Learning the parts of objects by nonnegative matrix factorization

D. D. Lee*
*Bell Laboratories
Lucent Technologies
Murray Hill, NJ 07974

H. S. Seung*,†
†Dept. of Brain and Cognitive Sciences
Massachusetts Institute of Technology
Cambridge, MA 02139

Is perception of the whole based on perception of its parts? There is psychological [1] and physiological [2, 3] evidence for parts-based representations in the brain, and certain computational theories of object recognition rely on such representations [4, 5]. But little is known about how brains or computers might learn the parts of objects. Here we demonstrate an algorithm called nonnegative matrix factorization (NMF) that is able to learn parts of faces and semantic features of text. This is in contrast to other algorithms like principal components analysis (PCA) and vector quantization (VQ), which learn holistic, not parts-based, representations. When all three algorithms are viewed as techniques for matrix factorization, NMF is distinguished from PCA and VQ by its use of nonnegativity constraints. These constraints lead to a parts-based representation because they allow only additive, not subtractive, combinations. When NMF is interpreted as a neural network learning algorithm, parts-based representations emerge by virtue of two properties: the firing rates of neurons are never negative and synaptic strengths do not change sign.

We have applied NMF, along with PCA and VQ, to a database of facial images. As shown in Figure 1, all three algorithms learn to represent a face as a linear combination of basis images, but with qualitatively different results. VQ discovers a basis consisting of prototypes, each of which is a whole face. The basis images for PCA are "eigenfaces", some of which resemble distorted versions of whole faces [6]. The NMF basis is radically different: its images are localized features that correspond better with intuitive notions of the parts of faces.

How does NMF learn such a representation, so different from the holistic representations of PCA and VQ? To answer this question, it is helpful to describe the three algorithms in a matrix factorization framework. The image database is regarded as an $n \times m$ matrix V, each column of which contains n nonnegative pixel values of one of the m facial images. Then all three algorithms construct approximate factorizations of the form $V \approx WH$, or

$$V_{i\mu} \approx (WH)_{i\mu} = \sum_{a=1}^{r} W_{ia} H_{a\mu} \ . \tag{1}$$

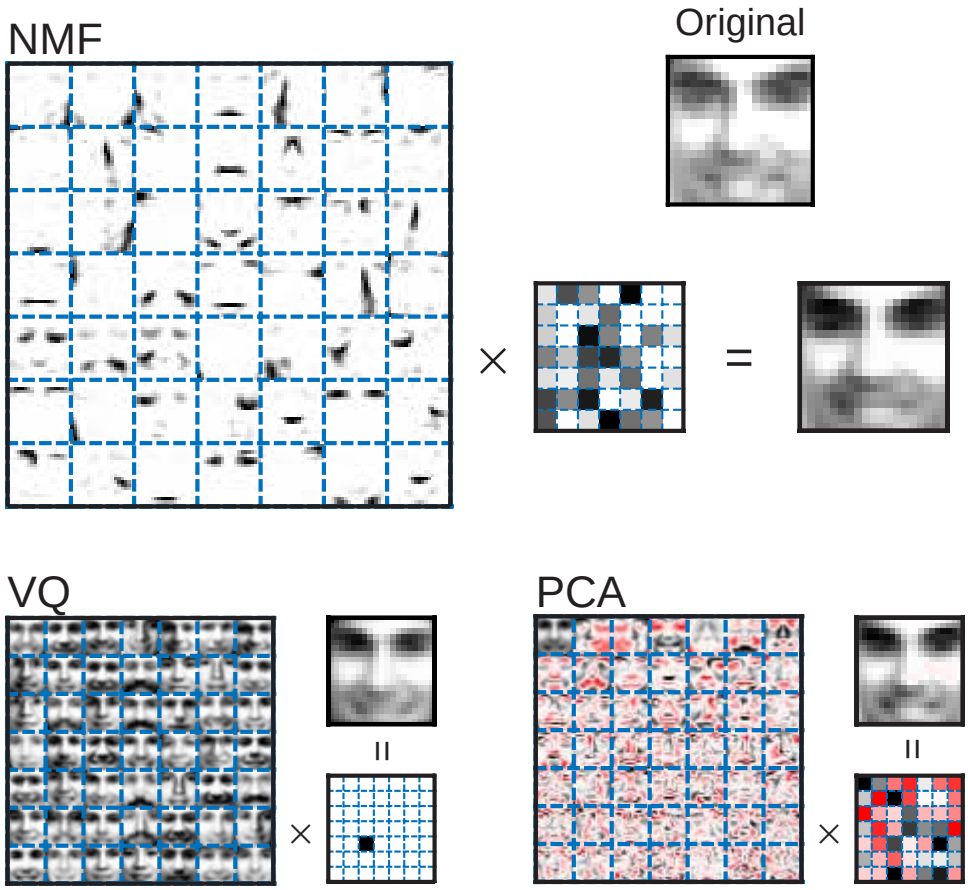

Figure 1: Nonnegative matrix factorization (NMF) learns a parts-based representation of faces, while vector quantization (VQ) and principal components analysis (PCA) learn holistic representations. The three learning algorithms were applied to a database of $m = 2429$ facial images, each consisting of $n = 19 \times 19$ pixels, and constituting an $n \times m$ matrix V. All three learning algorithms find approximate factorizations of the form $V \approx WH$, but with three different types of constraints on W and H, as described more fully in the main text and methods. As shown in the 7×7 montages, each algorithm has learned a set of $r = 49$ basis images. Positive values are illustrated with black pixels and negative values with red pixels. A particular instance of a face, shown at top right, is approximately represented by a linear superposition of basis images. The coefficients of the linear superpositions are shown next to each montage, in a 7×7 grid, and the resulting superpositions are shown on the other side of the equality sign. Unlike VQ and PCA, the NMF algorithm learns to represent faces with a set of basis images resembling parts of faces.

The r columns of W are called *basis images*. Each column of H is called an *encoding*, and is in one-to-one correspondence with a face in V. An encoding consists of the coefficients by which a face is represented with a linear combination of basis images. The dimensions of the matrix factors W and H are $n \times r$ and $r \times m$, respectively. The rank r of the factorization is generally chosen so that $(n + m)r < nm$, and the product WH can be regarded as a compressed form of the data in V.

The differences between PCA, VQ, and NMF arise from different constraints imposed on the matrix factors W and H. In VQ, each column of H is constrained to be a unary vector, with one element equal to unity, and the other elements equal to zero. In other words, every face (column of V) is approximated by a single basis image (column of W) in the factorization $V \approx WH$. Such a unary encoding for a particular face is shown next to the VQ basis in Fig. 1. This unary representation forces VQ to learn basis images that are prototypical faces.

PCA constrains the columns of W to be orthonormal and the rows of H to be orthogonal to each other. This relaxes the unary constraint of VQ, allowing a distributed representation in which each face is approximated by a linear combination of all the basis images, or eigenfaces [6]. A distributed encoding of a particular face is shown next to the eigenfaces in Fig. 1. While eigenfaces have a statistical interpretation as the directions of largest variance, many of them do not have an obvious visual interpretation. This is because PCA allows the entries of W and H to be of arbitrary sign. Since the eigenfaces are used in linear combinations that generally involve complex cancellations between positive and negative numbers, many individual eigenfaces lack intuitive meaning.

NMF does not allow negative entries in the matrix factors W and H. Unlike the unary constraint of VQ, these nonnegativity constraints permit the combination of multiple basis images to represent a face. But only additive combinations are allowed, because the nonzero elements of W and H are all positive. In contrast to PCA, no subtractions can occur. For these reasons, the nonnegativity constraints are compatible with the intuitive notion of combining parts to form a whole, which is how NMF learns a part-based representation.

The actual implementation of the NMF algorithm consists of the update rules for W and H given in Fig. 2. It can be shown that iteration of these update rules converges to a local maximum of the objective function

$$F = \sum_{i=1}^{n} \sum_{\mu=1}^{m} [V_{i\mu} \log (WH)_{i\mu} - (WH)_{i\mu}] \tag{2}$$

subject to the nonnegativity constraints described above. This objective function can be derived by interpreting NMF as an algorithm for constructing a probabilistic model of image generation. In this model, an image pixel $V_{i\mu}$ is generated by adding Poisson noise to the product $(WH)_{i\mu}$. The objective function in Eq. (2) is then related to the likelihood of generating the images in V from the basis W and encodings H.

The exact form of the objective function is not as crucial as the nonnegativity constraints for the success of the NMF algorithm in learning parts. A squared error objective function leads to a set of update rules for W and H different from those in Fig. 2 [9, 10]. These update rules yield results similar to those shown

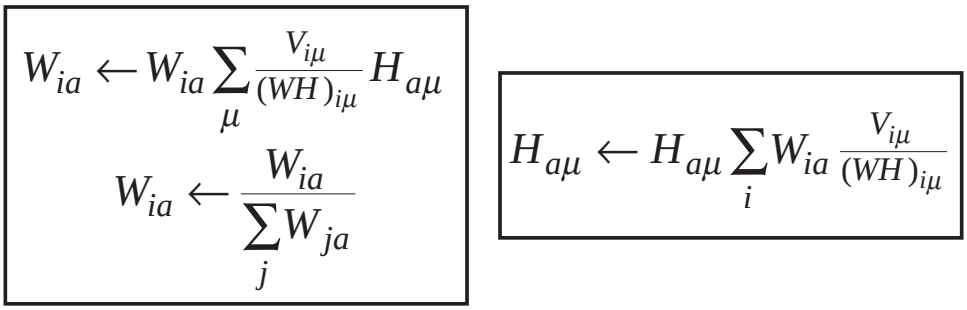

Figure 2: Iterative algorithm for nonnegative matrix factorization. Starting from nonnegative initial conditions for W and H, iteration of these update rules for nonnegative V finds an approximate factorization $V \approx WH$ by converging to a local maximum of the objective function given in Eq. (2). The fidelity of the approximation enters the updates through the quotient $V_{i\mu}/(WH)_{i\mu}$. Monotonic convergence can be proven using techniques similar to those used in proving the convergence of the EM algorithm [7, 8]. The update rules preserve the nonnegativity of W and H, and also constrain the columns of W to sum to unity. This sum constraint is a convenient way of eliminating the degeneracy associated with the invariance of WH under the transformation $W \to \lambda W$, $H \to \lambda^{-1} H$, where λ is a scalar.

in Fig. 1, but have the technical disadvantage of requiring the adjustment of a parameter controlling the learning rate. This parameter is generally adjusted through trial and error, which can be a time-consuming process if the matrix V is very large. Therefore, the update rules described in Fig. 2 may be advantageous for applications involving large databases.

It is helpful to visualize the dependencies between image pixels and encoding variables in the form of the network shown in Fig. 3. The top layer of nodes represents an encoding $h_1, \ldots, h_r$ (column of H), and the bottom layer an image $v_1, \ldots, v_n$ (column of V). The matrix element W_{ia} quantifies the amount of influence that the ath encoding variable h_a has on the ith image pixel v_i. A single encoding variable influences multiple image pixels, due to the fan-out of connections from the encoding variable. Because of the nonnegativity of W_{ia}, this influence is restricted to *coactivation* of image pixels, Intuitively, a parts-based representation should be learnable from observations of coactivation in V, as the image pixels belonging to the same part of the face are coactivated when that part is present. The NMF algorithm learns by adapting W_{ia} to generate the appropriate coactivations.

As can be seen from Fig. 1, the NMF basis and encodings contain a large fraction of vanishing coefficients, so both the basis images and image encodings are sparse. The basis images are sparse because they are non-global and contain several versions of mouths, noses, and other facial parts, where the various versions are in different locations or forms. The variability of a whole face is generated by combining these different parts. Although all parts are used by at least one face, any given face does not use all the available parts. This results in a *sparsely distributed* image encoding, in contrast to the unary encoding of VQ and the fully distributed PCA encoding [11, 12, 13].

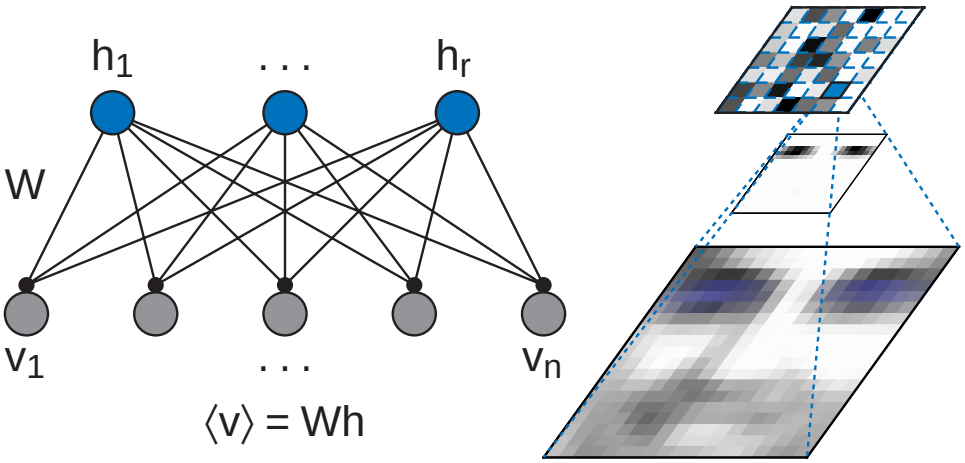

Figure 3: Probabilistic hidden variables model underlying nonnegative matrix factorization. The model is diagrammed as a network, depicting how the visible variables $v_1, \ldots, v_n$ in the bottom layer of nodes are generated from the hidden variables $h_1, \ldots, h_r$ in the top layer of nodes. According to the model, the visible variables v_i are generated from a probability distribution with mean $\sum_a W_{ia} h_a$. In the network diagram, the influence of h_a on v_i is represented by a connection with strength W_{ia}. In the application to facial images, the visible variables are the image pixels, while the hidden variables contain the parts-based encoding. For fixed a, the connection strengths $W_{1a}, \ldots, W_{na}$ constitute a specific basis image (right middle) that is combined with other basis images to represent a whole facial image (right bottom).

The preceding description of NMF has been specialized to images, but the algorithm is actually applicable to a wide variety of problem domains. More generally, NMF is a method for modeling the generation of directly observable visible variables V from hidden variables H [14, 15]. Each hidden variable coactivates a subset of visible variables, or "part." Activation of a constellation of hidden variables combines these parts additively to generate a whole. Seen in this light, NMF has a very broad range of potential applications. We illustrate this versatility by applying NMF to a completely different problem, the semantic analysis of text documents.

For this application, a corpus of documents is summarized by a matrix V, where $V_{i\mu}$ is the number of times the ith word in the vocabulary appears in the μth document [16]. These word counts can be regarded as a set of visible variables, and modeled as being generated from an underlying set of hidden variables. Application of the VQ, PCA, or NMF algorithms involves finding the approximate factorization of this matrix $V \approx WH$ into a feature set W and hidden variables H, in the same way as was done for faces.

In the VQ factorization, a single hidden variable is active for each document. If the same hidden variable is active for a group of documents, they are semantically related, because they have similar frequencies of word occurrence. Consequently, the hidden variables are called *semantic* variables, and VQ is accordingly used for automatic semantic indexing of documents by topic [16]. Each column of W, or semantic feature, consists of the word frequencies for the corresponding semantic variable.

VQ allows only one semantic variable to be active, which prevents more than one topic from being attributed to a document. PCA would seem to be a solution to this problem, as it allows activation of multiple semantic variables. While PCA has been successful in certain linguistic tasks [17], it generally results in semantic variables that are difficult to interpret, for much the same reason that the PCA representation of faces has no obvious visual interpretation. This is the result of two unrealistic aspects of the model: all semantic variables are used to represent each document, and negative values for semantic variables are allowed. Intuitively, it makes more sense for each document to be associated with some small subset of a large array of topics, rather than just one topic or all the topics. Since the sparsely distributed representation of NMF appears ideally suited for this purpose, we applied NMF to the semantic analysis of a corpus of encyclopedia articles.

Some of the $r = 200$ semantic features (columns of W) discovered by NMF are shown in Fig. 4. In each semantic feature, the algorithm has grouped together semantically related words. Each article in the encyclopedia is represented by additively combining several of these features. For example, to represent the article about the "Constitution of the United States," the semantic feature containing "supreme" and "court" and the one containing "president" and "congress" are coactivated.

In addition to grouping semantically related words together into semantic features, the algorithm uses context to differentiate between multiple meanings of the same word. For example, the word "lead" appears with high frequency in two semantic features shown in Fig. 4: it occurs with "metal", "copper", and "steel" in one, while it appears with "person", "rules", and "law" in the other. This demonstrates that NMF is able to deal with the *polysemy* of "lead" by disambiguating two of its meanings in the corpus of documents.

Although the NMF algorithm is successful in learning facial parts and semantic topics, this success does not imply that the algorithm can learn parts from any database, such as images of objects viewed from extremely different viewpoints, or highly articulated objects. Learning parts for these complex cases is likely to require fully hierarchical models with multiple levels of hidden variables, instead of the single level in NMF. While nonnegativity constraints may help such models to learn parts-based representations [15], we do not claim that they are sufficient in themselves. The NMF algorithm also does not learn anything about the "syntactic" relationships between parts. NMF assumes that the hidden variables are nonnegative, but makes no further assumptions about their statistical dependencies.

This is in contrast to independent components analysis (ICA), a variant of PCA which assumes that the hidden variables are statistically independent and non-Gaussian [18, 19]. Applying ICA to the facial images in order to make the encodings independent results in basis images that are holistic. The independence assumption of ICA is ill-suited for learning parts-based representations because various parts are likely to occur together. This results in complex dependencies between the hidden variables that cannot be captured by algorithms that assume independence in the encodings. An alternative application of ICA is to transform the PCA basis images in order to make the images rather than the encodings as statistically independent as possible [20]. This results in a basis that is non-global; however, in this representation all the basis images

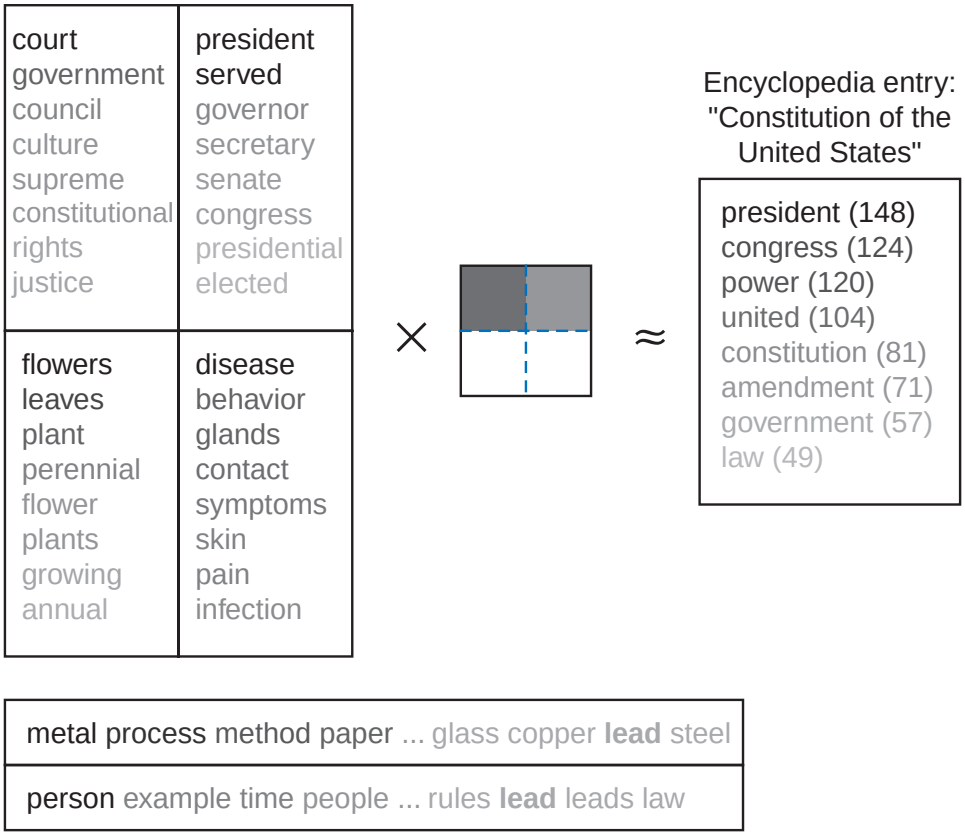

Figure 4: NMF discovers semantic features of $m = 30991$ articles from the Grolier encyclopedia. For each word in a vocabulary of size $n = 15276$, the number of occurrences was counted in each article, and used to form the 15276×30991 matrix V. Each column of V contained the word counts for a particular article, while each row of V contained the counts of a particular word in different articles. The matrix was approximately factorized into the form WH using the NMF algorithm described in Fig. 2. Just four of the $r = 200$ semantic features (columns of W) are described in the upper left. Since they are very high-dimensional vectors, each semantic feature is represented by a list of the eight words with highest frequency in that feature. The darkness of the text indicates the relative frequency of each word within a feature. On the right are the eight most frequent words and their counts in the encyclopedia entry on the "Constitution of the United States." This word count vector was approximated by a superposition that gave high weight to the upper two semantic features, and none to the lower two, as shown by the four shaded squares in the middle indicating the activities of H. The bottom of the figure exhibits the two semantic features containing "lead" with high frequencies. Judging from the other words in the features, two different meanings of "lead" are differentiated by the NMF algorithm.

are used in cancelling combinations to represent an individual face, and thus the encodings are not sparse. In contrast, the NMF representation contains both a basis and encoding that are naturally sparse, in that many of the components are exactly equal to zero. Sparseness in both the basis and encodings is crucial for a parts-based representation.

The algorithm of Fig. 2 performs both learning and inference simultaneously. That is, it both learns a set of basis images, and also infers values for the hidden variables from the visible variables. Although the generative model of Fig. 3 is linear, the inference computation is nonlinear due to the nonnegativity constraints. The computation is similar to maximum likelihood reconstruction in emission tomography [21], and deconvolution of blurred astronomical images [22, 23]. It is guaranteed to converge because the objective function in Eq. (2) is convex in H.

According to the generative model of Fig. 3, visible variables are generated from hidden variables by a network containing excitatory connections. A neural network that infers the hidden from the visible variables requires the addition of inhibitory feedback connections. NMF learning is then implemented through plasticity in the synaptic connections. A full discussion of such a network is out of the scope of this letter. Here we only point out the consequence of the nonnegativity constraints, which is that synapses are either excitatory or inhibitory, but do not change sign. Furthermore, the nonnegativity of the hidden and visible variables corresponds to the physiological fact that the firing rates of neurons cannot be negative. We suggest that the one-sided constraints on neural activity and synaptic strengths in the brain may be important for developing sparsely distributed, parts-based representations for perception.

Methods The facial images used in Fig. 1 consisted of frontal views hand-aligned in a 19×19 grid. For each image, the grayscale intensities were first linearly scaled so that the pixel mean and standard deviation were equal to 0.25, and then clipped to the range $[0, 1]$. NMF was performed with the iterative algorithm described in Fig. 2, starting with random initial conditions for W and H. The algorithm was mostly converged after less than 50 iterations; the results shown are after 500 iterations, which took a few hours of computation time on a Pentium II computer. PCA was done by diagonalizing the matrix VV^T. Displayed are the 49 eigenvectors with the largest eigenvalues. VQ was done via the k-means algorithm, starting from random initial conditions for W and H.

In the semantic analysis application of Fig. 4, the vocabulary was defined as the 15276 most frequent words in the database of Grolier encyclopedia articles, after removal of the 430 most common words, such as "the" and "and." Since most words appear in relatively few articles, the word count matrix V is extremely sparse, which speeds up the algorithm. The results shown are after the update rules of Fig. 2 were iterated 50 times starting from random initial conditions for W and H.

References

[1] Palmer, S. E. Hierarchical structure in perceptual representation. *Cogn. Psychol.* **9,** 441–474 (1977).

[2] Wachsmuth, E., Oram, M. W., & Perrett, D. I. Recognition of objects and their component parts: responses of single units in the temporal cortex of the macaque. *Cereb. Cortex* **4,** 509–522 (1994).

[3] Logothetis, N. K. & Sheinberg, D. L. Visual object recognition. *Annu. Rev. Neurosci.* **19,** 577–621 (1996).

[4] Biederman, I. Recognition-by-components: a theory of human image understanding. *Psychol. Rev.* **94,** 115–147 (1987).

[5] Ullman, S. *High-level vision: object recognition and visual cognition.* (MIT Press, Cambridge, MA, 1996).

[6] Turk, M. & Pentland, A. Eigenfaces for recognition. *J. Cogn. Neurosci.* **3,** 71–86 (1991).

[7] Dempster, A. P., Laird, N. M. & Rubin, D. B. Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Stat. Soc.* **39,** 1–38 (1977).

[8] Saul, L. & Pereira, F. Aggregate and mixed-order Markov models for statistical language processing. In C. Cardie and R. Weischedel (eds). Proceedings of the Second Conference on Empirical Methods in Natural Language Processing, 81–89. (ACL Press, 1997).

[9] Lee, D. D. & Seung, H. S. Unsupervised learning by convex and conic coding. In *Proceedings of the Conference on Neural Information Processing Systems 1996*, 515–521 (Morgan Kaufmann, 1997).

[10] Paatero, P. Least squares formulation of robust non-negative factor analysis. *Chemometr. Intell. Lab.* **37,** 23–35 (1997).

[11] Field D. J. What is the goal of sensory coding? *Neural Comput.* **6**, 559–601 (1994).

[12] Foldiak, P. & Young, M. Sparse coding in the primate cortex. *The Handbook of Brain Theory and Neural Networks*, 895–898. (MIT Press, Cambridge, MA, 1995).

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

[14] Nakayama, K., Shimojo, S. Experiencing and perceiving visual surfaces. *Science* **257,** 1357–1363 (1992).

[15] Hinton, G. E., Dayan, P., Frey, B. J. & Neal, R. M. The "wake-sleep" algorithm for unsupervised neural networks. *Science* **268,** 1158–1161 (1995).

[16] Salton, G. & McGill, M. J. *Introduction to Modern Information Retrieval* (McGraw-Hill, New York, NY, 1983).

[17] Landauer, T. K. & Dumais, S. T. The latent semantic analysis theory of knowledge. *Psychol. Rev.* **104,** 211–240 (1997).

[18] Jutten, C. & Herault, J. Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture. *Signal Proc.* **24,** 1–10 (1991).

[19] Bell, A. J. & Sejnowski, T. J. An information maximization approach to blind separation and blind deconvolution. *Neural Comput.* **7,** 1129–1159 (1995).

[20] Bartlett, M. S., Lades, H. M. & Sejnowski, T. J. Independent component representations for face recognition. *Proc. SPIE*, **3299**, 528–539 (1998).

[21] Shepp, L. A. & Vardi, Y. Maximum likelihood reconstruction for emission tomography. *IEEE Trans.* **MI-2,** 113–122 (1982).

[22] Richardson, W. H. Bayesian-based iterative method of image restoration. *J. Opt. Soc. Am.* **62,** 55–59 (1972).

[23] Lucy, L. B. An iterative technique for the rectification of observed distributions. *Astron. J.* **74,** 745–754 (1974).

Acknowledgments We acknowledge the support of Bell Laboratories. C. Papageorgiou and T. Poggio provided us with the database of faces, and R. Sproat with the Grolier encyclopedia corpus. We are grateful to L. Saul for convincing us of the advantages of EM-type algorithms. We have benefited from discussions with B. Anderson, K. Clarkson, R. Freund, L. Kaufman, E. Rietman, S. Roweis, N. Rubin, J. Tenenbaum, M. Tsodyks, T. Tyson and M. Wright.