A hierarchical Bayesian model for learning non-linear statistical regularities in non-stationary natural signals

Yan Karklin
yan+@cs.cmu.edu

Michael S. Lewicki*
lewicki@cnbc.cmu.edu

Computer Science Department &
Center for the Neural Basis of Cognition
Carnegie Mellon University

*To whom correspondence should be addressed. Submitted to *Neural Computation*.

Abstract

Capturing statistical regularities in complex, high-dimensional data is an important problem in machine learning and signal processing. Models such as PCA and ICA make few assumptions about the structure in the data, have good scaling properties, but are limited to representing linear statistical regularities and assume that the distribution of the data is stationary. For many natural, complex signals, the latent variables often exhibit residual dependencies as well as non-stationary statistics. Here we present a hierarchical Bayesian model that is able to capture higher-order non-linear structure and represent non-stationary data distributions. The model is a generalization of ICA in which the basis function coefficients are no longer assumed to be independent; instead, the dependencies in their magnitudes are captured by a set of *density components*. Each density component describes a common pattern of deviation from the marginal density of the pattern ensemble; in different combinations, they can describe non-stationary distributions. Adapting the model to image or audio data yields a non-linear, distributed code for higher-order statistical regularities that reflect more abstract, invariant properties of the signal.

1 Introduction

The goal of many algorithms in machine learning, signal processing, and computational perception is to discover and process intrinsic structures in the data. Extracting these from real signals is a difficult problem, because often the relationships among the observable variables are complex, and there is little a priori knowledge about the types of structures that exist. When some a priori knowledge is available, specialized algorithms can be designed, but this approach is generally less desirable, as it places restrictions on the type of structure that can be learned. Another difficulty is that the dimensionality of the data is often very high, and properties of interest lie in a relatively low dimensional subspace. Because of the inherent variability of most real-world signals, intrinsic regularities are statistical in nature, which makes them that much more difficult to learn.

One approach to learning statistical regularities is to formulate a probabilistic model of how the data are generated, and adapt its parameters to fit the observed distribution. The adapted parameters reflect the statistics of the data ensemble, while internal representations encode individual data patterns. These models make minimal assumptions about the data and can result in more general representations than those in algorithms tailored for specific tasks or types of data.

There are several ways in which data patterns are represented in probabilistic generative models. Distributed representations of linear componential models, such as those for PCA and ICA, are particularly useful for modeling complex high-dimensional data because they can capture independent regularities with independent internal parameters (Bell and Sejnowski, 1995). This makes it possible to model a continuum of different

statistical relationships and allows scaling of the algorithms to large numbers of dimensions. Current models, however, are limited in the type of structure they can represent; in order to understand these limitations, it is helpful to look at their mathematical formulation.

Linear componential models achieve a distributed representation by describing the data as a combination of linear basis functions (for a review see Hyvärinen et al., 2001b; Cichocki and Amari, 2002). This yields a probabilistic generative model in which the data (**x**) are generated as a linear combination of basis functions (**A**) weighted by coefficients (**u**),

$$\mathbf{x} = \mathbf{A}\mathbf{u}\,. \tag{1}$$

The likelihood of the observed data under this model is

$$p(\mathbf{x}) = p(\mathbf{u})/|\det(\mathbf{A})| \tag{2}$$

(Pearlmutter and Parra, 1996; Cardoso, 1997), and the basis function matrix **A** is adapted to maximize the data likelihood. The coefficients **u** are the unknown (latent) variables. They are assumed to be independent and identically distributed,

$$p(\mathbf{u}) = \prod_i p(u_i)\,. \tag{3}$$

The priors $p(u_i)$ are typically chosen to be fixed sparse distributions (although parameters of the prior may be adjusted to maximize data likelihood). Because basis function coefficients are assumed to be independent and identically distributed, the dependence among the data is represented solely by the learned matrix of basis functions.

The obvious limitation of this model is that its inherent linearity restricts the type of structure it can capture. Even simple, low-dimensional data often exhibit statistical

dependencies that cannot be captured by linear transformations. In many applications, data are complex and rich with statistical structure, and latent variables of linear models adapted to these data exhibit significant residual mutual dependence (Hyvärinen and Hoyer, 2000; Schwartz and Simoncelli, 2001; Karklin and Lewicki, 2003).

Another shortcoming of these models is that they assume that the statistical regularities in the data do not change, i.e. they describe stationary probability distributions. For example, once model parameters are adapted in ICA, both the prior and the basis functions are fixed, leading to a stationary distribution over the data. This does not depend on the form of the prior, and also applies to models with adaptive or entirely non-parametric priors. In many domains, however, the statistics of the data are known to change, as the physical properties of the environment or conditions for data acquisition vary. While the stationary prior assumption gives a valid approximation of true density over a large enough corpus of training data, it does not reflect the variation across contexts that is observed in many signals.

Figure 1 illustrates non-stationary statistics observed in images of natural scenes. ICA basis functions were adapted to 20×20 patches taken from an ensemble of natural images. Over the full ensemble of the training data, the basis function coefficients have marginal distributions that are consistent with the prior assumed by the model (not shown). However, computing coefficient histograms over particular image regions reveals systematic deviations from the (globally valid) stationary distribution. Patterns in the histograms suggest that basis functions of certain orientations are more active in some parts of the image (e.g. textured, oriented surface of the log) while in other regions different subsets tend to be activated. This is observed in other types of data as well: temporal basis functions adapted to speech also yield coefficients whose statistics

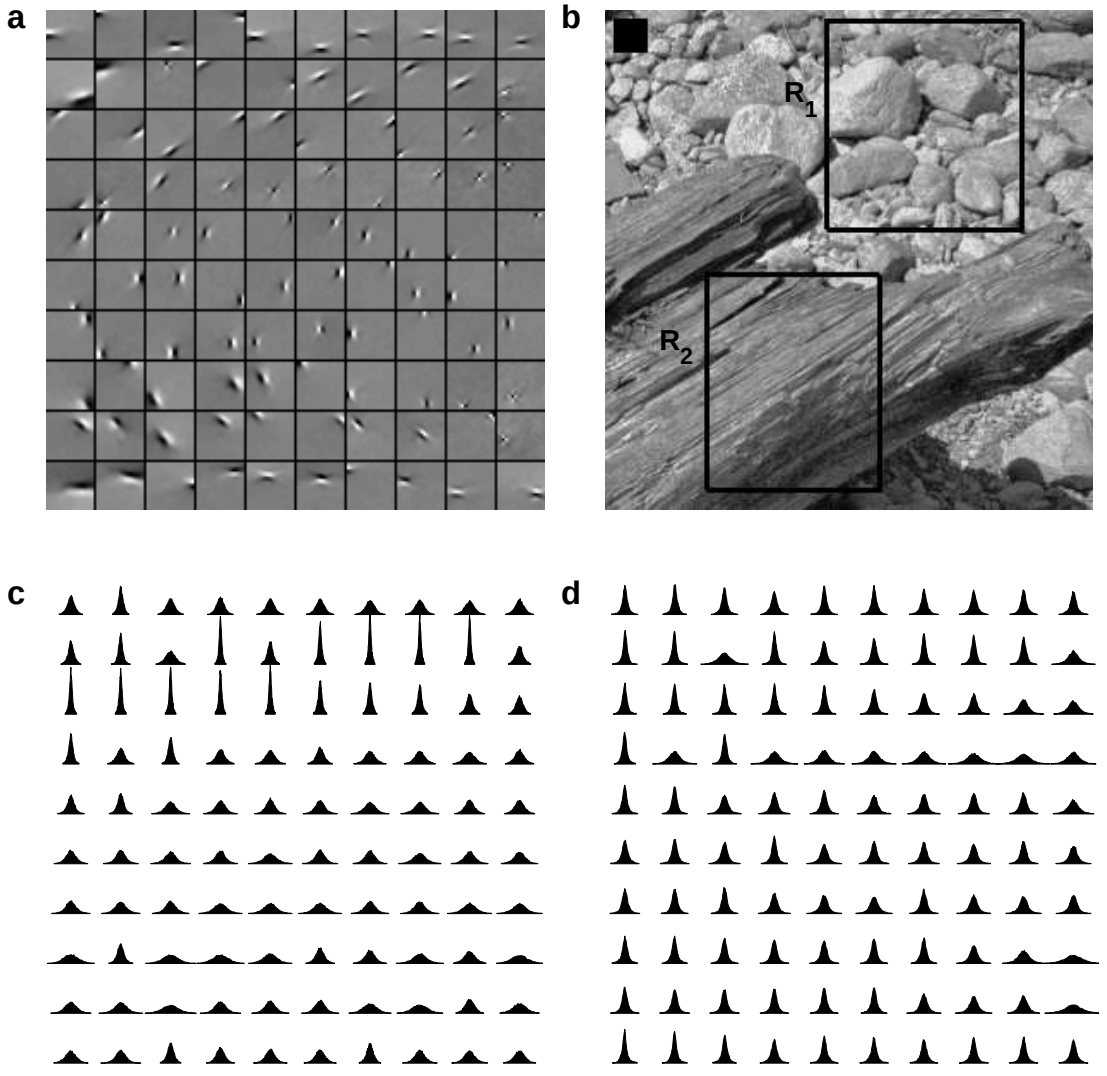

a
b
R1
R2
c
d

Figure 1: (*previous page*) The distribution of ICA basis function coefficients exhibits non-stationary statistics that reflect local image structure. (a) A subset of image basis functions learned from an ensemble of natural images, ordered by orientation. The small black square on the image indicates the size, relative to the image, of the learned basis functions. (b) Coefficients of independent components were computed over two regions of an image; (c,d) histograms of the coefficients for the two regions reveal patterns in the joint distributions. Each histogram in the 10×10 grid in (c) and (d) corresponds to a basis function at the same grid position in (a), and is normalized so that the filled area sums to 1. Different types of local image structure produce different patterns in the joint activities. For example, the image region containing the log yields higher coefficient variation for basis functions oriented along the grain and matching the approximate spatial frequency of the wood texture.

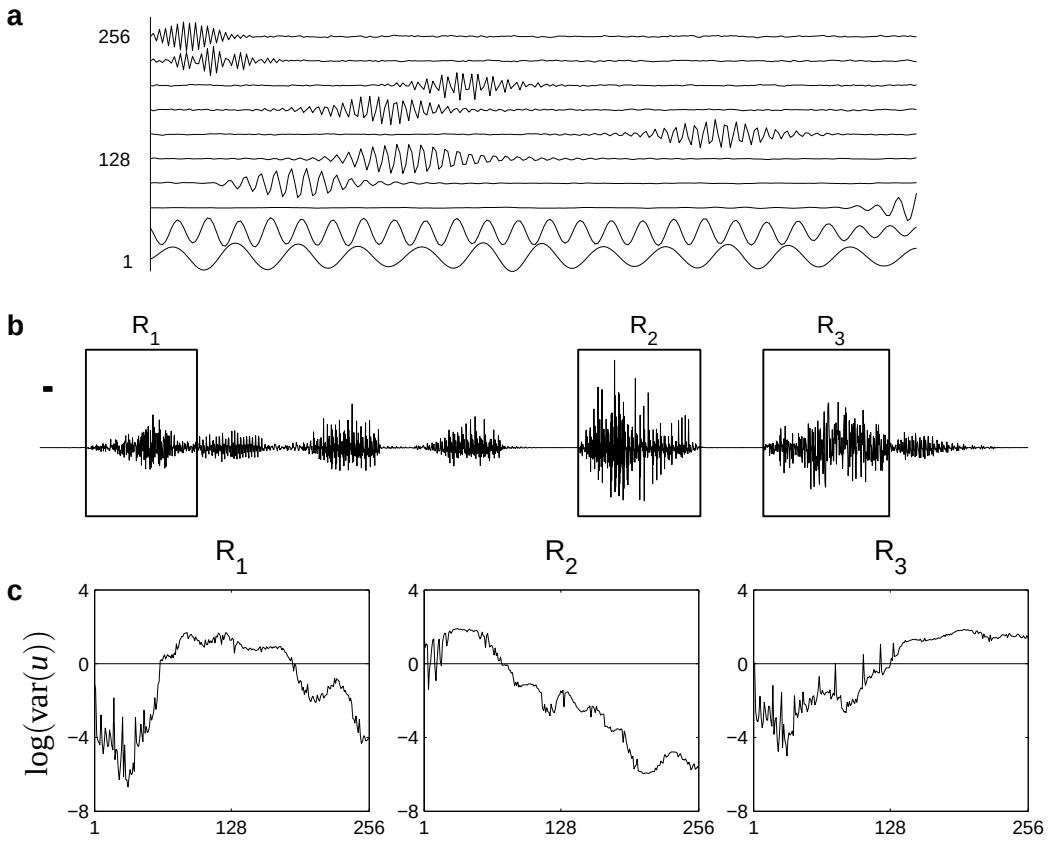


Figure 2: ICA basis functions adapted to speech data also exhibit non-stationary statistical dependencies. (a) A subset of 256 ICA-derived basis functions ordered by dominant frequency. (b) Each basis function was convolved with three different regions of a speech signal. The length of the basis function is indicated by the short bar above the start of the speech signal. (c) The variances of coefficients sampled over the three regions, with the 256 coefficients ordered as by frequency as in (a). Although all basis function coefficients have unit variance when sampled over the whole data ensemble, local regions show characteristic variance patterns that reflect local signal structure.

vary greatly across local regions of the signal (Figure 2).

Figures 1 and 2 give just a few examples of patterns in latent variable distributions that depend on the local context. In fact, there is a wide range of statistical regularities in complex data, a continuum of contexts that is as multi-dimensional as the physical properties of the environment that give rise to it. Local representations, as employed by clustering or mixture model techniques, assume that the contexts are discrete and thus cannot describe regularities that arise from a combination of different contexts. A model that captures this variation must form flexible, distributed representations of higher-order structure. Moreover, because the dimensions of the contexts are not known a priori, the model must be able to automatically discover this underlying structure. Finally, many previous models of non-stationary distributions have relied on the assumption that data statistics vary smoothly from sample to sample (Everson and Roberts, 1999; Pham and Cardoso, 2001) and computed local estimates of context-dependent variation. This assumption does not always hold; even spatially and temporally coherent data exhibit abrupt changes that cannot be modeled as slowly evolving processes.

Here we address the limitations of previous models with a hierarchical Bayesian model that forms a distributed code of higher-order statistical regularities and captures non-stationarities in the data distribution. The model is a generalization of ICA, thus we begin with a standard linear componential model in which the data are generated as a combination of linear basis functions. However, instead of assuming that the basis function coefficients are independent (and their joint prior distribution is factorable, equation 3), we explicitly model the dependence among hyperparameters of their priors. In order to capture variable, context-dependent activation of basis functions, the dependence is specified through the scale parameters governing the width of the prior

(and hence the variance of the coefficients). This dependence is modeled with a set of *density components*, a distributed code that describes the shape of the joint density of the linear coefficients and captures patterns in the variances of the coefficients as observed in the motivating examples.

Each density component describes a common underlying deviation from the standard assumption of independence (the i.i.d. joint prior) associated with a frequently encountered context. Using a weighted combination of density components, the model is able to represent a continuum of context-dependent changes in probability distributions. Adapting the set of density components and modeling their activation with a sparse prior yields a compact description of higher-order statistical regularities of the data ensemble. Unlike other recent methods, the model makes no assumptions of temporal or spatial coherence; it is able to infer, independently for each data sample, the higher-order code that describes the generating distribution.

Below we present the probabilistic framework for the model and describe the associated learning algorithms. Previously, we have used this model to discover higher-order structure in natural images (Karklin and Lewicki, 2003). Here, we describe the algorithm in more detail and frame it as a general method of statistical density estimation for high-dimensional non-stationary data. We verify the recovery of correct model parameters using a toy dataset, apply the learning algorithm to a wider range of data types, and show how the learned higher-order code accounts for observed dependencies. Below we provide results and analysis for photographs of natural scenes, scanned images of newspapers, and speech waveforms. However, the model is not tailored specifically to images or audio data, and can be used to automatically learn the non-linear statistical dependencies in any dataset with sufficiently rich structure.

2 A Hierarchical Model for Non-Stationary Distributions

Our model is a generalization of previous linear models, hence we begin by assuming that each data vector is generated as a combination of linear basis functions, $\mathbf{x} = \mathbf{A}\mathbf{u}$. As in standard ICA models (e.g. Cichocki and Amari, 2002), basis function coefficients are assumed to be sparsely distributed; here we use a generalized Gaussian distribution with zero mean:

$$p(u_i) = \mathcal{N}(0, \lambda_i, q_i) \quad (4)$$

$$= z_i \exp\left(-\left|\frac{u_i}{\lambda_i}\right|^{q_i}\right), \quad (5)$$

where $z_i = q_i/(2\lambda_i\Gamma[1/q_i])$ is a normalizing constant. The parameter q_i determines the weight of the distribution's tails, and can be estimated from the data; in many ICA applications the coefficients tend to be sparse, making their distributions supergaussian ($q_i < 2$). Typically, the scale parameter λ_i is fixed to a constant, since the basis functions in $\mathbf{A}$ can themselves scale to fit the data.

In order to capture residual dependence among coefficients $\mathbf{u}$, we must abandon the assumption of fixed, independent priors. The motivating examples suggested that intrinsic structures in the data give rise to patterns in the scales of the coefficients (similar dependencies have been observed previously in wavelet coefficients, Simoncelli, 1997). A natural way to model this is through the scale parameters of the prior, which we model as a non-linear transformation of latent higher-order variables. Specifically, we use a matrix of *density components* $\mathbf{B}$ and density component coefficients $\mathbf{v}$ to de-

scribe the logarithm of the scale parameter,

$$\log(\boldsymbol{\lambda}/c) = \mathbf{B}\mathbf{v}. \tag{6}$$

If we define the constant $c = \sqrt{\Gamma(1/q)/\Gamma(3/q)}$, the variance of the coefficients becomes 1 when the right side of the equation is 0 (this becomes convenient when a zero-centered prior is selected for the distribution of $\mathbf{v}$; see below).

The joint prior distribution of coefficients $\mathbf{u}$ can now be expressed as

$$-\log p(\mathbf{u}|\mathbf{B},\mathbf{v}) \propto \sum_i [\mathbf{B}\mathbf{v}]_i + \left| \frac{u_i}{c\exp([\mathbf{B}\mathbf{v}]_i)} \right|^{q_i}, \tag{7}$$

where $[\mathbf{B}\mathbf{v}]_i$ represents the i^{th} element of the vector $\mathbf{B}\mathbf{v}$ (see Appendix for the derivation).

Basis function coefficients are assumed to be independent *conditional* on the higher-order variables, $p(\mathbf{u}|\mathbf{v}) = \prod p(u_i|\mathbf{v})$. This accounts for the dependence in the magnitudes of basis function coefficients. The new form of the prior (7) implies that if $\mathbf{v}$ is 0, the model reduces to standard ICA in which the linear coefficients are independent and identically distributed with variance equal to 1. Non-zero values of $\mathbf{v}$ scale and combine density components (columns of $\mathbf{B}$) that define patterns in the distributions of $\mathbf{u}$. Because each v_i can be positive or negative, each density component represents contrast in the magnitudes of coefficients $\mathbf{u}$ (Figure 3).

We place a non-Gaussian, sparse prior on the latent variables $\mathbf{v}$ and infer their values for each data sample.[1] This means that *a priori* we assume that the activity of density component coefficients is sparse and relatively few components are needed to describe how the generating distribution associated with each data sample differs from the i.i.d.

[1] A Laplacian prior was used in the simulations, but other distributions may be more appropriate.

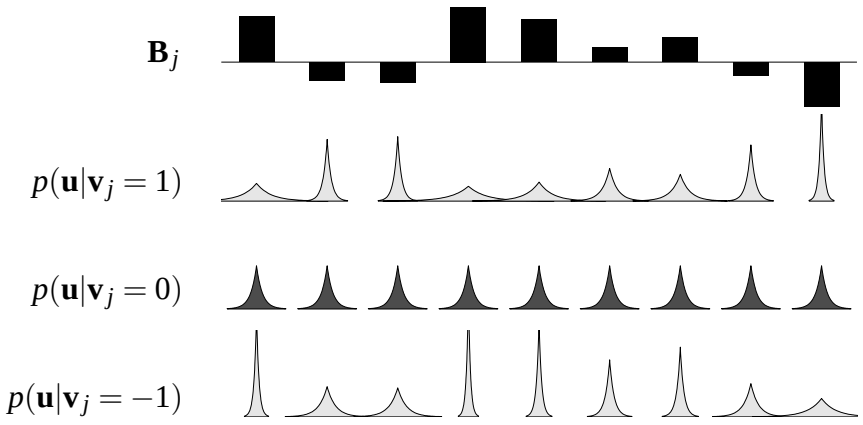
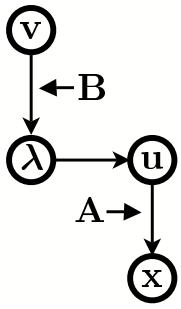


Figure 3: Each density component defines a pattern in the joint distribution $p(\mathbf{u})$. The plot at the top shows an example 9-dimensional density component $\mathbf{B}_j$, The distributions of coefficients $u_{1...9}$ are shown for different values of v_j. Here we show only a single density component $\mathbf{B}_j$, whereas the model adapts a set of them $\mathbf{B} = \{\mathbf{B}_1, \mathbf{B}_2, \ldots, \mathbf{B}_M\}$ to obtain a compact description for common scale patterns in the data.



$$
\begin{aligned}
v_i &\sim \mathcal{N}(0,1,q_i) \\
\lambda_j &= c\exp^{[\mathbf{Bv}]_j} \\
u_j|\lambda_j &\sim \mathcal{N}(0,\lambda_j,q_j) \\
x_k &= \sum_j \mathbf{A}_{kj}u_j
\end{aligned}
$$

Figure 4: A schematic of the hierarchical generative model. Sparsely distributed random variables **v** specify (through a non-linear transformation) the scale hyperparameters $\boldsymbol{\lambda}$ for the distribution of coefficients **u**. The data **x** are a linear combination of coefficients **u**. Matrices **A** and **B** are parameters that are adapted to the statistical distribution of the data.

ICA model. Using this parameterization, we adapt the density components to the entire data ensemble, which produces a compact description of higher-order statistical regularities.

The full generative model is shown in graphical form in Figure 4. There are two sets of random variables that give rise to the data, **v** and **u**, and two sets of parameters adapted to the data, the linear basis functions **A** and the density components **B**. A crucial difference between this generative form and several other models that account for higher-order dependence is that here, the density components specify a *distribution* over the coefficients, as opposed to exact values or pooled magnitudes, which have been used in other models (Hoyer and Hyvärinen, 2002; Welling et al., 2003). Thus, the model forms a hierarchical representation in which the lower level codes data values precisely and the higher level represents more abstract properties associated with the

shape of the data distribution.

3 Inference of Density Component Coefficients

For each data sample, it is necessary to compute the higher-order representation **v** that best describes the pattern in the scale of coefficients **u**. This transformation is non-linear, and cannot be expressed in closed form. Here, we compute the best value of **v** by maximizing the posterior distribution

$$\hat{\mathbf{v}} = \arg\max_{\mathbf{v}} p(\mathbf{v}|\mathbf{u},\mathbf{B}), \tag{8}$$

$$= \arg\max_{\mathbf{v}} p(\mathbf{u}|\mathbf{B},\mathbf{v})p(\mathbf{v}). \tag{9}$$

We assume that v_i's are independent ($p(\mathbf{v}) = \prod_i p(v_i)$) and sparsely distributed ($\log p(v_i) \propto -|v_i|$). For the simulations below, $\hat{\mathbf{v}}$ was derived by gradient ascent. We used second order methods (LeCun et al., 1998) to stabilize and speed up convergence to optimal estimates.

Because the prior is zero-centered and sparse, only a few non-zero values will contribute to the representation of each data sample. The inference of optimal density component coefficients is analogous to estimating sample variance based on single observations, but the problem is further constrained by the structure of the learned density components. Because the model is constrained to describe the pattern of variance with a sparse combination of density components, the value **v** for a typical pattern is usuall well-determined. In addition, the high dimensionality of the input facilitates the inference process, as it provides more directions of variation that make up the variance pattern.

4 Adapting Model Parameters to the Data

The linear basis functions and the density components are adapted to the data ensemble by maximizing the posterior $p(\mathbf{A},\mathbf{B}|\mathbf{X})$. We assume that samples in the data ensemble $\mathbf{X} = \{\mathbf{x}1,\ldots,\mathbf{x}_N\}$ are independent, so that

$$p(\mathbf{X}|\mathbf{A},\mathbf{B}) = \prod_{n=1}^{N} p(\mathbf{x}_n|\mathbf{A},\mathbf{B})\,. \tag{10}$$

For each data sample $\mathbf{x}$ the posterior distribution is

$$\begin{aligned} p(\mathbf{A},\mathbf{B}|\mathbf{x}) &\propto p(\mathbf{x}|\mathbf{A},\mathbf{B})p(\mathbf{A},\mathbf{B}) & (11)\\ &= p(\mathbf{u}|\mathbf{B})p(\mathbf{B})/|\det(\mathbf{A})|\,. & (12)\end{aligned}$$

Ideally, the marginal distribution $p(\mathbf{u}|\mathbf{B})$ would be computed by integrating over $\mathbf{v}$, but evaluating this integral for equation (7) is intractable. Here we approximate it using the maximum a posteriori estimate $\hat{\mathbf{v}}$:

$$\begin{aligned} p(\mathbf{u}|\mathbf{B}) &= \int p(\mathbf{u}|\mathbf{B},\mathbf{v})p(\mathbf{v})d\mathbf{v}\,, & (13)\\ &\approx p(\mathbf{u}|\mathbf{B},\hat{\mathbf{v}})p(\hat{\mathbf{v}})\,. & (14)\end{aligned}$$

Substituting this approximation into the posterior gives

$$p(\mathbf{A},\mathbf{B}|\mathbf{x}) \propto p(\mathbf{u}|\mathbf{B},\hat{\mathbf{v}})p(\hat{\mathbf{v}})p(\mathbf{B})/|\det\mathbf{A}|\,. \tag{15}$$

The prior on $\mathbf{B}$ places a small *a priori* bias for small values of $B_{i,j}$ and eliminates the problem of a degenerate case in which $\mathbf{B}$ grows without bounds while $\mathbf{v}$'s rescale to be smaller. For the results here, we assumed $B_{i,j}$ followed a Gaussian distribution. The matrices $\mathbf{A}$ and $\mathbf{B}$ can be optimized iteratively, by maximizing $p(\mathbf{A}|\mathbf{X},\mathbf{B})$, then maximizing $p(\mathbf{B}|\mathbf{X},\mathbf{A})$. In this case, the first step amounts to performing ICA in which

the priors incorporate the scale estimates $\hat{\mathbf{v}}$. Alternatively, we can assume that optimal linear basis functions are largely independent of the set of density components, and optimize $\mathbf{B}$ using a fixed $\mathbf{A}$. For computational efficiency, $\mathbf{A}$ and $\mathbf{B}$ were assumed to be independent and were adapted separately in the simulations described below. We verified the validity of this approach by training a model on data of reduced dimensionality and with fewer density components; results were qualitatively similar to optimizing the parameters independently.

In order to verify that the learning algorithm produces a valid solution, we adapted model parameters to an artificial dataset for which the optimal solution was known. The data were generated by constructing a set of density components and then sampling basis function coefficients according to $p(\mathbf{u}|\mathbf{B})$. An illustration of the process and the obtained results is shown in Figure 5. Optimizing density components from random initial values produced a matrix that was identical (up to a permutation of its columns) to the true model parameters (Figure 5a,b). The patterns in the learned density components specify non-linear dependencies among coefficient magnitudes; in fact, there are no linear correlations among basis function coefficients sampled from the model (even when the same $\mathbf{v}$ is used to generate the coefficients). Linear models like ICA are unable to recover these statistical regularities.

As a control, we adapted the density component model to a pure noise dataset in which coefficients $\mathbf{u}$ were random samples from independent sparse distributions. In this case, no regularities in the magnitudes of coefficients existed, and the resulting density components consisted of small, random values.

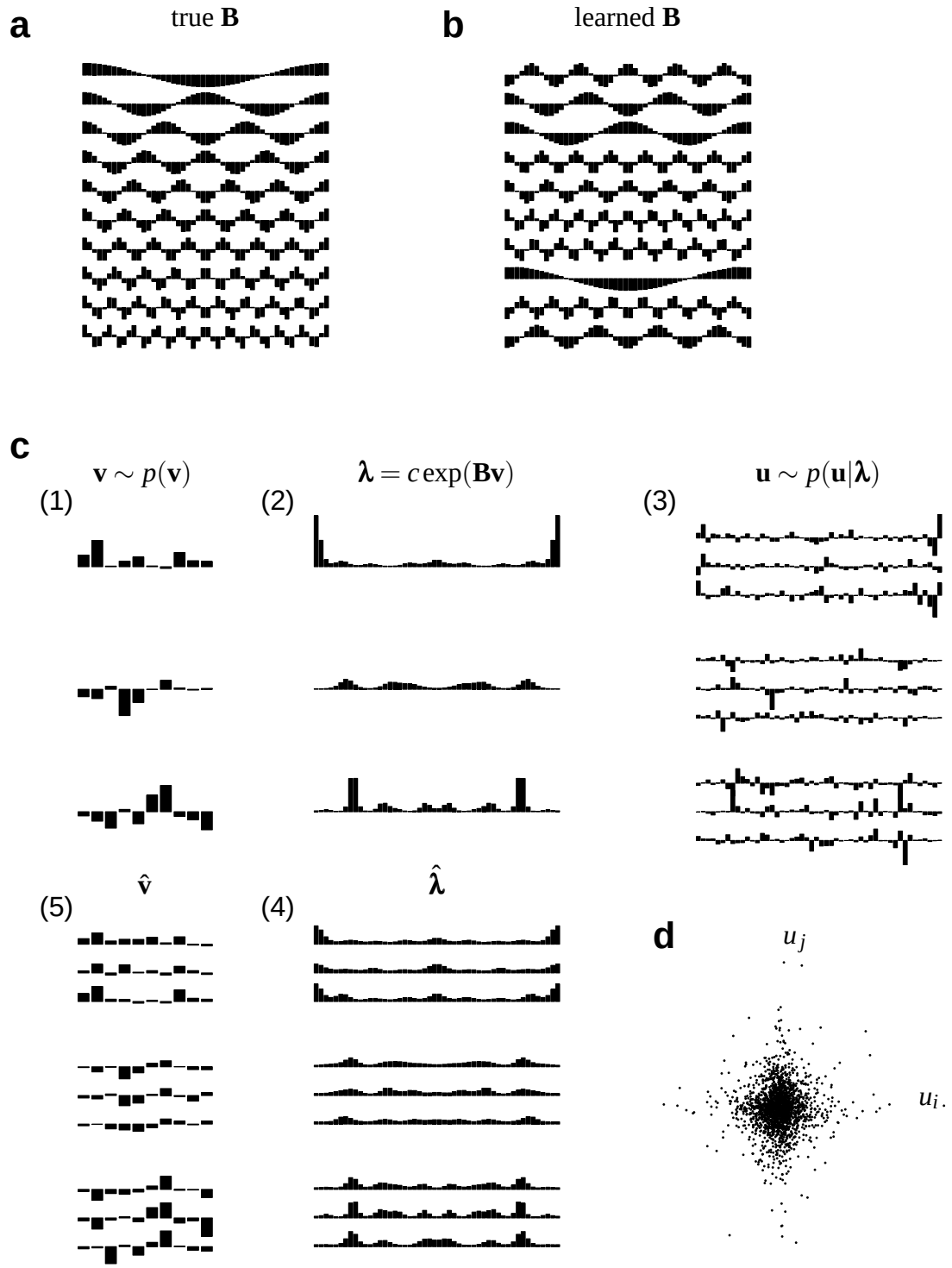

a
true B
b
learned B
c
v ~ p(v)
λ = c exp(Bv)
u ~ p(u|λ)
(1)
(2)
(3)
v̂
λ̂
(5)
(4)
d
u_j
u_i

Figure 5: The model correctly recovers the density components used to generate synthetic data. We constructed a 50×10 matrix $\mathbf{B}$ composed of 10 cosine-shaped density components (a). After 3000 iterations, the model recovers (up to a permutation) the correct density components (b). In (c) we illustrate the generative and inference steps of the algorithm. 1) Three 10-dimensional density component coefficients are drawn from a sparse distribution; 2) each $\mathbf{v}^{(i)}$ specifies a vector of scaling variables $\boldsymbol{\lambda}^{(i)}$ through the nonlinear transformation $\boldsymbol{\lambda}^{(i)} = c\exp(\mathbf{B}\mathbf{v}^{(i)})$. 3) The scaling variables are hyperparameters for non-stationary distributions $p(\mathbf{u})$, from which data samples $\mathbf{u}$ are drawn. In order to emphasize that each vector of scaling variables $\boldsymbol{\lambda}^{(i)}$ specifies a distribution, not fixed values of $\mathbf{u}$, we plotted several $\mathbf{u}$'s drawn from the distribution $p(\mathbf{u}|\boldsymbol{\lambda}^{(i)})$; in actual simulation each data point was generated independently. Using the learned density components estimates of 4) $\hat{\mathbf{v}}$ and 5) $\hat{\boldsymbol{\lambda}}$ were obtained for each data sample. Because the inference problem involves the estimation of density parameters from single data points, $\hat{\mathbf{v}}$ and $\hat{\boldsymbol{\lambda}}$ only approximately match true parameters. Although the complete hierarchical model includes another transformation $\mathbf{x} = \mathbf{A}\mathbf{u}$, the projection to data space $\mathbf{x}$ is linear and is not necessary for inference of $\hat{\mathbf{v}}$ when coefficients $\mathbf{u}$ are known. The scatter plot of 1000 samples of u_1 and u_2 drawn from the model (d) shows that there is no linear dependence among basis function coefficients.

5 Discovering Structure in Complex Data

5.1 Learned Density Components

We optimized model parameters on several datasets and analyzed the learned density components. For computational simplicity, in all the simulations the model was optimized in two stages. First a complete linear basis $\mathbf{A}$ was adapted to the data using standard methods; next, the density component matrix $\mathbf{B}$ was optimized on the coefficients of the fixed $\mathbf{A}$. Since the linear basis functions were learned using standard ICA methods, our analysis and discussion here is limited to the recovered matrix of density components. The density components were initialized to small random values and gradient ascent was performed on stochastically sampled batches of data. The MAP estimate $\hat{\mathbf{v}}$ was obtained using 20 steps of gradient ascent. Convergence of the gradient procedures for the optimization of $\mathbf{B}$ and estimation of $\hat{\mathbf{v}}$ was tested in a number of ways, including varying the step size, the number of iterations, and the initial conditions. The given optimization parameters yielded reasonable speed and accuracy as well as consistent solutions for different random initial conditions.

We first applied the learning algorithm to small (20×20) image patches sampled from a standard set of ten grayscale images of natural scenes (Olshausen and Field, 1996; Karklin and Lewicki, 2003). We used a complete set of 400 linear basis functions. The number of density components was set to 100 (although the algorithm is able to recover any number that yield a sparse distribution for coefficients $\mathbf{v}$). We used batches of 1000 samples for 35000 iterations of gradient ascent with a fixed step size of 0.3.

Statistical regularities of the data ensemble are captured in the matrix of density components. In order to analyze the structure described by this matrix, we need to

examine its weights as they relate to the basis functions whose distributions they affect. (Recall that each weight in a density component vector specifies how a particular $p(u_i)$ is rescaled). The initial ordering of basis functions in the learned matrix **A** is arbitrary, hence weights in **B** also appear random in their original ordering. However, we can rearrange the weights in **B** according to some property of the linear basis functions and examine whether the learned density components capture structure related to the chosen property. For example, ICA basis functions adapted to natural images are spatially localized; arranging density component weights according to the location of corresponding basis functions within the image patch reveals patterns in their organization (Figure 6). Thus, density components that appear structured in this arrangement specify dependence among *spatially related* linear basis functions. As parameters in the generative model, they describe common data distributions that reflect localized image structure. Some density components also appear random when arranged spatially, but these often show organization along other dimensions of the lower-order representation, such as orientation or spatial frequency (Karklin and Lewicki, 2003). Changing the number of density components does not affect the type of structure captured by the hierarchical model. A larger number of density components allows the model to represent more fine scale spatial regularities, as well as other statistical structure that is not as obvious to interpret.

We also applied the model to speech data from the TIMIT database. Linear basis functions were adapted to band-pass filtered speech segments of 256 samples (16 msec of 16kHz sound). The number of density components was set to 100, and the parameters were optimized using stochastic learning on data batches of 1000 for 10000 iterations. A representative set of the learned density components is shown in Figure

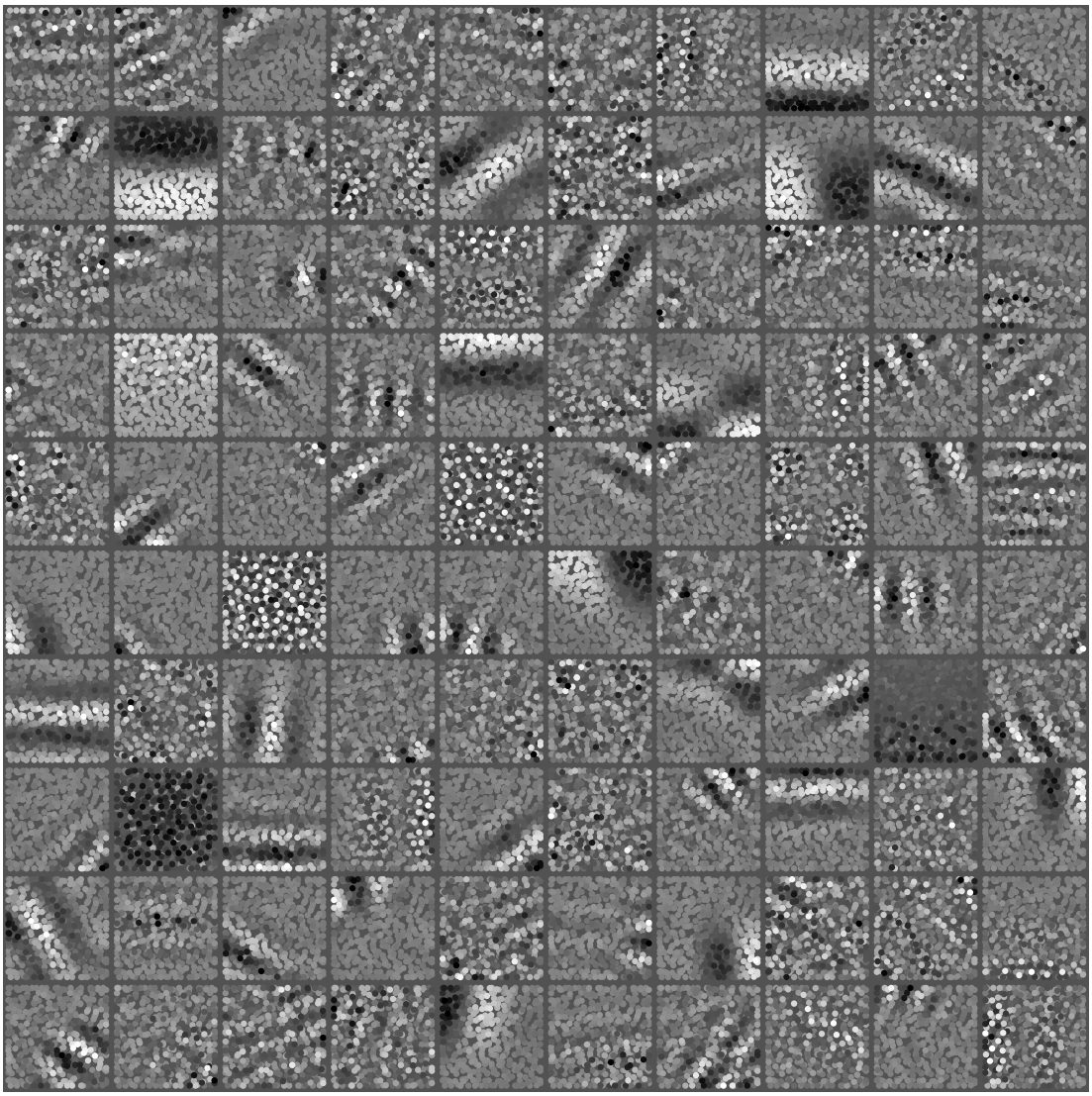


Figure 6: (*previous page*) Density components optimized on an ensemble of 20×20 image patches drawn from natural scenes. Each column of $\mathbf{B}$ is represented here as a square; its weights to 400 image basis functions are plotted as dots, placed in locations corresponding to the center of each image basis function in the image patch. Each dot is colored according to the value of the weight, with white indicating positive weights, black negative weights, and gray weights that are close to zero. Most density components describe spatial relationships and capture co-activation of linear basis functions localized to a particular area of the image patch. For example, the density component in the second row, second column describes whether contrast in the image patch is localized to the top or the bottom half. While most density components represent location, orientation, or spatial frequency regularities, the organization of some is not obvious.

7. In order to display the weights in the density components as they relate to the linear code, we first computed the Wigner distributions (WD) of the linear basis functions using the DiscreteTFDs Matlab package (O'Neill, 1999). The Wigner distribution of a basis function is a surface in the time-frequency space; we took a contour at 95% peak value for each basis function and drew all these contours on a single time-frequency plot (time on the horizontal axis, 0 to 16msec, and frequency on the vertical axis, 0 to 8kHz). Because the linear basis functions adapted to speech tile most of the the time-frequency space, the contours also exhibit relatively even tiling of the plots. In Figure 7, nine WD plots show the weights in nine density components to the same set of linear basis functions. Here, as in image density components, the shading of each patch corresponds to the value of the weight. Some density components describe coactivation of

linear basis functions of adjacent frequency bands, while others are localized in time within the sample window. Most density components capture periodic higher-order structure and regularities across multiple frequencies or time intervals, and a few are tuned specifically to subtle shifts in dominant frequency over the sample window.

5.2 Higher-order Code

In order to better understand the type of structure captured by the model, it is informative to look at the higher-order code – the coefficients of density components – and the statistical regularities it represents. Individual density component coefficients indicate the presence, in each data sample, of the type of structure represented in Figure 6. As a distributed code, their joint activity describes the data density whose shape reflects underlying structure in the data.

Figure 6 shows that among other statistical regularities, the higher-order code captures spatial relationships in the data. How does this representation compare to the lower-level, linear code for image structure? The activity of density component coefficients over contiguous regions of the data suggests that the higher-order representation captures more abstract properties of the data (Figure 8). When a sliding window is applied to a natural scene image, the resulting lower level representation changes rapidly from sample to sample – as would be expected from what are essentially outputs of linear filters. On the other hand, the higher-order representation varies more slowly over the image and captures more invariant properties of the data, such as overall image contrast or the dominance of certain spatial frequencies. Also shown in Figure 8 are the values of the linear and the density component coefficients for a model trained on images of newspaper text. Here too, the density component coefficients describe

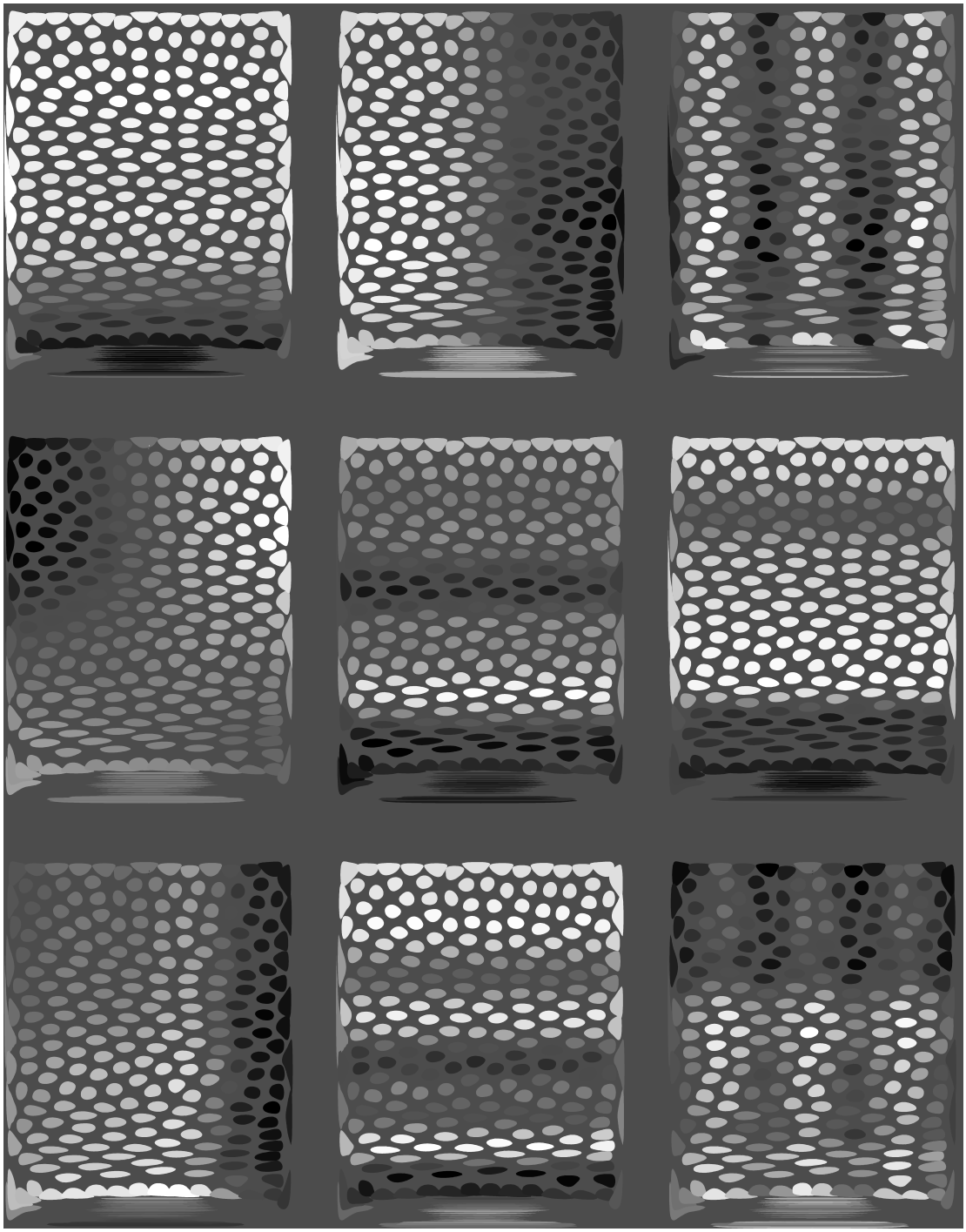


Figure 7: (*previous page*) A subset of density components of speech. The weights in a column of **B** are plotted as shaded patches in one of the nine panels. Each patch is placed according to the temporal and frequency distribution of the associated linear basis function and shaded according to the value of the weight, with white indicating positive weights, black negative weights, and gray weights that are close to zero. The axes represent time, 0 to 16msec, horizontally, and frequency, 0 to 8kHz, vertically. The density components form a distributed representation of the frequency of the signal and the location of energy within the sample window. Density components coding for multiple frequencies might capture harmonic regularities in the speech signal (see text for details).

more abstract properties: several combine to form a distributed representation of text line position in the image patch (the activity of one such coefficient is shown in the first panel of Figure 8f), while others represent commonly observed structures in the data, such as recurring shapes of letters or blank spaces between words.

Applied to audio data, the model also captures more abstract properties of the stimulus. In Figure 9, we plot an example audio signal (a), along with the activities of three linear coefficients (b) and three density component coefficients (c). We emphasize that, as for the images, the model is trained on segments drawn randomly from the dataset, and the values of the coefficients for each sample position in the signal shown in the figure are determined independently. The higher-order representation varies more slowly than responses of the linear filters and captures structural elements that extend well beyond the small sampling window. This may reflect a general property of natural signals – fast fluctuations in their exact values are caused by interactions of underlying physical

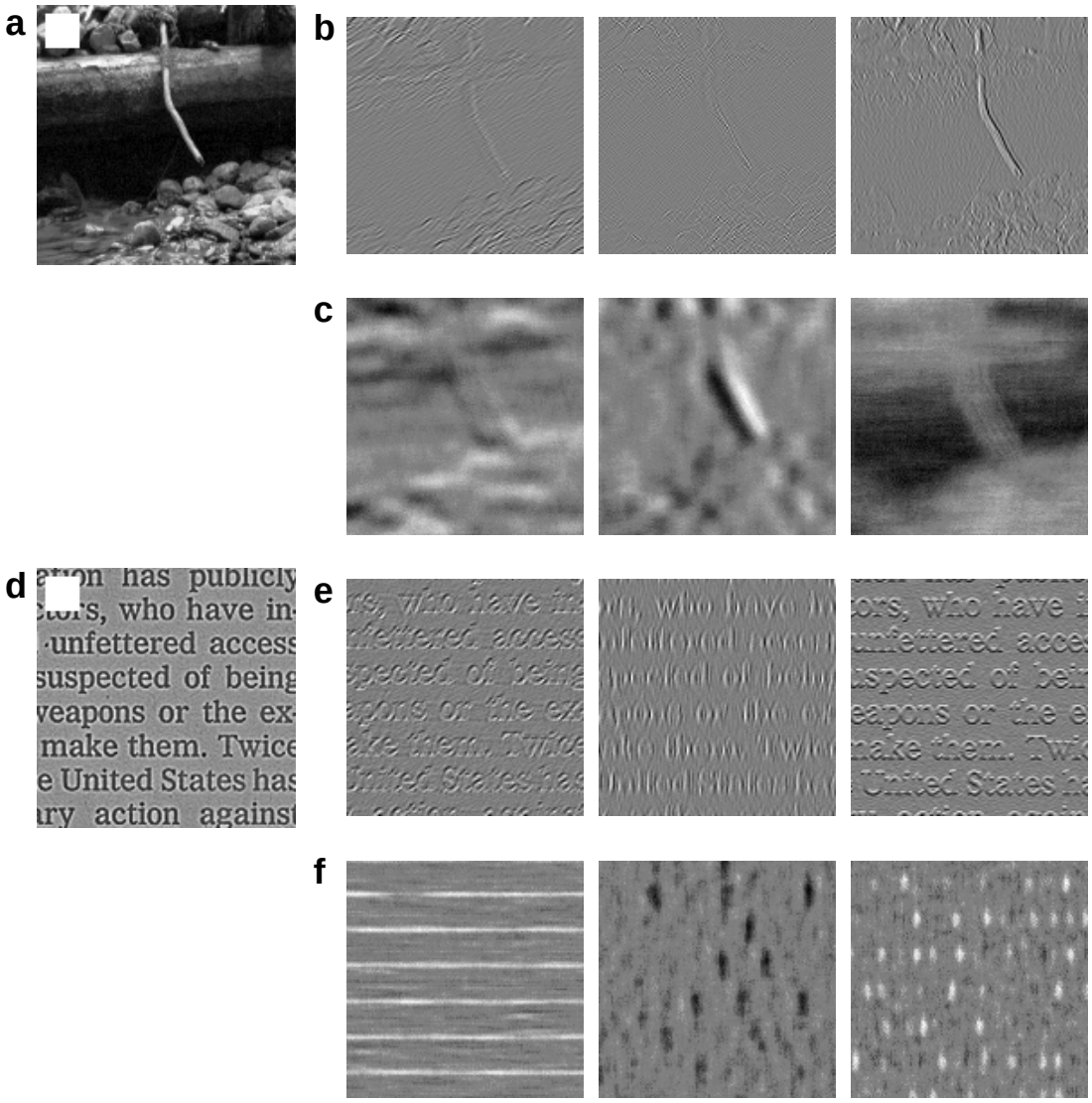

a
b
c
d
e
f

Figure 8: (*previous page*) The higher-order code captures more abstract properties of image data and therefore forms a more invariant representation than the coefficients of linear basis functions. We trained the model on natural images (panels a-c) and scanned newspaper clippings (d-f) and analyzed the representation formed by the model as it varied over the images. A sliding window (represented as white squares in the images) was applied over contiguous sections of the training data (a,d), and values of three linear coefficients u_i (b,e) and three higher-order coefficients v_j (c,f) were plotted as they varied over the signal. Although the model is trained on image patches selected randomly from the dataset, the higher-order code forms a representation that changes more slowly over space and captures properties of the data that extend beyond the sampling window, such as the overall contrast in natural images or the position of the text-line in newspaper images.

properties, which themselves change more slowly.

5.3 Modeling Residual Dependencies

The motivating examples (Figures 1 and 2) showed specific types of residual dependencies among the "independent" linear coefficients, such as the dependence among the scale of coefficients, which formed patterns that changed from context to context. The adapted hierarchical density component model is able to capture these dependencies. First, drawing from the model generates data with similar statistical regularities. Furthermore, the higher-order representation in the model defines an implicit normalization of the linear code, and the residual dependencies are no longer observed in the normalized code.

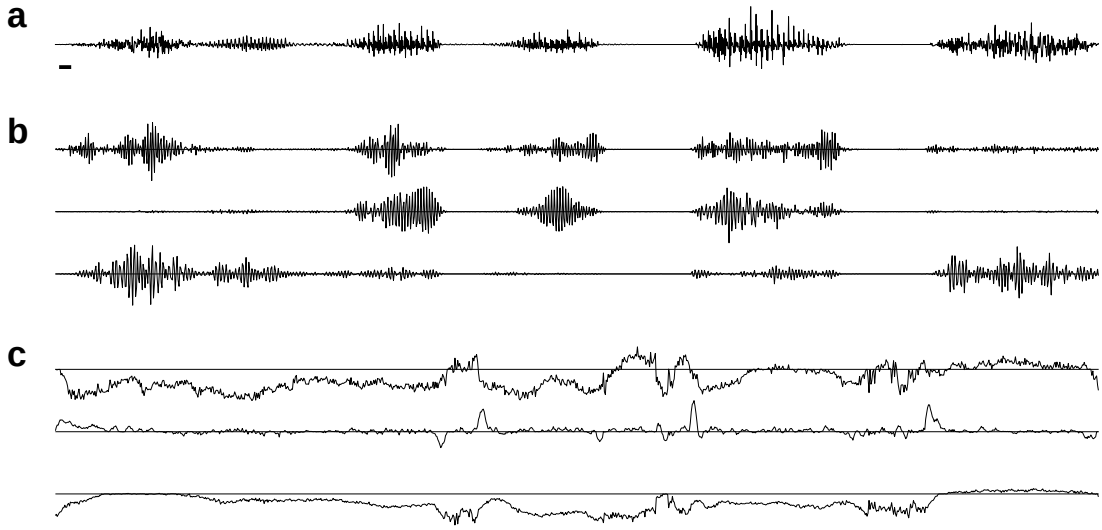


Figure 9: The higher-order representation formed by the hierarchical model trained on speech data is more invariant than simple outputs of linear filters. A sliding window was applied to a speech signal (a, size of window indicated by a short bar). At each point, the linear basis function coefficients **u** were computed (b) and the higher-order coefficients **v** were inferred (c). Values of **v** change slowly and represent more abstract properties, such as the presence of silence or the onset of vocalization.

Figure 10a shows the empirical joint distributions (top row) of two linear coefficients when sampled from the image regions R_1 or R_2 of Figure 1. In the two contexts, the shape of the distribution is different – the coefficients have high variance in one context but not in the other. The statistical properties in the two contexts are captured by the inferred density component coefficients. Fixing the density component coefficient to the empirical distributions and sampling the linear coefficients reveals the same type of statistical structure (middle row). At the same time, it is possible to use the estimated parameters of the generating distribution to normalize the data. Dividing the linear coefficients by the estimated scale parameters $\hat{\boldsymbol{\lambda}}$ results in joint distributions that are symmetric with uniform variance across different contexts and image regions (bottom row).

Another way to observe dependence among coefficient magnitudes is to draw a conditional histogram that plots distributions of one coefficient conditional on different values of another (Simoncelli, 1997; Schwartz and Simoncelli, 2001). While the joint histograms show that coefficient magnitudes are dependent on the sampling context, conditional histograms reveal pair-wise dependencies between coefficients across all contexts. For natural images, most linear coefficients show a positive magnitude dependence, i.e. the magnitude of one coefficient is positively correlated with the magnitude of another, (e.g. the left pair in figure 10b), but some exhibit the reverse pattern. Sampling from the model produces data with the same statistical dependencies (Figure 10b, middle row), while normalized linear coefficients show no conditional magnitude dependence (Figure 10b, bottom row).

Joint and conditional histograms illustrate pair-wise structure in the linear coefficients; global patterns in coefficients, such as those observed in Figures 1 and 2 are

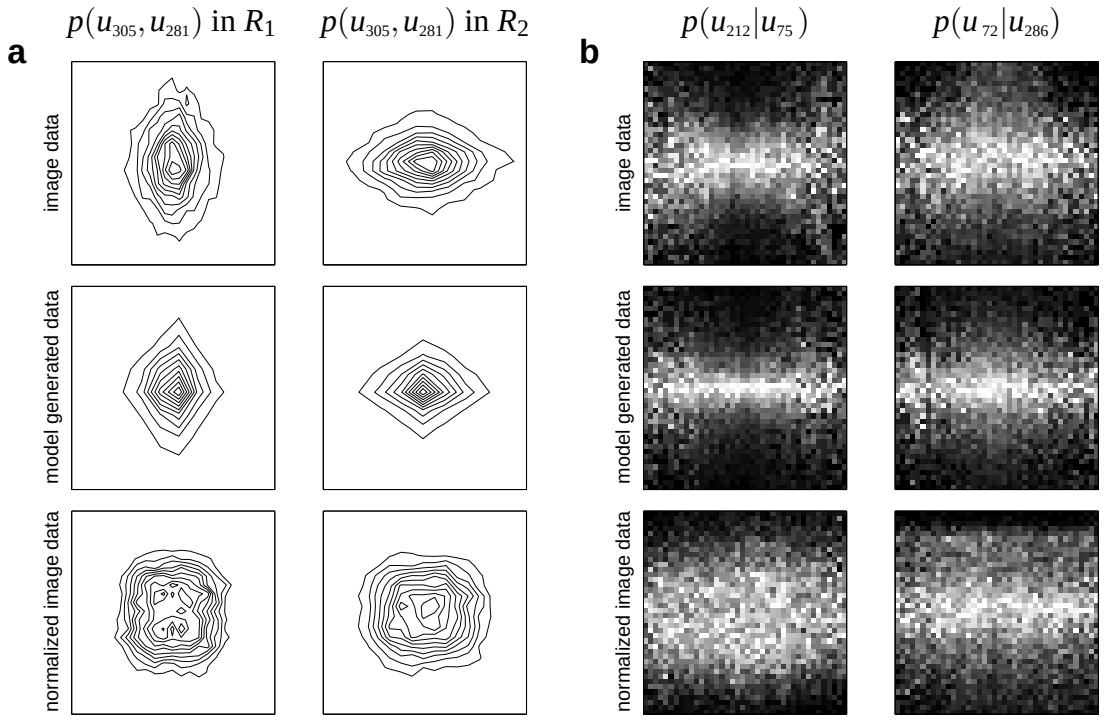


Figure 10: Dependence in the magnitudes of linear basis function coefficients are captured by the density component model. (a) The joint distributions of linear coefficients are different in the two image regions from Figure 1, i.e. the data distribution is not stationary. Sampling from the model under the estimated higher-order representation of each context results in similar distributions. Normalizing the image data by the estimated scale parameters, $\bar{u_i} = u_i/\lambda_i$, eliminates the non-stationarity. (b) Over the full data ensemble, empirical conditional histograms for pairs of coefficients show statistical dependencies in the magnitude. Sampling from the model adapted to this data ensemble produces similar dependencies, and normalizing by the estimated scale parameters removes the magnitude correlations. See text for more details.

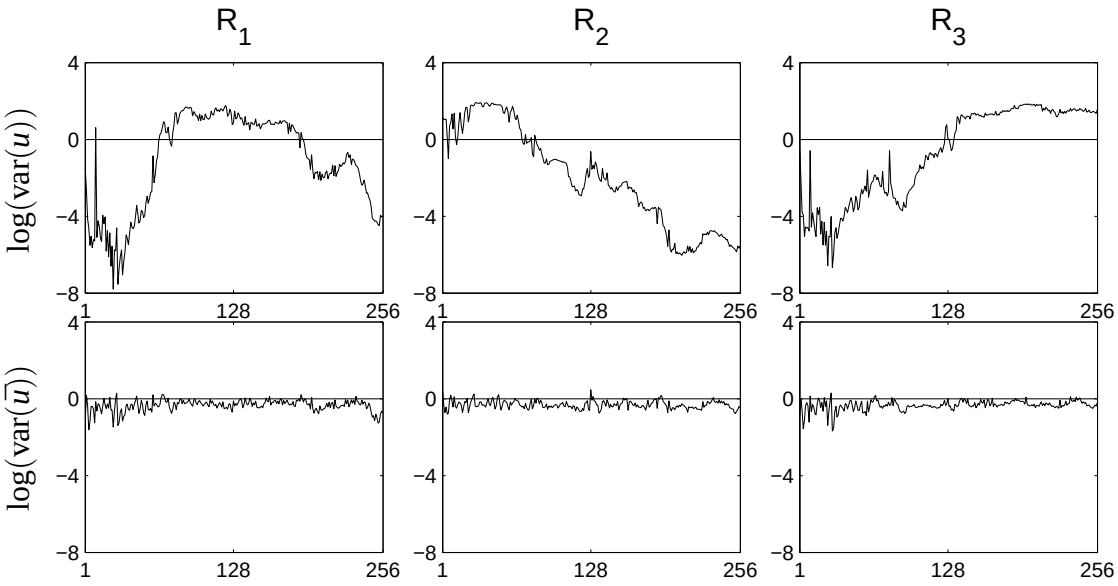


Figure 11: The model accounts for non-stationary statistics of coefficients. Top row: log variance of **u** for the three regions in the speech signal from Figure 2b. Each plot shows the log variance of 256 basis functions, sorted by dominant frequency (replotted from Figure 2)c. Bottom row: log variance of the normalized basis function coefficients $\bar{u_i} = u_i/\hat{\lambda}_i$.

also captured by the model. In the top row of Figure 11, we replot the statistics from Figure 2 that show variance patterns in different regions of the speech signal. Below we plot the same statistics for the coefficients normalized by the estimated scale parameters; after normalization, the statistics are stationary and the coefficients are identically distributed. The same global normalization effect is observed for natural images (plots not shown).

6 Discussion

6.1 Related Work

Some previous work has focused on extending linear probabilistic models. Mixtures of linear ICA models have been used to describe high-dimensional, non-Gaussian data drawn from distinct classes (Lee et al., 2000; Lee and Lewicki, 2002). In this approach, the number of classes is specified in advance and an optimal linear basis is learned for each class. This non-linear generative model describes different data distributions for different classes, but its higher-order representation is fundamentally local and does not scale well in domains where the variation in higher-order structure is continuous and high-dimensional. A key problem addressed by the model presented here is the presence and interaction of multiple instrinsic structures, and this is achieved by a continuous, distributed higher-order code.

Other models have extended ICA to handle non-stationary data distributions. Everson and Roberts (1999) proposed a model in which ICA basis functions evolve with time as a first-order Markov diffusion process. Similarly, Pham and Cardoso (2001) de-

veloped and Choi et al. (2002) extended algorithms for non-stationary models in which the variances of the sources modulate slowly in time. These are also related to models of time-varying mean and variance in economics (Bollerslev et al., 1994), and typically model data whose statistics change slowly over time or space. Alternatively, one can describe the variance with a sparse but temporally coherent latent variable (Hyvärinen et al., 2003). However, in many cases, real-world data are subject both to smooth and abrupt changes that do not follow diffusion dynamics or smooth amplitude modulation. In contrast to these approaches, the density component model makes no assumptions of temporal or spatial smoothness. It infers an optimal generating distribution for each data sample based only on the values of that sample, though the inference process is constrained by parameters adapted to the statistical regularities of the entire data ensemble. Thus it is able to capture both smooth and abrupt changes in the underlying structure.

Another approach to capturing intrinsic structures in the data has been to incorporate a specific non-linearity, such as the sum of squares (Krüger, 1998; Hoyer and Hyvärinen, 2002) or sigmoid functions (Lee et al., 1997). The drawback to these models is that the type of structure learned is limited by the specific choice of the non-linearity. Most of these methods also assume a fixed linear representation (e.g. a set of oriented, localized 2D basis functions for image models), and those that adapt the linear representation assume a more constrained form of the non-linear dependence (see below). In the model presented here, the linear basis is adapted to the data and maximizes the statistical independence of the linear representation. This ensures that the statistical regularities captured by the higher-order code represent fundamentally non-linear dependencies, rather than residual dependence resulting from the choice of

a sub-optimal linear basis. Furthermore, in some applications there is no clear choice of linear representation (such as Gabor filters or wavelets in image processing); in such cases, it is sensible to derive the linear code from the statistics of the data.

Several earlier models have explicitly represented the dependence among coefficients of linear basis functions. In the subspace ICA model (Hyvärinen and Hoyer, 2000), the linear basis functions are grouped into neighborhoods and adapted to maximize the independence of the vector norms of the neighborhoods. Basis functions within a neighborhood are no longer assumed to be independent; in fact, the energies of their coefficients are correlated. In the more generalized form of the model, called topographic ICA, the disjoint sets of dependent basis functions are replaced by a topographic arrangement that defines magnitude dependencies among basis functions (Hyvärinen et al., 2001a). The generative forms of subspace ICA and topographic ICA can be interpreted as more constrained versions of the density component model presented here. Neighborhood or topographic dependencies can be equivalently represented by density components whose weights are specified in advance to reflect tree-dependent or topographic relationships. The density component model, however, places no such constraints on the higher-order representation; thus, density components adapted to the data can capture non-topographic dependencies as well.

A related set of work has attempted to model the dependence among coefficients of a fixed linear transform, such as a multiscale wavelet decomposition. Romberg et al. (2001) used a set of discrete latent variables, propagated along a multiscale wavelet tree, to describe the distribution of each wavelet coefficient. The transition probabilities of the latent states were adapted to match the scale dependencies between adjacent nodes in the tree. Buccigrossi and Simoncelli (1999) computed a linear predictor of scale

for each coefficient as a function of the magnitudes of its neighbors. Wainwright et al. (2001) extended this approach by modeling the wavelet coefficients as observed variables in a Gaussian scale mixture, in which random Gaussian variables are multiplied by latent scaling variables. Dependence among coefficients adjacent on the wavelet tree is captured through the structure of a Gaussian process defined on the scaling variables. In addition to its reliance on a fixed linear representation (the drawbacks of this are outlined above), this model is limited in that it can only describe pair-wise dependencies between variables adjacent on the wavelet tree. Adapting a model to learn global statistical regularities, as opposed to local representations of class structure or pair-wise dependence, allows it to capture a wider range of intrinsic structures. Also, learning an efficient basis to describe these dependencies facilitates their interpretability and provides a better fit to the underlying structure.

6.2 Conclusions

We have introduced a hierarchical, generative Bayesian model that can be considered a non-linear extension to ICA. It uses parametric density estimation to learn statistical regularities from the data and makes no assumptions about the type of structure it expects to find. The model is general – it is not specific to any domain and can be applied to any dataset with rich statistical structure. Because the model forms distributed representations at all levels of its hierarchy, it scales well to large dimensional data.

Adapted to patches from natural images or samples from speech data, the density component model was able to learn non-linear statistical regularities. It yielded a distributed representation of context, which included higher-order spatial relationships for image data, and frequency and harmonic structure for audio data. Sampling from

the model produced data with the same statistical regularities observed in the training datasets; and the model's implicit normalization of the lower-order code accounted for the residual dependencies observed in various datasets.

Recently, it has been argued that higher-order properties of natural signals change slowly across time or space, and that this spatial and temporal *coherence* can be utilized to extract higher-order structure from the data (Foldiak, 1991; Kayser et al., 2001; Wiskott and Sejnowski, 2002; Hurri and Hyvärinen, 2003). We show that in some cases, simply learning higher-order statistical regularities in the data leads the model to recover more abstract properties that tend to vary slowly with time or space. This raises the possibility that the explicit computational goal of extracting coherent (slowly changing) parameters is helpful, but not necessary to learning intrinsic structures that underlie the variation in the data.

One result of learning global statistical regularities is that the learned structure is not necessarily obvious; for example, density components adapted to natural images describe a variety of statistical regularities, some of which are not easily interpreted. This is true for many unsupervised learning models that do not specify in advance the structure to be learned. For example, ICA applied to natural images yields a matrix of basis functions whose functional interpretation has ranged from edge detectors (Bell and Sejnowski, 1997) to models of biological sensory systems (van Hateren and van der Schaaf, 1998). The work presented here suggests that as more powerful unsupervised learning models are developed, the analysis of learned parameters and data representations will gain in importance.

The approach taken in this work is to attack a difficult problem – capturing intrinsic regularities in complex high-dimensional data – incrementally. Although the model is

able to capture some non-linear statistical regularities, the structure it learns is still quite low-level. This step-wise approach stands in contrast to other computational schemes that solve specific problems, such as perceptual invariance or scene segmentation. This may prove more tractable and robust because it does not rely on preconceived notions of intrinsic structures, but learns them from the data. This approach might also give more insight into the organization of biological perceptual systems, where each processing unit performs a relatively simple computational task, and many computational goals might be achieved incrementally and in parallel.

Acknowledgments

We thank Bruno Olshausen and Eero Simoncelli for helpful discussions. This work was supported by a Dept. of Energy Computational Science Graduate Fellowship to YK and National Science Foundation grant no. 0238351 to MSL.

Appendix

The value of $\hat{\mathbf{v}}$ for a given $\mathbf{u}$ was obtained by maximizing the log posterior distribution

$$L = \log p(\mathbf{v}|\mathbf{u},\mathbf{B}) \propto \log p(\mathbf{u}|\mathbf{B},\mathbf{v})p(\mathbf{v}) \tag{16}$$

We use the Laplace distribution for the prior on $\mathbf{v}$ and a generalized Gaussian distribution with the scale parameters $\boldsymbol{\lambda}$ for the likelihood $p(\mathbf{u}|\mathbf{B},\mathbf{v})$, so that

$$L \propto \log \prod_{i=1}^{N} z_i \exp\left(-\left|\frac{u_i}{\lambda_i}\right|^{q_i}\right) \prod_{j=1}^{M} z_j \exp\left(-\left|\frac{v_j}{c}\right|^{q_j}\right) \tag{17}$$

$$\propto \sum_{i=1}^{M}\left[\log \frac{q_i}{2\lambda_i \Gamma(1/q_i)} - \left|\frac{u_i}{\lambda_i}\right|^{q_i}\right] + \sum_{j=1}^{M}\left[\log \frac{q_j}{2\Gamma(1/q_j)} - \left|\frac{v_j}{c}\right|^{q_j}\right] \tag{18}$$

$$\propto \sum_{i=1}^{N}\left[-\log\lambda_i - \left|\frac{u_i}{\lambda_i}\right|^{q_i}\right] - \sum_{j=1}^{M}\left|\frac{v_j}{c}\right|^{q_j}, \tag{19}$$

where $z = q/(2\lambda\Gamma(1/q))$ is the normalization term, $\lambda_i = ce^{[\mathbf{Bv}]_i}$, and $c = \sqrt{\Gamma(1/q)/\Gamma(3/q)}$. For a given data sample, $\mathbf{u}$ is the $N \times 1$ vector of linear basis function coefficients and $\mathbf{v}$ the $M \times 1$ vector of density component coefficients. $\mathbf{A}$ is the $N \times N$ matrix of linear basis functions and $\mathbf{B}$ is the $N \times M$ matrix of density components. We use $[\mathbf{Bv}]_i$ to denote the i^{th} element of the vector $\mathbf{Bv}$, and $\mathbf{B}_i$ to denote the i^{th} row of the matrix $\mathbf{B}$.

The MAP estimate $\hat{\mathbf{v}}$ was obtained by gradient ascent,

$$\frac{\partial L}{\partial v_j} = \frac{\partial}{\partial v_j}\left[\sum_{i=1}^{N}\left[-\log\lambda_i - \left|\frac{u_i}{\lambda_i}\right|^{q_i}\right] - \sum_{j=1}^{M}\left|\frac{v_j}{c}\right|^{q_j}\right] \tag{20}$$

$$= \sum_{i=1}^{N}\left[-B_{ij} + q_i B_{ij}\left|\frac{u_i}{ce^{[\mathbf{Bv}]_i}}\right|^{q_i}\right] - sign(v_j) q_j \frac{|v_j|^{q_j-1}}{c^{q_j}}. \tag{21}$$

The gradient ascent procedure was sensitive to inital conditions and in some cases did not converge to a solution. We tried several alternatives, including a closed-form approximation to the MAP estimate. Ultimately, the most effective learning method was to adjust the step size ε by the stochastic estimate of the Hessian over each batch of data (LeCun et al., 1998):

$$\eta_j = \frac{\varepsilon}{\langle \frac{\partial^2 L}{\partial v_j^2}\rangle + \mu}, \tag{22}$$

where μ is a small constant that improves stability when the second derivative is very small. The second derivative for a data sample is given by

$$\frac{\partial^2 L}{\partial v_j^2} = -\sum_{i=1}^{N} q_i^2 B_{ij}^2 \left| \frac{u_i}{c e^{[\mathbf{Bv}]_i}} \right|^{q_i} - q_j(q_j-1)\frac{|v_j|^{q_j-2}}{c^{q_j}}. \tag{23}$$

The density component matrix **B** was estimated by maximizing the posterior over the data ensemble,

$$\log p(\mathbf{B}|\mathbf{x}_1,\ldots,\mathbf{x}_N,\mathbf{A}) \propto \log p(\mathbf{x}_1,\ldots,\mathbf{x}_N|\mathbf{A},\mathbf{B})p(\mathbf{B}) \tag{24}$$

$$\propto \sum_n \log p(\mathbf{x}_n|\mathbf{A},\mathbf{B})p(\mathbf{B}) \tag{25}$$

$$\propto \sum_n \log p(\mathbf{u}_n|\mathbf{B},\hat{\mathbf{v}}_n)p(\hat{\mathbf{v}}_n)p(\mathbf{B})/|\det \mathbf{A}|. \tag{26}$$

Let $\hat{L}_n = \log p(\mathbf{u}_n|\mathbf{B},\hat{\mathbf{v}}_n)p(\hat{\mathbf{v}}_n)p(\mathbf{B})$. We place a Gaussian prior on **B** and implement gradient ascent $\Delta B = \frac{1}{N}\sum_n \partial \hat{L}_n/\partial B_{ij}$, where the posterior for each data sample $\mathbf{x}_n$ is

$$\frac{\partial \hat{L}}{\partial B_{ij}} = \frac{\partial}{\partial B_{ij}}\left[\log p(\mathbf{u}_n|\mathbf{B},\hat{\mathbf{v}}_n) + \log p(\hat{\mathbf{v}}_n) + \log p(\mathbf{B})\right] \tag{27}$$

$$= \frac{\partial}{\partial B_{ij}}\left[\sum_{i=1}^{N}\left[-\log\lambda_i - \left|\frac{u_i}{\lambda_i}\right|^{q_i}\right] - \sum_{j=1}^{M}\left|\frac{v_j}{c}\right|^{q_j} - \sum_{i=1,j=1}^{N,M}\frac{B_{ij}^2}{2}\right] \tag{28}$$

$$= -v_j + v_j q_i \left|\frac{u_i}{c e^{[\mathbf{Bv}]_i}}\right|^{q_i} - B_{ij}. \tag{29}$$

References

Bell, A. J. and Sejnowski, T. J. (1995). An information maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159.

Bell, A. J. and Sejnowski, T. J. (1997). The 'independent components' of natural scenes are edge filters. *Vision Res.*, 37(23):3327–3338.

Bollerslev, T., Engle, R. F., and Nelson, D. B. (1994). ARCH models. In *Handbook of Econometrics*. Elsevier Science B.V., Amsterdam.

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

Cardoso, J.-F. (1997). Infomax and maximum likelihood for blind source separation. *IEEE Signal Processing Letters*, 4:109–111.

Choi, S., Cichocki, A., and Belouchrani, A. (2002). Second order nonstationary source separation. *Journal of VLSI Signal Processing*, 32(1-2):93–104.

Cichocki, A. and Amari, S.-I. (2002). *Adaptive Blind Signal and Image Processing : Learning Algorithms and Applications*. J. Wiley.

Everson, R. and Roberts, S. (1999). Non-stationary independent component analysis. *Proceedings of the 8th International Conference on Artificial Neural Networks*, pages 503–508.

Foldiak, P. (1991). Learning invariance from transformation sequences. *Neural Computation*, 3(2):194–200.

Hoyer, P. O. and Hyvärinen, A. (2002). A multi-layer sparse coding network learns contour coding from natural images. *Vision Res.*, 42:1593–1605.

Hurri, J. and Hyvärinen, A. (2003). Simple-cell-like receptive fields maximize temporal coherence in natural video. *Neural Computation*, 15(3):663–691.

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

Hyvärinen, A., Hoyer, P. O., and Inki, M. (2001a). Topographic independent component analysis. *Neural Comput.*, 13:1527–1558.

Hyvärinen, A., Hurri, J., and Väyrynen, J. (2003). Bubbles: a unifying framework for low-level statistical properties of natural image sequences. *Journal of the Optical Society of America A*, 20(7):1237–1252.

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

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

Kayser, C., Einhäuser, W., Dümmer, O., König, P., and Körding, K. (2001). Extracting slow subspaces from natural videos leads to complex cells. *Artificial Neural Networks*, 2130:1075–1080.

Krüger, N. (1998). Collinearity and parallelism are statistically significant second order relations of complex cell responses. *Neural Processing Letters*, 8:117–129.

LeCun, Y., Bottou, L., Orr, G., and Muller, K. (1998). Efficient backprop. In Orr, G. and K., M., editors, *Neural Networks: Tricks of the trade*. Springer.

Lee, T.-W., Koehler, B., and Orglmeister, R. (1997). Blind source separation of non-

linear mixing models. In *Proceedings of IEEE International Workshop on Neural Networks for Signal Processing*.

Lee, T.-W. and Lewicki, M. S. (2002). Unsupervised classification, segmentation and de-noising of images using ICA mixture models. *IEEE Trans. Image Proc.*, 11(3):270–279.

Lee, T.-W., Lewicki, M. S., and Sejnowski, T. J. (2000). ICA mixture models for unsupervised classification of non-Gaussian sources and automatic context switching in blind signal separation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(10):1078–1089.

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

O'Neill, J. C. (1999). DiscreteTFDs time-frequency analysis software. `http://tfd.sourceforge.net/`.

Pearlmutter, B. A. and Parra, L. C. (1996). A context-sensitive generalization of ICA. In *International Conference on Neural Information Processing*, pages 151–157.

Pham, D.-T. and Cardoso, J.-F. (2001). Blind separation of instantaneous mixtures of non stationary sources. *IEEE Trans. Signal Processing*, 49(9):1837–1848.

Romberg, J., Choi, H., and Baraniuk, R. (2001). Bayesian tree-structured image modeling using wavelet domain Hidden Markov models. *IEEE Transactions on Image Processing*, 10(7).

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

Simoncelli, E. P. (1997). Statistical models for images: Compression, restoration and synthesis. In *Proc. 31st Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, CA.

van Hateren, J. H. and van der Schaaf, A. (1998). Independent component filters of natural images compared with simple cells in primary visual cortex. *Proc. Royal Soc. Lond.* **B**, 265:359–366.

Wainwright, M. J., Simoncelli, E. P., and Willsky, A. S. (2001). Random cascades on wavelet trees and their use in analyzing and modeling natural images. *Applied Computational and Harmonic Analysis*, 11:89–123.

Welling, M., Hinton, G. E., and Osindero, S. (2003). Learning sparse topographic representations with products of student-t distributions. In *Advances in Neural Information Processing Systems*, volume 15, Cambridge, MA. MIT Press.

Wiskott, L. and Sejnowski, T. J. (2002). Slow feature analysis: Unsupervised learning of invariances. *Neural Computation*, 14(4):715–770.