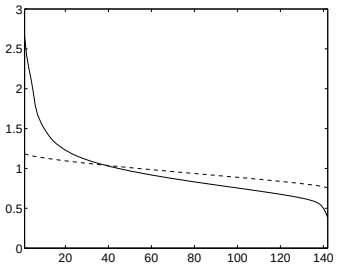

Figure 2: The average variances in PCA directions, when omitting the ICA direction used to pick the samples (the average variance in which was 7.81). Solid line: observed from data. Dashed line: what is expected when the ICA model holds.

3.1. Higher-Order Features

If we assume that that the interesting information is contained in the dimensions with the largest variance, two obvious ways of examining the data are available. We can examine the largest principal components, or perform ICA on the data, and order the components by the variance in the corresponding direction.

In figure 3 we have the PCA basis vectors corresponding to the data subset where the fourth IC feature on the second row of figure 1 is highly active. Note that as the data has been whitened before patch selection, these PCA components are completely produced by the selection process. The largest principal component is on the top left corner of the image and it corresponds to the IC used to select the data. The other PCs are in a descending order from left to right and from top to bottom. As one can see, the largest components have approximately the same size and location as the IC but start differing in other properties. The components in the middle of the basis apparently arise more due to random variation and not dependencies with the original IC. Lastly, the smallest PCA components appear to have sizes and orientations as different as possible from the original IC.

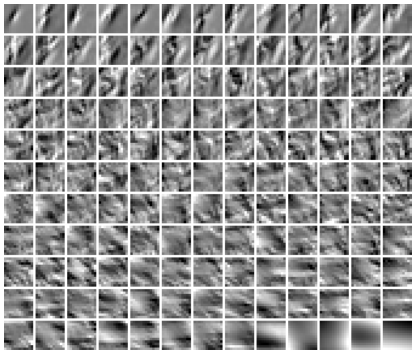

Figure 3: The PCA basis vectors for the data subset where the fourth ICA feature on the second row in figure 1 is active. The PCA components are ordered by (decreasing) variance from left to right and from top to bottom.

In figure 4 we have the ICA features found in the data subset where the fourth feature on the second row in figure 1 is highly active. We removed the component in the direction of the original IC as the distribution in this direction is by construction quite pathological. Again the largest ICs have positions, orientations and sizes quite close to the IC used to select the data. In the middle of the basis there are features that are not really related to the original IC and at the end are features with properties least like the original IC, just as was the case with PCA.

All this is quite natural as the activation of one IC tells that there is something (object, edge etc.) in that window that is quite well described by that feature, but as the IC cannot be expected to fully describe that object, the activation increases also in other similar features. The likelihood for the presence (in terms of contribution to variance) of completely different objects at the same time decreases especially as the patch variances are normalized.

In figure 5 we have the original ICA components or-

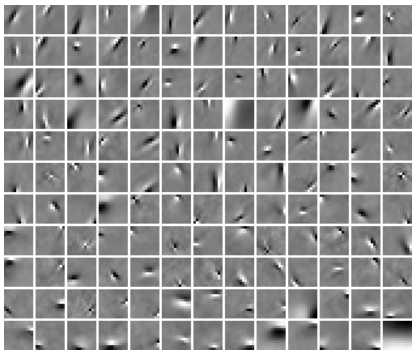

Figure 4: The ICA features for the data subset where the first feature on the top row in this image is highly active. The new ICA features are then ordered by decreasing variance from left to right and from top to bottom.

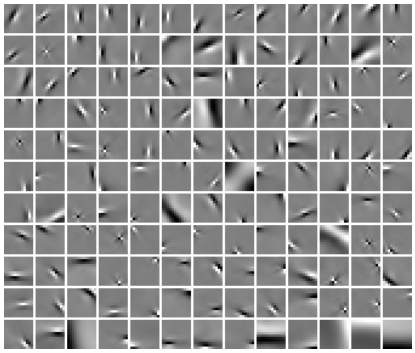

Figure 5: The original ICA features ordered by decreasing variance in the data subset where the fourth ICA feature on the second row in figure 1 is highly active.

dered by decreasing variance in the data subset where the feature used in patch selection is highly active. There is not a large qualitative difference between the figures, although it would appear that there are more features in figure 4 that are visually close to the original feature. However, as there is essentially an infinite amount of these ICs in image data, we surmise that the ICs do not change with the activation of a feature, only their prevalence (the subset found) changes slightly.

Finally, in figure 6 we have the eight largest ICA components for each of the data subsets where one of the original ICA components is active. For reasons of space, only the first quarter of the results are presented here. Again, ICA features with similar properties to the one used in the selection increase in activity.

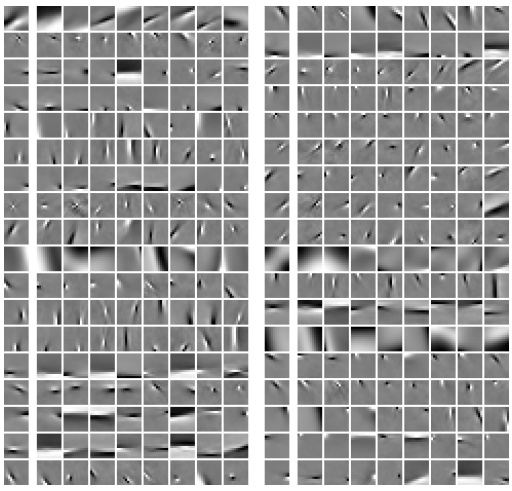

Figure 6: The eight largest ICA components obtained for data subsets where one (on the left of each group) of the ICA features in figure 1 is highly active. Only the first quarter of the results are shown.

4. CONCLUSIONS

Here we studied the dependencies between ICA components in image data by looking at what effect the activation of one feature has for the presence of other features in the data. We found that features with similar properties (position, orientation, size) increase in activity. We surmise that the features themselves do not change much and mostly only their activity level changes. We also believe that the procedures and results presented here facilitate further analysis and may prove to be useful in image processing.

5. REFERENCES

[1] P. Comon. Independent component analysis—a new concept? *Signal Processing*, 36:287–314, 1994.

[2] A. Hyvärinen, P. O. Hoyer, and M. Inki. Topographic independent component analysis. *Neural Computation*, 13(7), 2001. in press.

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

[4] D. Mumford and B. Gidas. Stochastic models for generic images. *Quarterly of Applied Mathematics*, LIX(1):85–111, 2001.

[5] B. A. Olshausen and D. J. Field. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, 37:3311–3325, 1997.

[6] E. P. Simoncelli, W. T. Freeman, E. H. Adelson, and D. J Heeger. Image compression via joint statistical characterization in the wavelet domain. *IEEE Transactions on Image Processing*, 8(12):1688–1701, 1999.