Natural Image Patch Statistics Conditioned on Activity of an Independent Component

Mika Inki INKI@CIS.HUT.FI
Neural Networks Research Centre
Helsinki University of Technology
P.O. Box 5400, FI-02015 HUT, Finland

Abstract

In this paper we offer a relatively comprehensive look at how the activation of one independent component analysis (ICA) feature changes the first and second order statistics in orthogonal directions in whitened image patches. Essential here is normalizing image patch lengths and normalizing the patches by the sign of the active (conditioning) component. First order statistics for high activation are shown to extend the original features even outside the original windows. Changes in second order statistics can be argued to be linked to the 'errors' made in describing the actual object in the image patches by the active feature.

1. INTRODUCTION

Independent component analysis has been used successfully in analyzing image data, even though the model is fundamentally insufficient for describing images. In ICA the observed data is expressed as a linear transformation of latent variables that are nongaussian and mutually independent. We can 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 latent variables called the independent components (ICs) 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 Comon (1994), and several methods for estimation of the classic ICA model have been proposed in the literature, see Hyvärinen et al. (2001b) for a review.

The assumption of independence is fundamental in ICA. Most types of natural data (and natural image data in particular) do not, however, have such independent (linear) features that ICA attempts to find. However, the components found by ICA are maximally independent in the mutual information sense, which means their marginal distributions are maximally nongaussian, or sparse with image data. This maximal independence (sparseness) means that the ICA basis is the best basis for component-wise compression. It is important and interesting to know about the data structures not captured by the ICA model, i.e. dependencies between the components. An adequate description of these dependencies can be useful when building a nonlinear (or multi-layer) model on top of the ICA model for image data, or when simply searching for an efficient coding scheme for the data.

The usefulness of this approach can be motivated by the link between ICA features and cortical simple cell receptive fields, see Olshausen and Field (1996), Stork and Wilson (1990). This supports the hypothesis that the primary goal of early sensory (visual) processing is to find an efficient representation for the sensory input Barlow (1961).

There have been efforts at extending the ICA model for image data to cases where the sources (independent components) are, in fact, not independent, see Hyvärinen et al. (2001a), Hyvärinen and Hoyer (2000), Karklin and Lewicki (2003). This is basically what we will be concentrating on here, but we will not start by heuristically building a model and then fitting it, we will be more concentrated on examining the dependencies. Here we are interested in knowing what we can infer about the content of a patch from the knowledge that a certain IC contributes a certain portion of it. Note, that if the data had been generated according to the ICA model, knowledge of the value of one independent component would not reveal anything about other independent components.

Another reason why the ICA model is insufficient for describing images is that ICA components are usually restricted to be orthogonal to each other in the whitened space, which means that the dimensionality of the sources does not exceed the dimensionality of the data (that is, $n \leq m$). However, no such limitation naturally exists in image data. There have been several attempts to build algorithms for the case where the number of sources is higher than data dimensionality, see Hyvärinen and Inki (2002), Lee et al. (1999), Lewicki and Sejnowski (2000). The precise way in which the components are constrained to be in the space can also affect the shapes of the features, or the distribution of their sizes. However, we will not address this problem here. After finding the independent components, we analyze the dependencies by taking one or two components at a time, so the distribution of the features (in the feature space) mostly only matters w.r.t. the effect it has on the shapes of the individual features.

A third third reason for the insufficiency of the ICA model is that, for computational reasons, the data is usually examined only in small image patches and the features are often 'cut' by the edges of the patches. It is important to know how the results for such small patches can be extended or generalized into larger areas. We will revisit this problem later in this work.

Of course, there are other ways of extending the results of using ICA for images. One can look at stereo or color images (Hoyer and Hyvärinen, 2000), at image sequences (van Hateren and Ruderman, 1998), or at three dimensional images (Inki, 2003b). But in these cases, the extension is in the data, and not in the model. The same questions about dependencies have to be answered also in these cases.

In Figure 1 we have an example of the scenario we are interested in. On the left side we have an ICA feature, and on the right side 20 patches, where the feature contributes more than one sixth of the variance of the patch (in the whitened space). These patches are visualized with their surrounding areas.

As one can see, in all of the patches, the object that activates the feature is quite prominent, yet the feature does not describe the object completely in any of the patches. In most patches the edge that the feature corresponds to extends significantly further, even outside of the patch, so other features are needed to describe the rest of the edge. In some patches the edge is not perfectly aligned with the feature, so other features are needed to describe the 'error'. In a few patches, the edge starts to bend, while in others it is partly

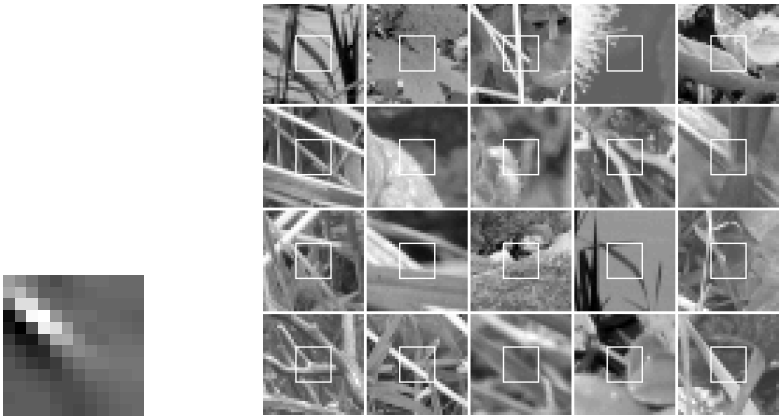

Figure 1: A typical ICA feature on the left, and framed on the right are locations where it is highly active. For visual clarity, the patches in this figure (and in all further figures) have been normalized so that the mean value is zero (corresponding to mid-gray), and either the darkest (smallest) value in the patch is black or the brightest (biggest) is white.

occluded by another edge. In several patches the edge is accompanied by an other edge, so that the object has the shape of a line of varying width. Note also that the object seldom 'fades out' even though the feature does.

There are also objects in the patches that are apparently (nearly) independent of the existence of the edge. However, with the knowledge about the activity of a feature, one can actually infer many things about other features in the patch. Some of these structures are nearly always present, some only sometimes, others nearly never, but the probability of their presence is often quite different compared to the case when nothing is known of the patch. We will study the qualities and sizes of these dependencies in this paper. These results can suggest useful models for image data, for example for coding and noise removal purposes, possibly even when the bases used are not exactly ICA bases, but have some similarities as with certain types of wavelet bases.

This paper is organized as follows. In Section 2 we discuss how we collected the image data used here. In Section 3 we discuss what statistical properties of image patches we will examine, and how to process the data to make these properties easy to analyze. In Section 4 we analyze the image patch statistics given the knowledge about the activity of an independent component, whereas in Section 5 we fit a parametric model to the dependencies between two ICs (conditioning and conditioned component). In Section 6 we show a way to extend the features outside the original image patches. Finally, in Section 7 some conclusions are made of the results and future directions for the research are discussed.

2. DATA SELECTION

As data we used 24 images of landscapes, plants and animals. These images can be seen in Figure 2. The images were chosen so that they have features of different scales in them; some are taken up close to an object, some further away; in some a typical leaf covers only a few pixels, in others they take up a good portion of the image; some have blurriness due to depth-of-field, most do not; some are very complex, others simpler; some clouds are visible in the images, but no blue sky. We attempted to avoid saturation in the images, as it produces artificially smooth areas (even more so than blue sky or some man-made structures) pretty much useless for further analysis. Still, a little over one pixel in ten thousand is saturated, most of which are on the back of the white horses in the third image from the left on the third row in Figure 2.

The images were taken with a digital camera (Canon Ixus 400), converted to grayscale (using the Y component of the YIQ color space), and block averaged in four by four pixel blocks (and downscaled by the same factor). An area of 512 by 384 pixels was then selected of each of the downscaled images. The downscaling should pretty much negate artifacts brought by color interpolation, noise reduction and even compression.

The original images were compressed using the smallest level of compression available with the camera (i.e. compression level with the best quality), and no artifacts were visible to the naked eye. The average size of the (2272 by 1704 pixel) original JPG-images was 2.55 megabytes, i.e. about 5.3 bits per pixel (depending on the image, this was as low as 3.2 bpp or as high as 6.6 bpp). Due to the nature of the camera sensor, 'under' each pixel there is a sensor for only one of the color components, and thus the actual information content in a 24-bit uncompressed image obtained with such a sensor cannot really be more than 8 bits per pixel (including noise), with realistic estimates significantly below that. Thus the compression was quite close to being lossless. The compression is naturally also concentrated on the higher frequencies removed in the downscaling. As JPEG-compression works on 8 by 8 pixel blocks, downscaling also limits the possible compression artifacts to neighboring pixels. The downscaling should also remove possible slight unsharpness in the images and even reduce noise.

The images were saved in a 16-bit grayscale TIFF-format (after the histograms were stretched to cover the 16-bit space). The conversion to grayscale results in more gray levels than it is possible to describe using eight bits, and the averaging in 16-pixel blocks also brings additional gray levels. As we did not want to lose any information that could be used later, we chose a 16-bit intermediate format. It can also be argued that it is better that color gradients degrade to noise (as it is likely in a 16-bit format) than to a series of discrete steps (as is possible with eight bits). The original images, and the 16-bit TIFFs can be found at the web address **http://www.cis.hut.fi/inki/images/**.

Next we sampled 200000 12 by 12 pixel patches from these images. Note that different parts of the same original image can be considered as different images from a similar setting. The patches do not necessarily have to be completely non-overlapping either, but definitely more data would always be useful especially when estimating the stochastic variation in the statistics. Patch mean was subtracted from each patch. The orientation of the patches was normalized so that 'up' in the patches corresponds to 'up' in the original scene. Sometimes, due to the fact that usually the upper parts of the images are the brightest (e.g. parts of

Figure 2: The 24 images used in this work. Each image is 16-bit grayscale, 512 by 384 pixels. Not at full resolution here.

the sky are visible), even though the mean value has been subtracted from each patch, the pixels still do not have zero mean. Even though this was not a particular problem with the dataset used here, we randomly assigned a new sign to each of the patches. This reduced the pixel means on average by more than forty percent.

A small sampling of the patches can be seen in image 3. We have ordered these by increasing kurtosis in the corresponding direction in the whitened space, and as one can see, the 'simplest' patches (i.e. the smoothest patches, with fewest apparent edges, some of which are somewhat out of focus) are the most kurtotic. The word 'kurtotic' is interchangeable with the word 'sparse' for all intents and purposes here. Generally the patches do not

resemble ICA features, barring the most kurtotic, which correspond to the largest ICA features. To describe these 'simple' patches, one needs only a few of the largest ICA features, whereas to describe the more complex patches, one needs a plethora of smaller features. As we have normalized patch variances (norms), the few active features in the simpler patches also have higher activity levels than the numerous active features in more complex patches. This in turn explains why the largest features (patches) are more kurtotic, and why one should expect the dependencies to be more apparent with the simpler patches (features). Note that even though the more complex patches appear to be quite 'noisy', their non-downscaled versions show that there are actual objects and structures in them.

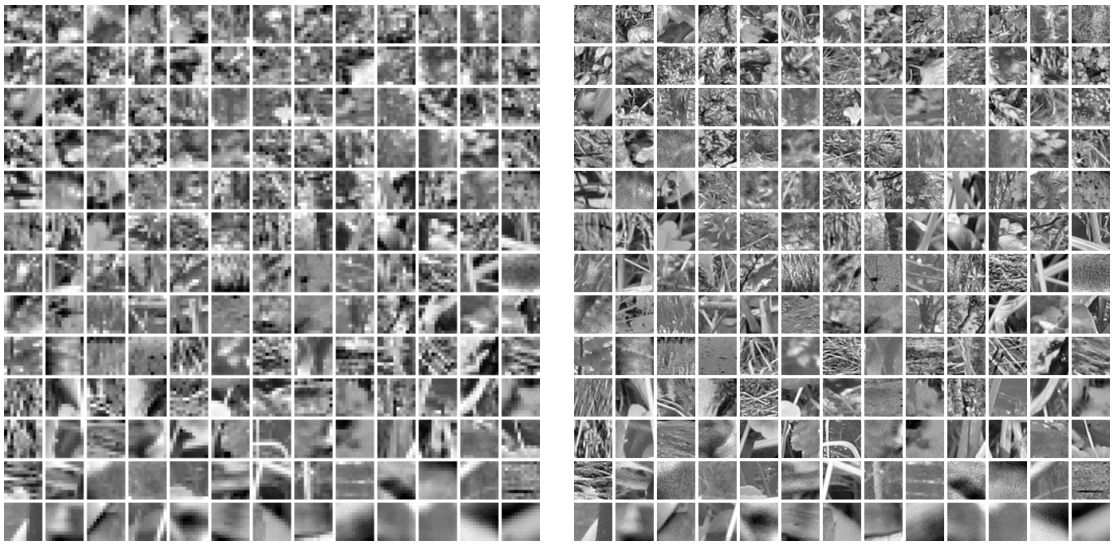

Figure 3: Some of the patches used in this work. Left: Patches ordered by kurtosis. Right: Same patches without downscaling. Nearly all the patches show underlying structures, and very little artifacts or noise.

3. ANALYSIS OF CONDITIONAL DEPENDENCIES

We will here be investigating how the statistics of the image patches change, when we know one (specific) independent component is highly active. By activity we mean that the absolute value of the estimated independent component ($|\mathbf{y}_i|, \quad \mathbf{y}_i = \mathbf{w}_i^T \mathbf{x}$, where $\mathbf{w}_i^T \mathbf{a}_j$ is one when $i = j$, otherwise zero.) exceeds some threshold α. That is, in these cases the component can be considered to describe something essential appearing in the patch, i.e. part of an edge or line. We denote the whole data set as

$$\mathbf{X} = \{\mathbf{x}(t) \quad | \quad t = 1, \ldots, N\}, \tag{2}$$

where the pixels of the patch with index t are stacked as a vector $\mathbf{x}(t)$, as in equation (1). We denote the indexes for which independent component i exceeds α by

$$I_{\alpha,i} = \{t \quad | \quad |y_i(t)| > \alpha\}, \tag{3}$$

and the subset of the whole data associated with $I_{\alpha,i}$ by

$$\mathbf{X}_{\alpha,i} = \{\mathbf{x}(t) \quad | \quad t \in I_{\alpha,i}\}. \tag{4}$$

3.1 PREPROCESSING

There are two essential steps of preprocessing we do here that are important for observing the statistics. The first is normalizing patch variances or patch norms (these differ by an irrelevant scaling factor). With this, and with the reduction of the mean and whitening of the data, the contrast variations between the patches are mostly eliminated. If this normalization was not done, it would be difficult to say whether a highly active feature actually describes something essential in the patch or whether the variance of the patch (contrast) is just high. With this normalization, speaking of activation is more meaningful in the sense that a certain level of activation of a feature means that the feature contributes a specified portion of the content (variance) of the patch.

As the patch variances have been normalized, each of the patches also corresponds to a direction in the high-dimensional space or a point on the surface of an origin-centered hypersphere. Therefore high activation of a component means we are looking at a segment (all the patches in the segment) on the surface of the origin-centered sphere within a fixed angle of a point (the active ICA feature). Actually, one ICA feature corresponds to two points on the surface of the sphere, but our other preprocessing step simplifies this (see below). Note that we demand that the patch variances (patch length in any orthogonal basis) equal unity and that the data is white simultaneously. The requirement for whitening can be written as

$$\mathrm{cov}(\mathbf{x}) = \mathbf{I} - \frac{1}{n}\mathbf{1}, \tag{5}$$

where the latter term on the right side results from eliminating the patch means [1]. For fixed t we require

$$\mathrm{mean}_i(x_i(t)) = 0, \qquad \mathrm{var}_i(x_i(t)) = 1 \quad \Longleftrightarrow \quad \|\mathbf{x}(t)\| = \sqrt{n}. \tag{6}$$

As the patch variance normalization affects the covariance of the data, whitening and patch normalization were repeated (alternately) a total of ten times. After this, the largest (nonzero) eigenvalue of the covariance matrix was less than one millionth larger than the smallest ($7.86 * 10^{-7}$ to be exact). Before the process had begun, the largest value was 370 times larger than the smallest, after the first whitening/normalization it was 1.7 times bigger, after the second 0.2 times etc. Note that this whole process can be described as whitening and patch normalization, as it corresponds to multiplying the data with a matrix (product of all the whitening matrices) and then normalizing the patches. The matrix has essentially the same eigenvectors as the original whitening matrix, but slightly different eigenvalues.

The second step of preprocessing before examining the dependencies is normalization by the sign of the active component. There is no inherent sign attached to an image patch as one component can be positive and another negative mostly regardless of each other in the

1. Projection matrix $\frac{1}{n}\mathbf{1}$ projects all the data to the space spanned by the mean value. If $\mathrm{cov}(\mathbf{y}) = \mathbf{I}$, then it is easy to show that $\mathrm{cov}((\mathbf{I} - \frac{1}{n}\mathbf{1})\mathbf{y}) = \mathbf{I} - \frac{1}{n}\mathbf{1}$.

patches. But closely related features (for instance two collinear edge detectors) may very well exhibit consistently the same signs when either one is highly active as this corresponds to having an edge in the patches to which both react. We will denote the sign-normalized data associated with independent component i and threshold α as $\mathbf{Z}_{\alpha,i}$:

$$\mathbf{Z}_{\alpha,i} = \{\mathbf{z}(t) \quad | \quad \mathbf{z}(t) = \mathbf{x}(t)\text{sign}(y_i(t)), \quad t \in I_{\alpha,i}\}. \tag{7}$$

So, with the possible exception of sign, $\mathbf{Z}_{\alpha,i}$ has the same elements as $\mathbf{X}_{\alpha,i}$ in Equation (4). We will denote the vectors in $\mathbf{Z}_{\alpha,i}$ as $\mathbf{z}_{\alpha,i}$.

3.2 FIRST AND SECOND ORDER STATISTICS

We will keep the analysis of the dependencies as simple as possible, while still hopefully exposing the underlying dependencies. For one, we look at the mean value of $\mathbf{z}_{\alpha,i}$:

$$\boldsymbol{\mu}_{\alpha,i} = E\{\mathbf{z}_{\alpha,i}\} \approx 1/T \sum_{t=1}^{T} \mathbf{z}_{\alpha,i}(t). \tag{8}$$

By denoting the matrix which projects the data into a subspace spanned by component i as $\mathbf{P}_i$ (if $\mathbf{a}_i$:s are of unit length, and the data has been whitened, $\mathbf{P}_i = \mathbf{a}_i\mathbf{a}_i^T$), we can write the portion of $\boldsymbol{\mu}_{\alpha,i}$ in the direction of component i as $\mathbf{P}_i\boldsymbol{\mu}_{\alpha,i}$. The rest of $\boldsymbol{\mu}_{\alpha,i}$ can be written as $(\mathbf{I}-\mathbf{P}_i)\boldsymbol{\mu}_{\alpha,i}$. We naturally use sample means when calculating expectations from the data.

We will also look at changes in the second order statistics as a function of α. Because the statistics in the direction of the active component are pathological by construction, we exclude this dimension from our analysis. If this dimension was kept, most of our methods presented here should be modified to take the effects of the patch selection and sign normalization in to consideration. The covariance matrix can be written as:

$$\mathbf{C}_{\alpha,i} = \text{cov}((\mathbf{I}-\mathbf{P}_i)\mathbf{z}_{\alpha,i}) = (\mathbf{I}-\mathbf{P}_i)(E\{\mathbf{z}_{\alpha,i}\mathbf{z}_{\alpha,i}^T\} - \boldsymbol{\mu}_{\alpha,i}\boldsymbol{\mu}_{\alpha,i}^T)(\mathbf{I}-\mathbf{P}_i)^T, \tag{9}$$

and we denote the (nonzero) eigenvalues of $\mathbf{C}_{\alpha,i}$ as $d_1^2, \ldots, d_{n-1}^2$ from the largest to the smallest ($d_i > 0, \forall i$), and the corresponding eigenvectors as $\mathbf{e}_1, \ldots, \mathbf{e}_{n-1}$. The corresponding components in the data are called principal components. Note that when $\alpha = 0$, $E\{\mathbf{z}_{0,i}\mathbf{z}_{0,i}^T\} = E\{\mathbf{x}\mathbf{x}^T\} = \mathbf{I} - \mathbf{1}\frac{1}{n}$, so the eigenvalues of $\mathbf{C}_{0,i}$ can differ from zero or one only in the dimension spanned by $\boldsymbol{\mu}_{0,i}$.

3.3 SOME DISCUSSION

First order dependencies between independent components have not been examined in this manner in the literature earlier. In order to observe these dependencies one needs large quantities of data, and one needs to normalize the signs of the patches by the sign of the active component. Indeed, this analysis most directly relates to the work done in psychology in relation to 'good continuation' of a feature (see Wertheimer, 1938, Geisler et al., 2001, Sigman et al., 2001), as the mean value exposes what usually accompanies (or completes) the corresponding feature.

Dependencies between second order statistics have been explored in the literature earlier. The dependencies between the variances of similar wavelets have been identified earlier (see

Buccigrossi and Simoncelli, 1999, Schwartz and Simoncelli, 2001), and extensions of ICA, where there is a correlation structure between the energies (squares) of the components, have been introduced earlier, see Hyvärinen et al. (2001a), Hyvärinen and Hoyer (2000), Karklin and Lewicki (2003). However, our analysis has many aspects not explored earlier. These arise from our way of looking at the dependencies and the ways we process the data. One could say that $\boldsymbol{\mu}_{\alpha,i}$ describes the 'static' part of the dependencies associated with independent component i and threshold α, and $\mathbf{C}_{\alpha,i}$ can be seen as describing the 'variable' part, at least partially.

One might ask, why do we look at the dependencies when the activation is greater than α, and not between α_1 and α_2. This would be an alternative to our analysis here, and especially when $\alpha_1 - \alpha_2 \to 0$, the analysis would be directly linked to the conditional distribution of the data given the value of IC i. However, as the interval $\alpha_1 - \alpha_2 \to 0$, the amount of data needed for this analysis would be enormous (without smoothing). Also, it would be difficult to speak of the highest activations (as $\alpha_2 \to \sqrt{n}$) due to the lack of data, so the interval (or smoothing) would have to increase for higher values of α. In our analysis, all the data above a threshold is included in the statistics (with associated probabilities), so we don't have this problem.

Note also, that it is possible to say something about the statistics when the activation is less than α. In this case elementary mathematics shows the mean value to be:

$$\boldsymbol{\mu}'_{\alpha,i} = \frac{\boldsymbol{\mu}_{0,i} - p(|y_i| \geq \alpha)\boldsymbol{\mu}_{\alpha,i}}{p(|y_i| < \alpha)}.$$

So, the mean value is a linear combination of $\boldsymbol{\mu}_{0,i}$ and $\boldsymbol{\mu}_{\alpha,i}$. The covariance is:

$$\mathrm{cov}(\mathbf{z}'_{\alpha,i}) = E\{\mathbf{z}'_{\alpha,i}\mathbf{z}'^T_{\alpha,i}\} - \boldsymbol{\mu}'_{\alpha,i}\boldsymbol{\mu}'^T_{\alpha,i} = \frac{\mathbf{I} - \frac{1}{n}\mathbf{1} - p(|y_i| \geq \alpha)E\{\mathbf{z}_{\alpha,i}\mathbf{z}^T_{\alpha,i}\}}{p(|y_i| < \alpha)} - \boldsymbol{\mu}'_{\alpha,i}\boldsymbol{\mu}'^T_{\alpha,i}.$$

Therefore, with the possible exception of the dimension spanned by $\boldsymbol{\mu}'_{\alpha,i}$, the eigenvectors of $\mathrm{cov}(\mathbf{z}'_{\alpha,i})$ are the same as they are for $\mathrm{cov}(\mathbf{z}_{\alpha,i})$. The eigenvalues are different, and their order (w.r.t. size) is inverted. So, the components that increase activity when the activation of an IC is greater than α decrease activity when activation is less than α.

The description of the changes in these statistics goes a long way in describing the changes brought on by the activation of a component. The skewness, kurtosis and other higher order statistics also change, but their change is partially linked to the change in mean and variance. For example, the shift in the mean value for high values of α may be visible in the skewness with lower values of α as the small amount of patches where the mean is clearly shifted may form a tail for the distribution. Note also, that we have normalized the variances (to unity) and mean values (to zero) when no components are active, but we have not normalized any other statistics. Therefore quantifying which portion of these other statistics (or of their change) is due to the activation of an independent component is essentially harder.

Also, the fundamental way of describing first order statistics of a group of variables is by a vector, second order statistics by a (covariance) matrix, and higher order statistics by higher dimensional tensors. So, there are generally no inherent bases associated with higher order statistics. This also means that it is very hard to normalize skewness and kurtosis (in

all bases simultaneously), even if such a procedure was reasonable. Note also, that one of the reasons for the popularity of ICA is the difficulty of analyzing higher-order statistics. ICA can reveal some interesting structures, even if the actual generative model for the data would be quite far from the assumptions of ICA.

Of course, one could incorporate higher order statistics in the analysis by searching for a new ICA basis when a component is highly active, as we did in Inki (2003a) without normalizing the data by the sign of the active component, but it appears that the basis is not in any significant way different from the ICA basis found in the entire data set. Most of the changes can be attributed to the changes in PCA values, i.e. there are more basis vectors close to the highly active component, less further away from it.

4. STATISTICS CONDITIONED ON THE ACTIVITY OF AN IC

FastICA (Hyvärinen, 1999) in symmetric mode with the hyperbolic tangent nonlinearity was then used to perform ICA on the data. The basis vectors we found can be seen on the top left in Figure 4. The features are represented here in the original, not whitened space. The components are very similar to those we obtained with a different data set in Inki (2003a). For comparison, on the top right side in Figure 4 we have the components found in the same data set without patch variance normalization. They appear to be quite similar, though there are maybe a little less very small features in the latter, and the shapes are a bit less compact.

One should note that the convergence properties of the algorithm do not matter for our purposes here, as long as it converges to a local extremum of the objective function (minimum of mutual information, maximum of likelihood). The nonlinearity used in FastICA (and many other ICA algorithms) is related to the source distribution, and can have an effect on the results, especially in cases where the data has not been generated according to the ICA model. Hyperbolic tangent nonlinearity relates to a supergaussian source distribution somewhat similar to what has been observed in natural images. If we use a cubic nonlinearity, which corresponds to a subgaussian distribution and weights large values much more, one obtains features that are somewhat longer and more step-function like as can be seen on the bottom of Figure 4. This difference is arguably linked to the changes we notice in this work in the mean values of the patches, when a certain feature is highly active. Nevertheless, we only use the top left basis of Figure 4 in our experiments here, because using a hyperbolic tangent is sensible with natural image data, and arguably as good a choice as any other.

Note also that different starting values produce different bases for image data. In addition to the normal ICA indeterminacies regarding the permutation, sign and variance of the features in a basis, different starting values usually yield bases that have similar properties, but the individual features are different (suggesting that an overcomplete description of image data could be appropriate). Still, any such set of features can be obtained from a small set of 'basic' Gabor-like features using translation, rotation and scaling. Again, we only use the results from one run of FastICA in this work. This basis should be about as good as any other basis FastICA would find from our data set using different starting values.

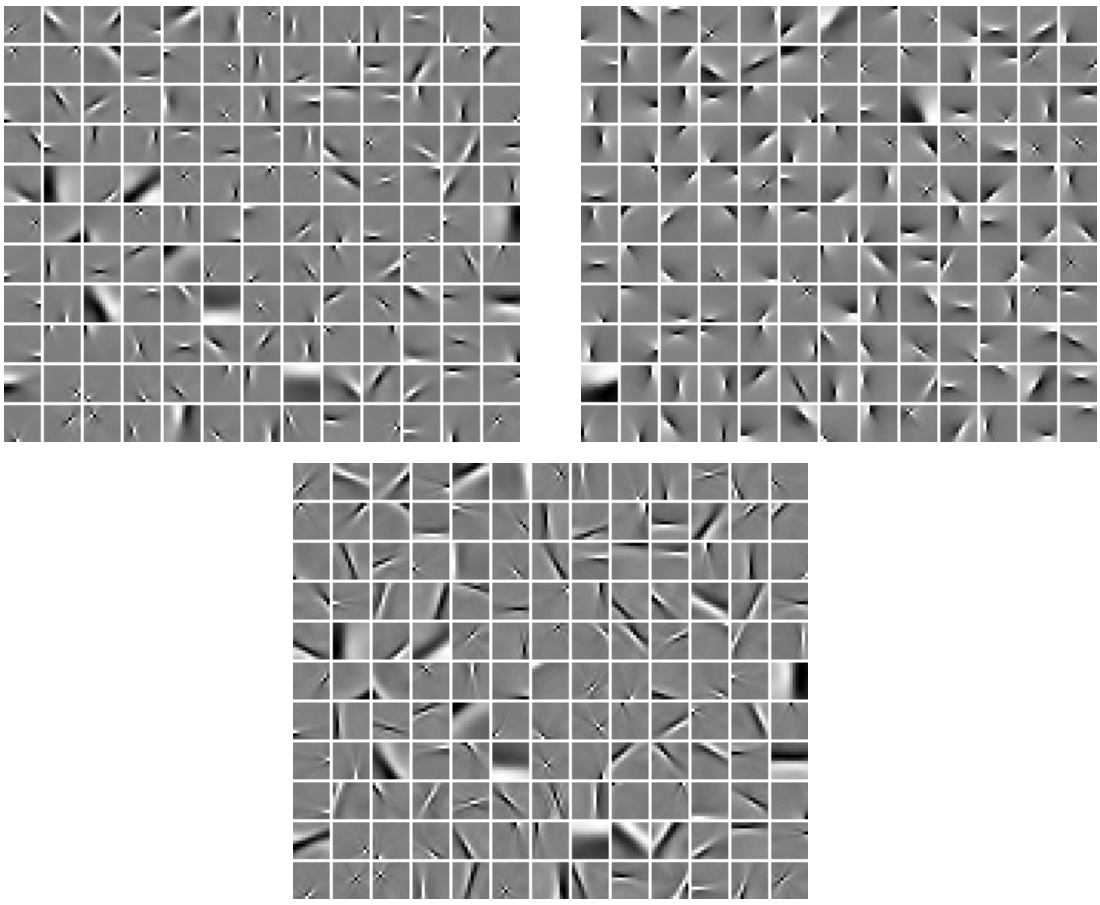

Figure 4: ICA bases. Top left: ICA basis found in normalized data using hyperbolic tangent nonlinearity. Top right: ICA basis found in data without patch variance normalization. Bottom: ICA basis with third power nonlinearity and using the top left basis as the starting point.

4.1 EFFECT OF THRESHOLD ON AVAILABILITY OF DATA

First a few notes on the effect of the threshold α (Equation 3) on the number of available patches. With $\alpha = 0$, all the independent components (ICs) have 200000 sample points. As α increases the number of available patches starts to decline approximately exponentially. On the other hand, as α increases, the changes in the statistics become more apparent. Note also that the maximum value of α with 12 by 12 pixel patches was $\sqrt{n} = \sqrt{143} \approx 11.96$.

In Table 1 we have the numbers of patches available for the IC with the most patches and the IC with the least patches, and the average number of patches. The differences between components in w.r.t. the number of available patches increase as α increases. With $\alpha = 2$, all the components have approximately the same amount of patches (about 5% of the total number of patches) but with a threshold of three, the component with most patches

	$\alpha = 1$	$\alpha = 2$	$\alpha = 3$	$\alpha = 4$	$\alpha = 5$	$\alpha = 6$	$\alpha = 7$
maximum	53618	11861	4328	1895	774	303	114
mean	49671.6	11281.3	2657.0	674.4	185.7	53.8	16.0
minimum	36546	10061	2052	305	38	3	0
percent of variance	0.70	2.80	6.29	11.1	17.5	25.2	34.3
patches included (%)	100	100	87.7	37.4	12.2	3.75	1.14

Table 1: The numbers of patches available for the component with most and least patches, and the average number of patches for different values of α. Also listed are the (minimum) percentages of the variance of the patch that the feature contributes for a given value of α, as well as the percentage of patches that is included in at least one of the active sets $I_{\alpha,i}$ (Equation 3).

has more than twice the number of patches of the component with least patches. With a threshold of five, this difference is about 2000%.

Also, in order to have as many samples to estimate the statistics with $\alpha = 5$ as one has with $\alpha = 3$, one would on average need more than fourteen times the amount of data, although for the components with least available patches the factor is more than fifty. One can see that the amount of data needed in this analysis soon becomes prohibitive as α increases. Note also that the components that have less than the average number of patches for small values of α (less than about 2) have more than the average number for larger values of α, i.e. are more kurtotic.

We will have to limit our analysis of the statistics mostly to only a few different values of α. As a compromise between the number of samples, and size of the dependencies, we selected $\alpha = 3$ as the baseline value for the analysis. Almost all of the patches are still used in the analysis with this value of α, as one can see in Table 1. The patches that do not get used include mostly only the most complex or least kurtotic (cf. Figure 3). We also use values $\alpha = 0$, $\alpha = 1$ and $\alpha = 5$ where possible. Using values higher than $\alpha = 5$ is quite difficult due to our limited sample size.

4.2 FIRST ORDER STATISTICS

As we have normalized each patch by the sign of the active component, one may assume that the changes in activity level between components is visible even in the first order statistics. Assuming you have two collinear features, i.e. they both describe the same straight edge, one may also assume that the activation of one is often accompanied with the activation of the other so that they together form a consistent edge. Therefore the signs of the components are in a way 'interlocked' and normalizing the patches by the sign of the active component shifts the mean value of the other, even though the components are orthogonal.

In Figure 5, we have the mean values $\boldsymbol{\mu}_{\alpha,i}$ (Equation 8) for the patches in the original (not whitened) space for each of the ICs. On the left side of each group, we have the original features. The next four features correspond to the mean values for $\alpha = 0$, $\alpha = 1$, $\alpha = 3$, and $\alpha = 5$, respectively. On the right side of each group is the portion of the mean

value for $\alpha = 5$ not in the direction of the original feature. The mean values appear to be actually closest to the original features with a threshold of one, and with higher thresholds, the mean value features are much longer and more step-function -like, especially for the larger features. That is, even though the transition between the positive and negative lobes in the features is not essentially slower, the activation persists much further from the zero crossing.

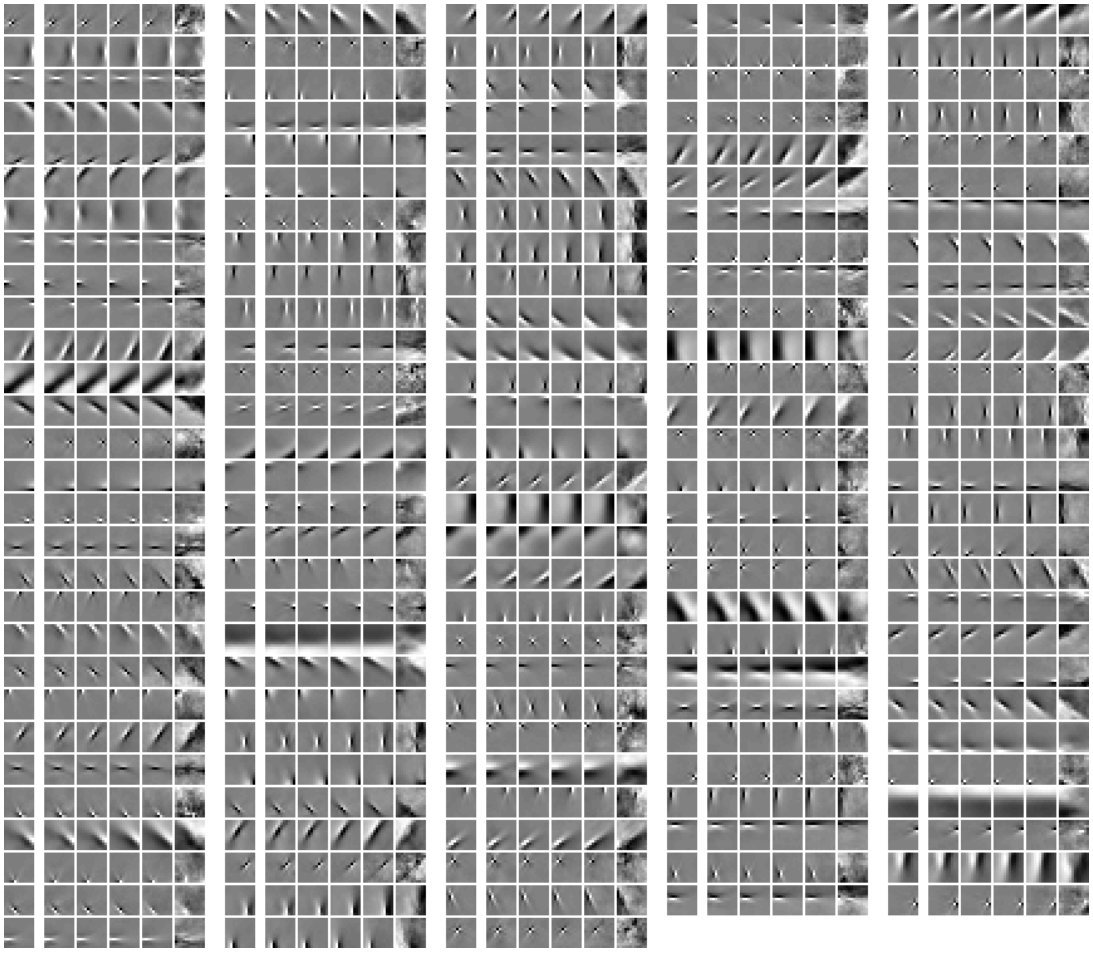

Figure 5: Conditional mean values of the whitened patches in the original space, when normalizing by the sign of the active component and using different thresholds. If the data had been generated according to the ICA model, these mean values would be identical to the basis vector on the left side of each group for any threshold. The remaining features in each group from left to right: α=0 , α=1, α=3, α=5 and the portion of the mean value orthogonal (in the whitened space) to the original feature for $\alpha = 5$.

In Figure 6, we have the length of the mean value vector, $\|(\mathbf{I}-\mathbf{P}_i)\boldsymbol{\mu}_{\alpha,i}\|$, as a function of α. Note that the data has been whitened, so the standard deviation in any direction is one without activation. On the right side of Figure 6, we have divided the length of the mean value by the size of the mean value in the direction of the active component, i.e. $\|(\mathbf{I}-\mathbf{P}_i)\boldsymbol{\mu}_{\alpha,i}\|/\|\mathbf{P}_i\boldsymbol{\mu}_{\alpha,i}\|$. This relates directly to how far visually the mean values in Figure 5 are from the original features. For later use in this work, we have also plotted a second version of each of the graphs (with dotted lines) that depict how the situation changes, when about 30 percent of the smallest components are removed from consideration.

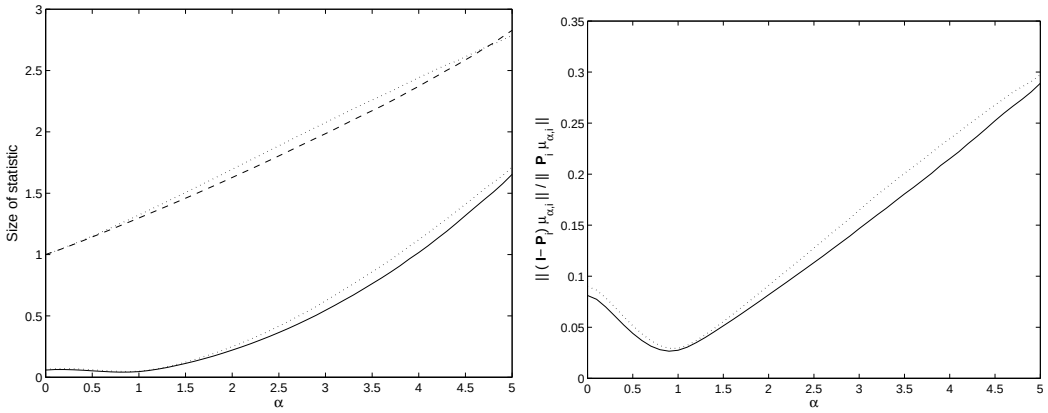

Figure 6: Sizes of the statistics. The solid line on the left graph represents the average size of the mean value of the patches as a function of the threshold used in patch selection. The dashed line represents the standard deviation in the maximal direction (i.e. d_1) as a function of threshold. The direction used for patch selection is removed before this analysis. On the right side: The length of the mean value in orthogonal directions to the original ICA direction divided by the length of the mean value in the ICA direction, averaged over all components. Dotted lines: The same graphs when about 30% of the smaller components are removed from consideration.

As one can see, the changes in the statistics are significant for larger values of α. The shift in the mean value equals the standard deviation without activation when $\alpha = 4$. However, if this mean value corresponded to the whole content of the patches, its value would be $\sqrt{n} \approx 11.96$. Of course, the direction of the active component captures some of the content of the patch, and the value converges to zero when $\alpha \to \sqrt{n}$. However, for reasonable values of α, orthogonal dimensions correspond to larger portions of the mean value for larger values of α, as can be seen on the right side of Figure 6. Again, $\alpha \approx 1$ stands out as the mean values are closest to the original features near it.

4.3 SECOND ORDER STATISTICS

Second order statistics are harder to visualize, because the basis is so essential in the visualization. For each active component second order statistics correspond to a matrix

(covariance matrix), not a vector of numbers (mean values). This matrix is completely described by its eigenvectors and eigenvalues, i.e. the second order statistics are fully described by the principal component analysis (PCA) of the data $\mathbf{Z}_{\alpha,i}$ (Equation 7). However, the principal components are different for each active component (and threshold), which makes the analysis for a large number of independent components hard. One can also examine the second-order structures in the (fixed) ICA basis, but then visualizing and quantifying the cross-correlations is harder.

Here we examine the standard deviations of the PCA components and investigate how the sizes change as the activation increases. We also look at the PCA basis vectors for one particular IC and one particular threshold.

We have plotted the standard deviation of the largest principal component, i.e. d_1 (Equation 9), as a function of α on the left side in Figure 6. The change in d_1 is essentially linear as a function of α. Also noticeable is that it increases at about the same rate as the length of $\boldsymbol{\mu}_{\alpha,i}$ for large values of α. It equals the standard deviation (without activation) when $\alpha = 3$. Note, however, that d_1 also converges to zero when $\alpha \to \sqrt{n}$.

In Figure 7, we have the standard deviations of the PCA components $(d_1, \ldots, d_{n-1})$ plotted for several different values of α, averaged over all the possible active ICs. As one can see, normalizing by the sign of an IC does not in itself change the variance structure of the data at all ($\alpha = 0$). As we know from Equation (9), this structure can only change in the direction of $\boldsymbol{\mu}_{0,i}$, and here $\|\boldsymbol{\mu}_{0,i}\|^2$ is very small. When $\alpha = 1$, there is a noticeable increase at the high end (for similar features) and a drop-off at the low end (very different features). For $\alpha = 3$ the largest changes in variance are mostly concentrated to the high end. This means that for low thresholds, the question is equally "what is left out of the data", as "what is included", whereas for higher thresholds it is more the latter. This supports the idea that that in patches where a feature is highly active, the feature is essential in describing an object or edge, instead of just being more active than some other feature(s).

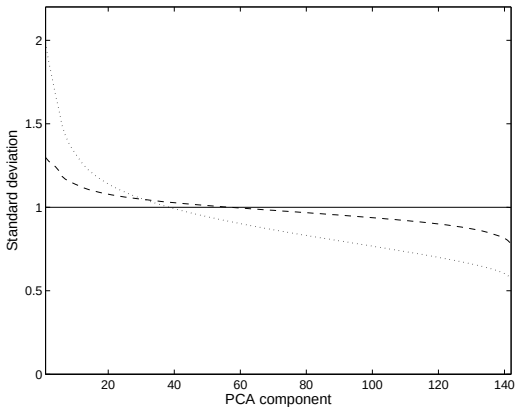

Figure 7: Standard deviations of the PCA components for different thresholds averaged over all ICs. Solid line: $\alpha = 0$. Dashed line: $\alpha = 1$. Dotted line: $\alpha = 3$.

Next in Figure 8 we have the PCA eigenvectors $\mathbf{e}_1, \ldots, \mathbf{e}_{n-1}$ in ascending order when the second component on the first row of the ICA basis on the left in Figure 4 is active, for two different values of α. The largest components (top left) appear to be active in about the same position and orientation as the original component. The very smallest components, i.e. components that decrease in activity, appear to have properties as far away from the original component as possible. Otherwise the components appear to be 'noisy' rather than linked to the activation of the original component. It is quite difficult to give a definite qualitative or quantitative analysis of the dependencies from this graph. The standard deviations of the PCA components are quite close to the averages in Figure 7.

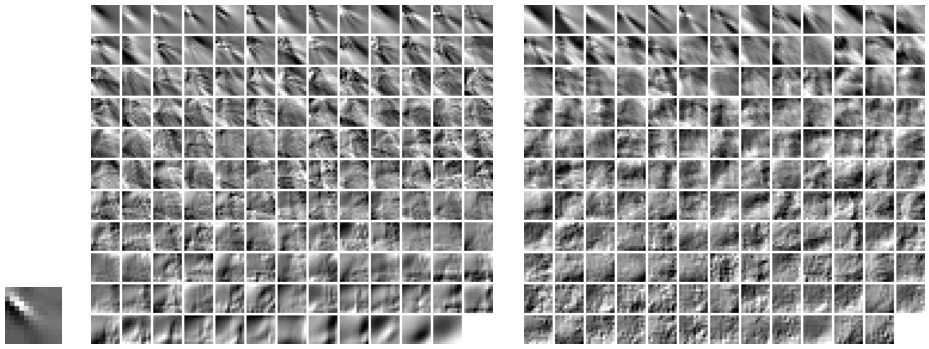

Figure 8: PCA components ordered by variance, when the second independent component on the left side of Figure 4 is active. Leftmost: the active feature. Center: $\alpha = 1$. Right: $\alpha = 3$.

Even though it is difficult to give a thorough interpretation of qualities of the second order dependencies, some speculations can be made. Note that the component we selected as an example here is the same one we used as an example in the introduction. As we observed from Figure 1, the actual edge can, for example, have a slightly different orientation to the feature in question, it may start to bend, or even though the orientations are similar, the edges in the patch and the feature might not align perfectly. However, as such 'errors' can occur in any direction, the integral over all of them is basically zero. Therefore such considerations will not show up in the mean value $\boldsymbol{\mu}_{\alpha,i}$. The probabilities of the different possibilities is, however, reflected in the second order statistics. For example, the variance of adjacent orientations (with similar sizes and locations) could be assumed to be higher, which can be argued to be supported by the basis functions in Figure 8. As the analysis of the qualities of the second order dependencies given here is somewhat unsatisfactory, we will later attempt to assess them in the ICA basis.

4.4 ASSESSING RANDOMNESS IN THE STATISTICS

As we could infer from in Figure 8, estimation error due to limited sample size (seen as noise) has a noticeable effect on the second order statistics we obtained. We will now try to quantify the amount of estimation error, or randomness due to limited sample size, in our estimates of the statistics.

We made a new version of the data, where by construction the value of the conditioning (active) IC does not affect other components. We call this $\mathbf{U}_{\alpha,i}$. For each component i and every patch $\mathbf{z}_{\alpha,i}(t)$ in Equation (7), we keep the active component, and select randomly a patch $\mathbf{x}(t_2)$ from which we take the other components. We multiply these other components so that the variance of the new patch is normalized. That is:

$$\mathbf{u}_{\alpha,i}(t) \leftarrow \mathbf{P}_i \mathbf{z}_{\alpha,i}(t) + (\mathbf{I} - \mathbf{P}_i)\mathbf{x}(t_2) \frac{\sqrt{n - \|\mathbf{P}_i \mathbf{z}_{\alpha,i}\|^2}}{\|(\mathbf{I} - \mathbf{P}_i)\mathbf{x}(t_2)\|}. \tag{10}$$

Figure 9 is an illustration of what portion of the observed dependencies can be considered to be a result of the limited sample size. The solid line depicts $d_1 - \text{mean}_l(d_l)$ for $\mathbf{u}_{\alpha,i}$ averaged over i divided by the same for $\mathbf{z}_{\alpha,i}$ averaged over i. The dashed line depicts $\|(\mathbf{I} - \mathbf{P}_i)\boldsymbol{\mu}_{\alpha,i}\|$ for $\mathbf{u}_{\alpha,i}$ averaged over i divided by the same for $\mathbf{z}_{\alpha,i}$. If these quantities hovered around one, randomness would be a sufficient explanation for the quantities of the statistics we observe in the context of this analysis, but we have already seen that there are definite structures.

As one can see, despite our relatively big sample size, estimation error is still considerable in the statistics. This is probably one reason why some of the results in this paper have not been presented earlier elsewhere: in order to make the dependencies visible, one needs a lot of data. Near $\alpha = 1$, the first order statistics in the random version of the data are actually bigger than the observed dependencies. It can thus be said that near $\alpha = 1$, the mean values do not statistically differ from the basis vectors in the context of this analysis. For the second order statistics, the randomness does not appear to be as bad, but that is just part of the story, as we compare here only the sizes of the largest PCA components. Note that near $\alpha = 0$, the results from this analysis are no longer reasonable for second order statistics, as we have so exactly whitened our data sample and $\|\mu_{\alpha,i}\|^2$ is small.

In Figure 10 we compare the average sizes of $d_1, \ldots, d_{n-1}$ for $\mathbf{z}_{3,i}$ and $\mathbf{u}_{3,i}$. Even though the difference for d_1 is large, it rapidly shrinks, and most of the PCA components can be seen as mostly noise, as was apparent in Figure 8.

On the left side of Figure 11, we have the ICA features $\mathbf{a}_i$ sorted by $\|(\mathbf{I} - \mathbf{P}_i)\boldsymbol{\mu}_{3,i}\|$ in ascending order. It is readily apparent that the largest shifts in mean values occur for the largest, most kurtotic components, and actually the components are ordered pretty much by their size. On the right side, we have the a scatterplot of $\|(\mathbf{I} - \mathbf{P}_i)\boldsymbol{\mu}_{3,i}\|$ in the observed data, versus the same in control data. It would appear that the portion of this length that can be explained by random variation does not increase as the observed length increases, actually a small decrease is noticeable. Thus the observed statistics for the largest components are quite reliable, whereas for the smallest they can mostly be explained by randomness. This is related to the higher kurtosis (sparseness) of the larger features.

In Figure 12 we have the same plots for the second order statistics. Again, the ICA features are essentially sorted by their size, and the statistics for the largest components are the least 'noisy'. Thus one can conclude that we can be quite confident that the statistics we observe for the largest components are quite accurate, even though in general the estimation error in the statistics is higher than one would wish. To observe the statistics better, one would need more data (samples), and also possibly more images to sample the data from.

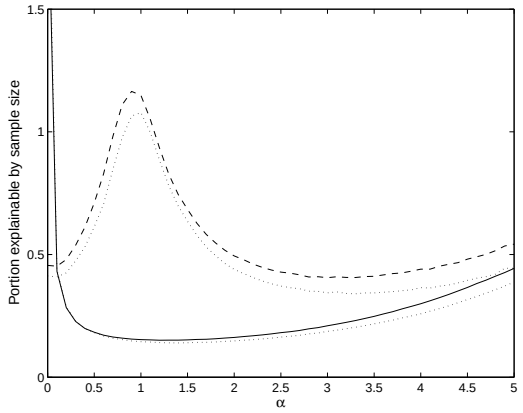

Figure 9: Comparing the statistics to the random case. Values on y-axis relate to the proportion of the observed dependencies that can be attributed to finite sample size, values of α are on the x-axis. Solid line: Comparison for second order statistics. Dashed line: Comparison for first order statistics. Dotted lines: Same graphs with about 30% of the smallest components removed.

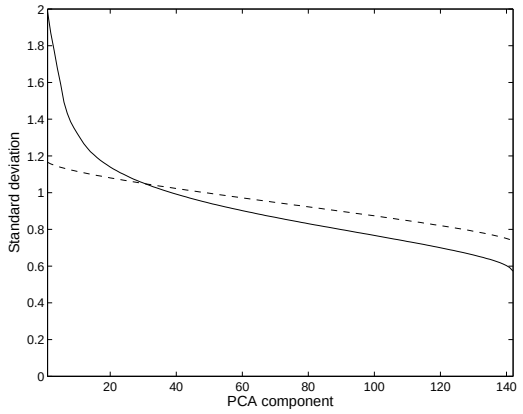

Figure 10: Standard deviations of PCA components compared to a random control case. Values of d_i on the y-axis. Solid line: observed from real data. Dashed line: observed from control data.

5. A PARAMETRIC MODEL OF DEPENDENCIES BETWEEN ICA FEATURES

In order to better analyze the qualities of the dependencies, we decided to estimate a model where the properties of the features predict the change in the statistics for a given value of α. So, the inputs are calculated from the properties of the active and the 'reactive' feature

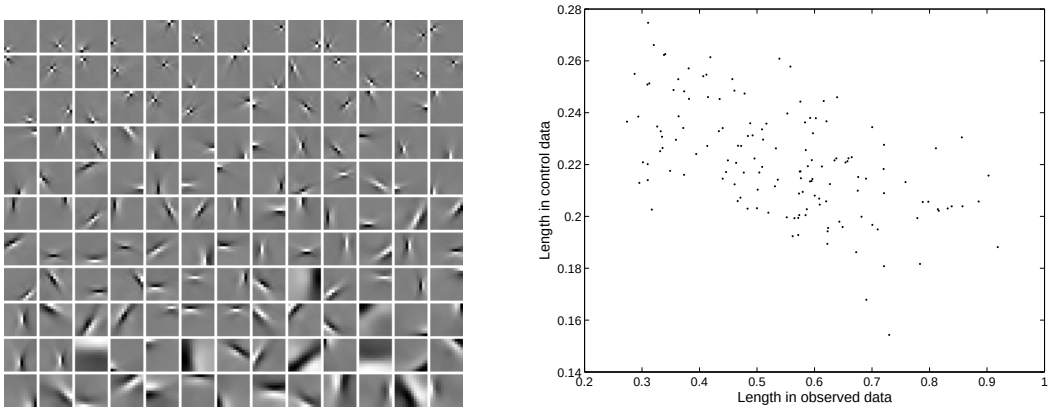

Figure 11: On the left: ICA basis sorted by size of first order statistics at $\alpha = 3$. Right: the length of $(\mathbf{I} - \mathbf{P}_i)\boldsymbol{\mu}_{3,i}$ plotted for each i in the control data versus observed data.

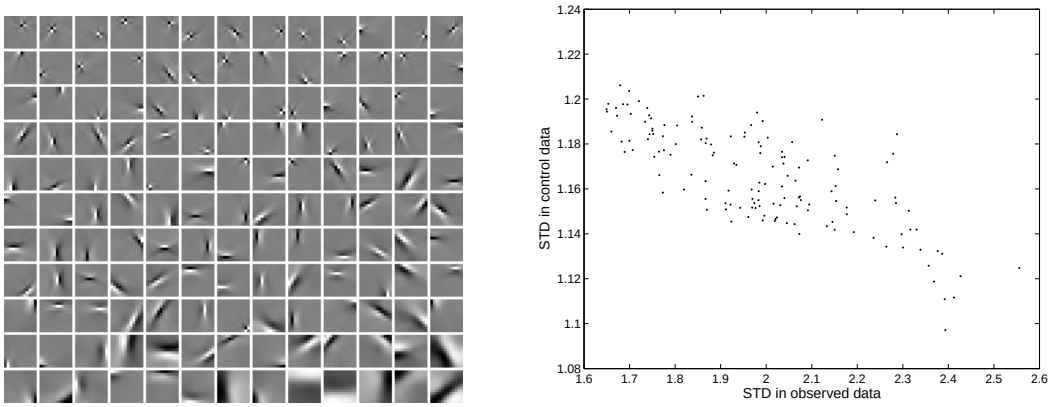

Figure 12: On the left: ICA features i sorted by square root of the largest PCA eigenvalue (d_1) associated with $\mathbf{X}_{3,i}$. Standard deviation (i.e. d_1) plotted for each component i in control data versus observed data.

(or the conditioning and conditioned feature), and the output should be the statistic that interests us, e.g. variance of the reactive component.

We will restrict our analysis to the case when all the conditioning and conditioned features are from the same basis. In order to extract all the information available in the PCA analysis of the previous Section, we would need to analyze all possible conditioned components (orthogonal to the activating feature), not restricted to one ICA basis or even any basis where the features are Gabor-like. The properties of first order statistics are fully described using only a single basis for the conditioned components, although these properties are possibly more apparent in the mean value features of Figure 5.

5.1 A MODEL FOR ANALYZING THE DEPENDENCIES

In order to estimate the properties of the features, we fitted Gabor functions to the ICA features in the top left basis in Figure 4. Even though the features are not exactly Gabor functions, many of their properties (size, orientation, position) can be described using Gabor functions. As Gabor functions we used real-valued two-dimensional functions, see Daugman (1988):

$$g(\mathbf{r}) \propto \exp(-\sum_{i=1}^{2} \frac{r_i^2}{2b_i^2}) \cos(2\pi\omega r_1 + \theta). \tag{11}$$

Here $\mathbf{r}$ is the two dimensional position vector, b_i:s are the widths in corresponding dimensions of $\mathbf{r}$, θ is the phase and ω the frequency. Any Gabor function can now be obtained with a rotation, translation, and scaling of $g(\mathbf{r})$. Let us denote the angle of this rotation as β.

We used four properties of the Gabors as the input. First input parameter was the logarithmic difference between the widths of the components, $G_{i,j}^{WI} = \log(b_i/b_j)$. Second input, $G_{i,j}^{OR}$ is the difference between the orientations of the conditioning component i and conditioned component j, $G_{i,j}^{OR} = |\sin(\beta_i - \beta_j)|$. The third was a measure of collinearity: new features are obtained by ignoring the attenuation of the Gabors along r_2, i.e. $g^{new}(\mathbf{r}) = g(\mathbf{r}) \exp(r_2^2/2b_2^2)$, these new features are normalized (w.r.t. inner product with themselves), and $G_{i,j}^{CO}$ is the absolute value of the inner product between these new features for the conditioning and conditioned components. The fourth input variable $G_{i,j}^{OL}$ was obtained by discarding the cosine part of the Gabors, i.e. $g^{new}(\mathbf{r}) \cos(2\pi\omega r_1 + \theta) = g(\mathbf{r})$, normalizing, and taking the inner product of these new features, i.e. it was an overlap measure that depends on the positions and sizes of the features.

Our model was of the type

$$R'_{i,j} = f(f_{WI}(G_{i,j}^{WI}) f_{OR}(G_{i,j}^{OR}) f_{CO}(G_{i,j}^{CO}) f_{OL}(G_{i,j}^{OL})), \tag{12}$$

where the functions on the right side are adjusted so that the squared error between $R'_{i,j}$ and the observed statistic $R_{i,j}$ is minimized over all conditioning ICs i and conditioned ICs j. As $\{f_{WI}, f_{OR}, f_{CO}, f_{OL}\}$, we used functions consisting of evenly spaced five points (the smallest of which is at the smallest value of corresponding $G_{i,j}$, the largest at the largest value) that were interpolated with piecewise cubic Hermite interpolation as implemented in Matlab version 6.5.

The function f consisted of eleven unevenly spaced points. The first point is at the minimum value of its input, i.e. the zero percent mark, the second at the five percent mark (where five percent of its input are smaller), third at the ten percent mark, then 20%, 35%, 50%, 65%, 80%, 90% 95% and the final one at the 100% mark. Additionally, f was required to be monotonically increasing and positive. With $\{f, f_{WI}, f_{OR}, f_{CO}, f_{OL}\}$ the model had a total of 31 free parameters. Of these 31 parameters, four are actually redundant, as the scaling in f can offset the scaling in any (and all) of the other functions.

As $R_{i,j}$ we used the variances of the conditioned components, as well as the absolute values of the mean values of the conditioned components. When fitting Gabor functions to the ICA features, we excluded the smallest features that cannot be so well described as Gabor functions from the analysis. We picked 98 of the best fitting features, so there were a total of 9506 ($= 98^2 - 98$) examples of $R_{i,j}$ for further analysis. We presented some

statistics with only these components in Figures 6 and 9, and even though there is a slight shift, the basic properties of these statistics do not essentially change. We used threshold $\alpha = 3$ in these experiments.

Note that of these only the first parameter G^{WI} can capture nonsymmetric properties of the dependency, i.e. if the places of the conditioning and conditioned component are exchanged, the dependency can change. Actually, this variable is antisymmetric in the sense that if this exchange is performed, the value of the variable changes sign but not size.

5.2 RESULTS

We fitted the model in Equation (12) to the statistics for $\alpha = 3$. The minimum of $\sum_{i,j}(R_{i,j} - R'_{i,j})^2$ was searched by Matlab's fminsearch -function which uses the Nelder-Mead method not requiring derivatives. Although the convergence properties of the search are out of the scope of this paper, it should be noted, that the function needed to be run several times (using the result of the previous stage as the starting point) in order to find a stable result. The error we ended up with was 0.2656 (i.e. 26.6%) of the variance of $R_{i,j}$ for second order statistics, and 0.4787 of the variance of $R_{i,j}$ for first order statistics.

We have plots of f_{WI}, f_{OR}, f_{CO}, and f_{OL} in Figure 13. The functions for second order statistics have been plotted with solid lines. The first function f_{WI} shows a maximum at zero, i.e. when the functions are of the same size. The function is also slightly nonsymmetric, so the overall model is slightly nonsymmetric. The second variable has a maximum at zero, i.e. when the orientations of the components are identical (or differ by π). The third component shows a maximum when the modified features overlap the most, i.e. are in a sense collinear. The fourth shows a maximum when the overlap of the features is the greatest. These functions are plotted in logarithmic scale, and their precise values do not matter, for which reason their maximum values have been normalized. The differences between their ranges does matter, and one can see that this is the biggest in the case of the fourth function, and smallest in the case of the first function. This would strongly suggest that the fourth parameter is the most important for the fit, and the first is the least important.

We have also plotted the same functions for first order statistics with dashed lines in Figure 13. As one can see, these functions have similar shapes to the corresponding functions for second order statistics, yet there are differences. In the middle range of the variables, the third and fourth functions appear to be 'flat'. Again, judging by the ranges of the functions, the fourth variable (overlap) appears to be the most important for the fit, but now the importance of the second variable (difference of orientation) seems to be smaller, and the third variable (collinearity) appears to be more important.

In Figure 14, we have scatterplots, where on the x-axis are the values of $f^{-1}(R'_{i,j}) = f_{OR}(G^{OR}_{i,j}) f_{WI}(G^{WI}_{i,j}) f_{CO}(G^{CO}_{i,j}) f_{OL}(G^{OL}_{i,j})$ and on the y-axis the observed values $R_{i,j}$. Both axes are in logarithmic scale. On the left side we have the scatterplot for the first order statistics, and on the right side the scatterplot for second order statistics. Also plotted in the figures with a solid line is the corresponding function f. The interpolation of function f was done on logarithmic scale.

It appears that for second order statistics, the function f has a somewhat sigmoidal shape, i.e. if $f^{-1}(R'_{i,j})$ is very high or very low, the actual estimate of the dependency is

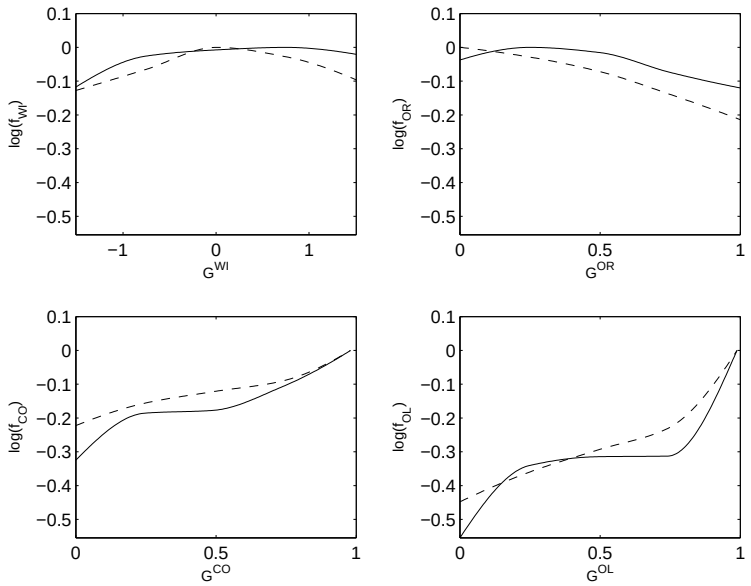

Figure 13: Plots of the logarithms of f_{WI}, f_{OR}, f_{CO}, and f_{OL}. Top left: logarithm difference in the sizes of the components. Top right: Difference in angle. Bottom left: Collinearity measure. Bottom right: measure of how much the features overlap. Solid lines: first order statistics. Dashed lines: second order statistics.

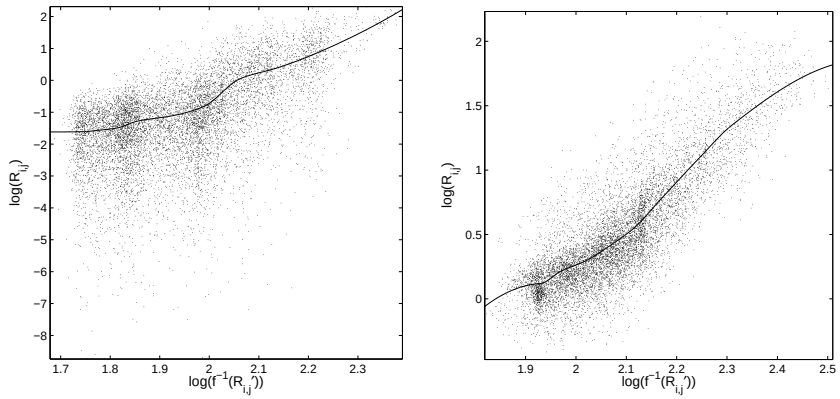

Figure 14: Plots of $R_{i,j}$ against $f^{-1}(R'_{i,j})$. Left side: First order statistics. Right side: Second order statistics. Also plotted in the figures with solid line is the corresponding f.

shifted towards the average. This is because very low values indicate that all the individual functions give a low value, but already one or two substantial differences in the properties

(say, overlap of the functions is insignificant) is enough to make them virtually independent. Similar argument could be made of very high values.

It is harder to say anything of the shape of function f in the case of first order statistics, possibly due to the lesser fitness of the model. Still, the smallest values are shifted towards the average. The model appears to fit quite well to the largest values, but not so well to the smaller values.

Another way of exploring how important the different variables are for the fit is by excluding one variable (and the corresponding funtion) from the model, and fitting it again. For the second order statistics, excluding the first variable G^{WI} produced an error of 0.2869, excluding the second G^{OR} 0.3668, excluding the third G^{CO} 0.3084, and excluding the fourth G^{OL} 0.4775. This supports our earlier conjecture that the fourth variable is the most important, and the first the least important in the fit. Now one can also identify that the second variable is also quite important for the fit.

These results we basically support our earlier conjectures that the changes in the second order statistics for the most part describe the 'error' between the feature, and the actual object (edge) that activates it. It can be argued that nearby orientations are active, because the orientation of the feature is not necessarily the same as that of the edge. Same argument can be made of size difference. Overlap is important because the activating object is (always) present near the feature (when looking at high activations), and the probability for its presence diminishes as the distance grows.

For the first order statistics, the errors were 0.4923, 0.5128, 0.5846, and 0.6844, for excluding the first, second, third, and fourth variables respectively. This supports our earlier reasoning that the second variable G^{OR} is not so important for the fit as the third G^{CO}, which makes sense as we have earlier argued that collinearity is important for the size of the mean value of the conditioned feature. Orientation is not so important because similar orientation without collinearity does not produce a consistent edge.

So, one can say that the most important factor for the size of the dependency (first or second order) is the overlap of the features. Note that the way in which we measure overlap depends on distance between the features and also on size difference. One must note, however, that all the variables we used here improved the fit somewhat, so one can't say that overlap is the only factor to be considered. We also attempted to use additional parameters (more than four) in the model estimation, but could not achieve essentially better fits.

Furthermore, note that by using a fixed ICA basis, noise is not so prominent in the statistics as it was in our earlier analysis using PCA and mean value vectors. This is because one does not need to estimate the basis, and essentially only needs to estimate two parameters per component (mean and variance). One can use the data set $\mathbf{U}_{\alpha,i}$ in Equation 10 to estimate the amount of noise. Similarly as in Figure 9, we have plotted in Figure 15 the estimated amount of noise in the statistics. For the mean value we used the IC with the largest absolute value of the mean value, and not the length of the mean value vector. These values are lower than in Figure 9, but it is also noteworthy that highest values of variance have been dispersed among greater number of components, as one can see in the center and rightmost image in Figure 15. Actually, with PCA, the energy is by construction distributed among the smallest possible number of components. Note also that

in our earlier analysis, we looked at the length of the mean value vector (single dimension), so these analyses are not directly comparable.

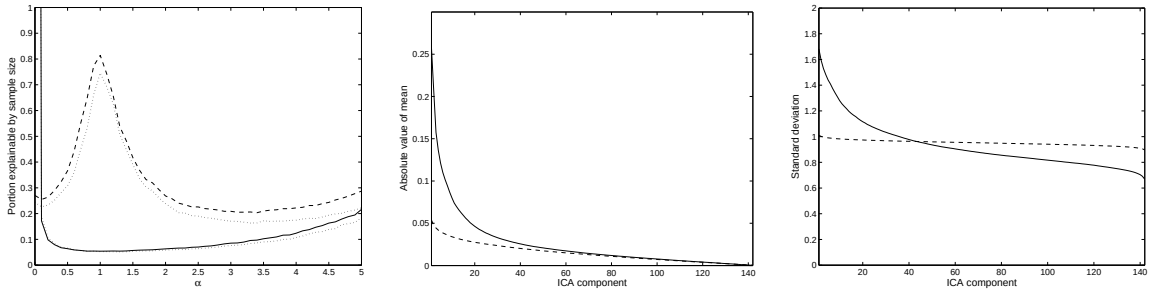

Figure 15: Our estimate of the portion of finite sample size in the observed ICA statistics. On the left: The average (estimated) portion of the observed statistics explainable by finite sample size. Compare to Figure 9. Center: Sorted average absolute values of the means of the ICs in the observed data and in the control data for $\alpha = 3$. Right: Sorted average standard deviations of the ICs in the observed data and in the control data for $\alpha = 3$. Compare to Figure 10.

5.3 ASSESSING VALIDITY OF THE MODEL STRUCTURE

In order to estimate the validity of our model, we also fitted a multilayer perceptron (MLP) network to the same variables. An MLP should be able to fit into the dependencies between the parameters, whereas in our model the parameters are essentially independent w.r.t. their contribution to the dependency, barring for the effect of f. We used Matlab's Neural Network Toolbox for creating and training the MLP. The input and target variables were the same as we used in our model.

For second order statistics, with five hidden layer neurons (and the same number of free parameters as in our model), we obtained a very similar error measure: 0.2625. But with an MLP it is harder to interpret the properties of the fitted model. With fifty hidden layer neurons, i.e. a total of 301 free parameters, we obtained an error of 0.200. This is sufficiently close to the error we obtained with our model with significantly less free parameters (and less chance of overfitting) for us to say that most of the information available in the four variables is captured by our model. When we added also the x- and y- coordinates of the Gabors and their explicit widths to the input parameters for the MLP, the error dropped with five neurons to 0.2179, with fifty neurons to 0.1142. In this case the model had 61 and 601 free parameters with five and fifty hidden layer neurons, respectively.

For first order statistics, with five hidden layer neurons, we obtained an error measure of 0.4580. With fifty hidden layer neurons, the error was 0.3492. Again, these are close enough to the error obtained with our model to say that most of the information available in the four variables was captured by our model. Adding the position for the x- and y-coordinates of the Gabors and their explicit widths to the input parameters produced an error of 0.2554 with fifty hidden layer neurons.

Note also that we can estimate the amount of noise (stochastic variation) in $R_{i,j}$ using $\mathbf{U}_{\alpha,i}$ in Equation (10), thus obtaining a lower bound for our fit without overfitting. This is as low as 0.0061 for second order statistics, even though our best fit with MLP was only 0.1142. For the estimation of the MLP network, and the model in Equation (12), we have the added difficulty of Gabor parameter estimation, and choosing the Gabor parameters for further use. This is quite significant for the results, as the ICA features are not perfectly Gabor functions. We can assume that this is for a large part responsible for the difference between the best error and our noise estimate. Also, our relatively dense sampling from a small number of images may be partially responsible for this difference. For first order statistics this lower bound for the fit was 0.0468, which is still significantly lower than our best fits.

6. EXTENDING FEATURES OUTSIDE PATCH BOUNDARIES

As a final note, the approach we have used here makes it easy to study how the features extend outside the patch edges. Extending the features outside the edges is still nontrivial, but the first order statistics can be easily calculated for bigger areas. As we know which patches belong to $I_{\alpha,i}$ and we know where they have been sampled from, we can pick bigger patches from the same positions, omitting the patches that are too close to the edges. Here we chose to sample patches of size 36 by 36 pixels, where the centermost 12 by 12 pixels form the original patch. Calculating the mean for these bigger patches (which have been normalized by the sign of the active component, as well as norm of the original patch after using the cumulative whitening matrix), we get the plots in Figure 16. Remember that the mean values are closest to the original features when $\alpha = 1$.

The noise is apparent in the smaller features for larger values of α, especially outside the original patch edges. The selection process has an innate tendency to suppress everything (other than the active component) inside the patch boundaries, but not outside them. As the number of samples increases, even the noise outside the patch boundaries should tend to zero. Note that there are some 11% less applicable patches than when the calculations were done for the original patches in Figure 5, due to the constraints set by the edges of the images in Figure 2.

It is also easy to see here (probably easier than in Figure 5) how the mean value features are larger than the original features for large values of α. The features are both wider along the edge and level off slower far from the zero crossing (orthogonal to the edge).

7. CONCLUSIONS AND FUTURE WORK

Here we studied the residual dependencies in the ICA model for image data by looking at what effect the activation of one feature has for the first and second order statistics of the data. It can be argued that first order statistics describe the static part of the changes brought by the activation of a component, and second order statistics most of the variable part.

We found that by normalizing by the sign of the active component, the mean values can be seen as extending the features into orthogonal dimensions in the whitened space. The mean value features for higher activations become wider and more step-function like.

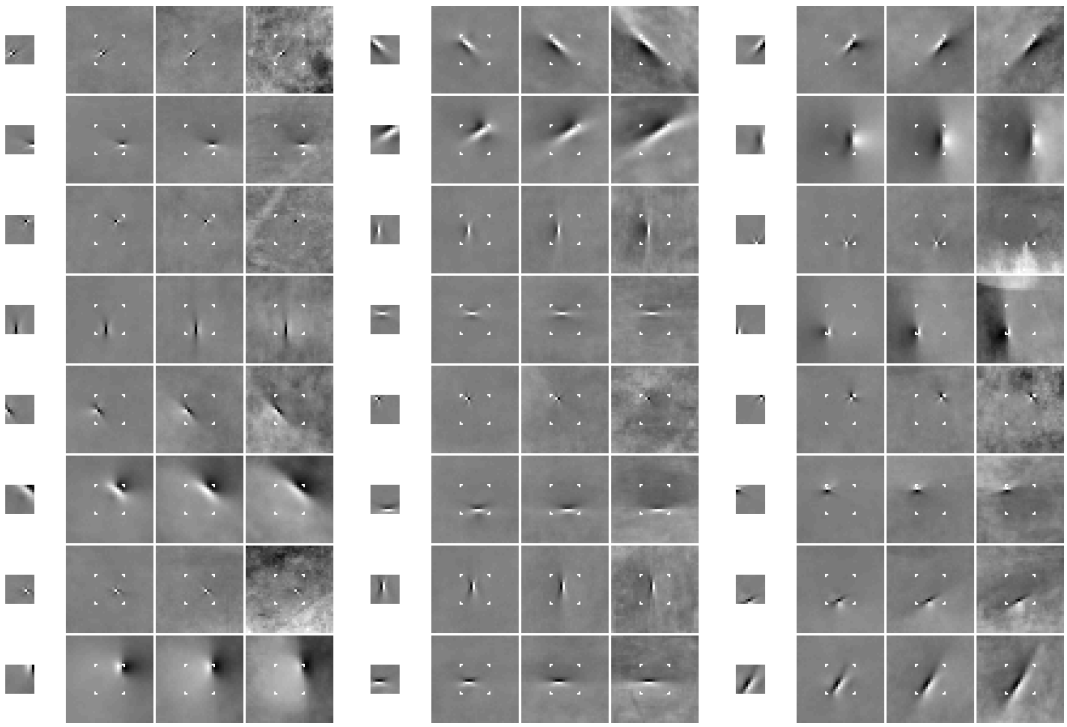

Figure 16: Mean values of the extended patches, when normalizing by the sign of the active component and using different thresholds. For each feature, from left to right: Original feature, $\alpha = 1$, $\alpha = 3$ and $\alpha = 5$. The area corresponding to the original feature has been explicitly marked for clarity.

Uncorrelated features which overlap the active feature and are possibly collinear were shown to have nonzero mean values.

Whereas first order statistics can be seen as indicating how the feature on average continues for a given level of activation, second order statistics can be argued to be linked to the possible misfits between the active feature and the actual activating object in the patches, especially to misfits that can occur in any direction, thus producing zero mean. Examples of these can be differences in orientation, position, and size. Overlapping (uncorrelated) features that possibly have similar orientations were shown to be active alongside the activating (conditioning) feature.

We presented a relatively comprehensive analysis of these statistics, but the quality of the results is partly diminished by the amount of data required for this analysis. Even with 200000 samples of 12 by 12 patches, noise is quite prominent in the estimates of the statistics. This problem was more aggravated with the smaller ICA features. One may note that simply storing the data required about 230 megabytes of memory in Matlab, so using significantly larger data sets for estimation of the dependencies is difficult. Increasing the patch size would increase memory requirements further. Also, in order to ensure that the

new patches are useful, one would need to have more images from which to sample them. Even for our current analysis, it could be useful to have more images to sample from.

Possibly the conceptually simplest way of extending the analysis here is to examine how the activation of two (or more) components changes the statistics. However, the number of different pairings and activation levels to examine would explode, so the implementation and interpretation of the analysis would be nontrivial.

One possibly useful alternative starting point for analyzing the dependencies would be to use a group of basis functions whose shape is mathematically defined and possesses the good qualities of an ICA basis, such as sparseness (high independence) and edge detecting properties. Gabor functions (or a certain subset of Gabor functions) are one such group of functions. Then the difficulty of parameter estimation from the basis functions would be avoided and the results could be more easily generalized. However, if these functions retained their edge detecting properties, they would generally no longer be orthogonal to each other in the whitened space in addition to not being optimally sparse. Exposing and analyzing the dependencies might require more work.

The results obtained here can work as a starting point for developing models for image analysis, compression and denoising, as a usable description of the dependencies between components should lead to better coding schemes. These results may also offer a small insight into the workings of biological visual systems. Especially the results for first order statistics can lead to better mechanisms for estimating how to 'complete' an image given certain parts of it. The method for extending the features outside the patch edges may prove valuable in achieving this.

Acknowledgments

This work has been supported by the Helsinki Graduate School in Computer Science and Engineering, the Finnish Foundation for Advancement of Technical Sciences and the Emil Aaltonen foundation. I would also like to thank Dr. Aapo Hyvärinen for his help with the manuscript.

References

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

R.W. Buccigrossi and E.P. Simoncelli. Image compression via joint statistical characterization in the wavelet domain. *IEEE Transactions on Image Processing*, 8(12):1688–1701, 1999.

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

J. G. Daugman. Complete disrete 2-D Gabor transforms by neural networks for image analysis and compression. *IEEE Trans. on Acoustics Speech and Signal Processing*, 36(7):1169–1179, 1988.

W. S. Geisler, J. S. Perrya, B. J. Superb, and D. P. Galloglya. Edge co-occurrence in natural images predicts contour grouping performance. *Vision Research*, 41(6):711–724, 2001.

P. O. Hoyer and A. Hyvärinen. Independent component analysis applied to feature extraction from colour and stereo images. *Network: Computation in Neural Systems*, 11(3):191–210, 2000.

A. Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans. on Neural Networks*, 10(3):626–634, 1999.

A. Hyvärinen and P. O. Hoyer. Emergence of phase and shift invariant features by decomposition of natural images into independent feature subspaces. *Neural Computation*, 12(7):1705–1720, 2000.

A. Hyvärinen, P. O. Hoyer, and M. Inki. Topographic independent component analysis. *Neural Computation*, 13(7), 2001a.

A. Hyvärinen and M. Inki. Estimating overcomplete independent component bases for image windows. *Journal of Mathematical Imaging and Vision*, 17(2):139–152, 2002.

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

M. Inki. Examining the dependencies between ICA features of image data. In *Proc. of ICANN/ICONIP 2003*, Istanbul, Turkey, 2003a.

M. Inki. ICA features of image data in one, two and three dimensions. In *Proc. Fourth International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003)*, Nara, Japan, 2003b.

Y. Karklin and M. S. Lewicki. Learning higher-order structures in natural images. *Network: Computation in Neural Systems*, 14:483–499, 2003.

T.-W. Lee, M.S. Lewicki, M. Girolami, and T.J. Sejnowski. Blind source separation of more sources than mixtures using overcomplete representations. *IEEE Signal Processing Letters*, 4(5), 1999.

M. Lewicki and T.J. Sejnowski. Learning overcomplete representations. *Neural Computation*, 12: 337–365, 2000.

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.

O. Schwartz and E. P. Simoncelli. Natural signal statistics and sensory gain control. *Nature Neuroscience*, 4(8):819–825, 2001.

M. Sigman, G. A. Cecchi, C. D. Gilbert, and M. O. Magnasco. On a common circle: Natural scenes and gestalt rules. *Proc. National Academy of Sciences (USA)*, 98(4):1935–1940, 2001.

D.G. Stork and H.R. Wilson. Do gabor functions provide appropriate descriptions of visual cortical receptive fields. *J. Opt. Soc. Am. A*, 7(8):1362–1373, August 1990.

J. H. van Hateren and D. L. Ruderman. Independent component analysis of natural image sequences yields spatiotemporal filters similar to simple cells in primary visual cortex. *Proc. Royal Society, Ser. B*, 265:2315–2320, 1998.

M. Wertheimer. *Laws of organization in perceptual forms*. Harcourt, Brace & Jovanovitch, London, 1938.