larvae collected randomly in the field (2° 48.12′ N, 41° 40.33′ E) by SCUBA. Between 5 and 10 juveniles were recruited successfully in each of 15, 1 l polystyrene containers ($n = 15$), the bottom of which was covered with an acetate sheet that served as substratum for sponge attachment. Containers were then randomly distributed in 3 groups, and sponges in each group were reared for 14 weeks in 3 different concentrations of Si(OH)$_4$: 0.741 ± 0.133, 30.235 ± 0.287 and $100.041 \pm 0.760 \, \mu M$ (mean $\pm$ s.e.). All cultures were prepared using 0.22 μm polycarbonate-filtered seawater, which was collected from the sponge habitat, handled according to standard methods to prevent Si contamination[29] and enriched in dissolved silica, when treatments required, by using Na$_2$SiF$_6$. During the experiment, all sponges were fed by weekly addition of 2 ml of a bacterial culture (40–60×10^6 bacteria ml^{-1}) to each container[30]. The seawater was replaced weekly, with regeneration of initial food and Si(OH)$_4$ levels. The concentration of Si(OH)$_4$ in cultures was determined on 3 replicates of 1 ml seawater samples per container by using a Bran-Luebbe TRAACS 2000 nutrient autoanalyser. After week 5, the accidental contamination of some culture containers by diatoms rendered subsequent estimates of Si uptake by sponges unreliable, so we discarded them for the study.

For the study of the skeleton, sponges were treated according to standard methods[30] and examined in a Hitachi S-2300 scanning electron microscope (SEM).

1. Hartman, W. D., Wendt, J. W. & Wiedenmayer, F. Living and fossil sponges. Notes for a short course. *Sedimentia* **8,** 1–274 (1980).
2. Ghiold, J. The sponges that spanned Europe. *New Scient.* **129,** 58–62 (1991).
3. Leinfelder, R. R. Upper Jurassic reef types and controlling factors. *Profil* **5,** 1–45 (1993).
4. Wiedenmayer, F. Contributions to the knowledge of post-Paleozoic neritic and archibental sponges (Porifera). *Schweiz. Paläont. Abh.* **116,** 1–147 (1994).
5. Lévi, C. in *Fossil and Recent Sponges* (eds Reitner, J. & Keupp, H.) 72–82 (Springer, New York, 1991).
6. Moret, L. Contribution à l'étude des spongiaires siliceux du Miocène de l'Algérie. *Mém. Soc. Géol. Fr.* **1,** 1-27 (1924).
7. Vacelet, J. Indications de profondeur donnés par les Spongiaires dans les milieux benthiques actuels. *Géol. Méditerr.* **15,** 13–26 (1988).
8. Maldonado, M. & Young, C. M. Bathymetric patterns of sponge distribution on the Bahamian slope. *Deep-Sea Res. I* **43,** 897–915 (1996).
9. Lowenstam, H. A. & Weiner, S. *On Biomineralization* (Oxford Univ., Oxford, 1989).
10. Maliva, R. G., Knoll, A. H. & Siever, R. Secular change in chert distribution: a reflection of evolving biological participation in the silica cycle. *Palaios* **4,** 519–532 (1989).
11. Nelson, D. M., Tréguer, P., Brzezinski, M. A., Leynaert, A. & Quéguiner, B. Production and dissolution of biogenic silica in the ocean: revised global estimates, comparison with regional data and relationship to biogenic sedimentation. *Glob. Biochem. Cycles* **9,** 359–372 (1995).
12. Tréguer, P. *et al.* The silica balance in the world ocean: a reestimate. *Science* **268,** 375–379 (1995).
13. Calvert, S. E. in *Silicon Geochemistry and Biogeochemistry* (ed. Aston, S. R.) 143–186 (Academic, London, 1983).
14. Lisitzyn, A. P. Sedimentation in the world ocean. *Soc. Econ. Palaeon. Mineral. Spec. Pub.* **17,** 1–218 (1972).
15. Weinberg, S. *Découvrir la Méditerranée* (Éditions Nathan, Paris, 1993).
16. Maldonado, M. & Uriz, M. J. Skeletal morphology of two controversial poecilosclerid genera (Porifera, Demospongiae): *Discorhabdella* and *Crambe. Helgoländer Meeresunters.* **50,** 369–390 (1996).
17. Hinde, G. J. & Holmes, W. M. On the sponge remains in the Lower Tertiary strata near Oamaru, New Zealand. *J. Linn. Soc. Zool.* **24,** 177–262 (1892).
18. Kelly-Borges, M. & Pomponi, S. A. Phylogeny and classification of lithisthid sponges (Porifera: Demospongiae): a preliminary assessment using ribosomal DNA sequence comparisons. *Mol. Mar. Biol. Biotech.* **3,** 87–103 (1994).
19. Mostler, H. Poriferenspiculae der alpinen Trias. *Geol. Paläont. Mitt. Innsbruck.* **6,** 1–42 (1976).
20. Palmer, T. J. & Fürsich, F. T. Ecology of sponge reefs from the Upper Bathonian of Normandy. *Palaeontology* **24,** 1–23 (1981).
21. Burckle, L. H. in *Introduction to Marine Micropaleontology* (eds Haq, B. U. & Boersma, A.) 245–266 (Elsevier, Amsterdam, 1978).
22. Austin, B. Underwater birdwatching. *Canadian Tech. Rep. Hydro. Ocean. Sci.* **38,** 83–90 (1984).
23. Koltun, V. M. in *The Biology of the Porifera* (ed. Fry, W. G.) 285–297 (Academic, London, 1970).
24. Tabachnick, K. R. in *Sponges in Time and Space* (eds van Soest, R. M. W., van Kempen, T. M. G. & Braekman, J. C.) 225–232 (A. A. Balkema, Rotterdam, 1994).
25. Harper, H. E. & Knoll, A. H. Silica, diatoms and Cenozoic radiolarian evolution. *Geology* **3,** 175–177 (1975).
26. Hartman, W. D. in *Silicon and Siliceous Structures in Biological Systems* (eds Simpson, T. L. & Volcani, B. E.) 453–493 (Springer Verlag, New York, 1981).
27. Pisera, A. Upper Jurassic siliceous sponges from Swabian Alb: taxonomy and paleoecology. *Palaeont. Pol.* **57,** 1–216 (1997).
28. Reincke, T. & Barthel, D. Silica uptake kinetics of *Halichondria panicea* in Kiel Bight. *Mar. Biol.* **129,** 591–593 (1997).
29. Grasshoff, K., Ehrardt, M. & Kremling, K. *Methods of Seawater Analysis* (Verlag Chemie, Weinheim, 1983).
30. Maldonado, M. & Uriz, M. J. An experimental approach to the ecological significance of microhabitat-scale movement in an encrusting sponge. *Mar. Ecol. Prog. Ser.* **185,** 239–255 (1999).

Learning the parts of objects by non-negative matrix factorization

Daniel D. Lee* & H. Sebastian Seung*†

* *Bell Laboratories, Lucent Technologies, Murray Hill, New Jersey 07974, USA*
† *Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA*

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 for non-negative matrix factorization that is able to learn parts of faces and semantic features of text. This is in contrast to other methods, such as principal components analysis and vector quantization, that learn holistic, not parts-based, representations. Non-negative matrix factorization is distinguished from the other methods by its use of non-negativity constraints. These constraints lead to a parts-based representation because they allow only additive, not subtractive, combinations. When non-negative matrix factorization is implemented as a neural network, 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 non-negative matrix factorization (NMF), together with principal components analysis (PCA) and vector quantization (VQ), to a database of facial images. As shown in Fig. 1, all three methods 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 methods in a matrix factorization framework. The image database is regarded as an $n \times m$ matrix V, each column of which contains n non-negative pixel values of one of the m facial images. Then all three methods construct approximate factorizations of the form $V \approx WH$, or

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

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. Although 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. As 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 non-negativity constraints permit the combination of multiple basis images to represent a face. But only additive combinations are allowed, because the non-zero elements of W and H are all positive. In contrast to PCA, no subtractions can occur. For these reasons, the non-negativity constraints are compatible with the intuitive notion of combining parts to form a whole, which is how NMF learns a parts-based representation.

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[7–9].

We implemented NMF with the update rules for W and H given in Fig. 2. Iteration of these update rules converges to a local maximum of the objective function

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

subject to the non-negativity constraints described above. This objective function can be derived by interpreting NMF as a method 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 equation (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 non-negativity constraints for the success of NMF in learning parts. A squared error objective function can be optimized with update rules for W and H different from those in Fig. 2 (refs 10, 11). These update rules yield results similar to those shown 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.
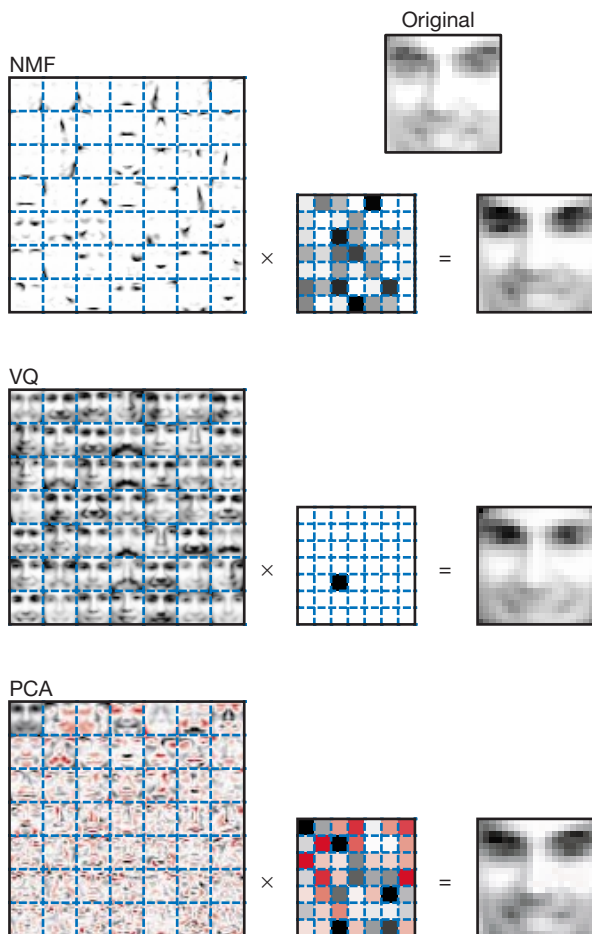

Figure 1 Non-negative matrix factorization (NMF) learns a parts-based representation of faces, whereas vector quantization (VQ) and principal components analysis (PCA) learn holistic representations. The three learning methods were applied to a database of $m = 2{,}429$ facial images, each consisting of $n = 19 \times 19$ pixels, and constituting an $n \times m$ matrix V. All three 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 method 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 superposition 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, NMF learns to represent faces with a set of basis images resembling parts of faces.

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, owing to the fan-out of connections from the encoding variable. Because of the non-negativity 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. NMF learns by adapting W_{ia} to generate the appropriate coactivations.

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 modelling the generation of directly observable visible variables V from hidden variables H (refs 12, 13). 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[14]. These word counts can be regarded as a set of visible variables and modelled as being generated from an underlying set of hidden variables. Application of VQ, PCA or NMF 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[14]. 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. Although PCA has been successful in certain linguistic tasks[15], 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. Because 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 semantic features ($r = 200$, 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, whereas it appears with 'person', 'rules' and 'law' in the other. This demonstrates that NMF can deal with the polysemy of 'lead' by disambiguating two of its meanings in the corpus of documents.

Although NMF is successful in learning facial parts and semantic topics, this success does not imply that the method 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. Although non-negativity constraints may help such models to learn parts-based representations[13], we do not claim that they are sufficient in themselves. Also, NMF does not learn anything about the 'syntactic' relationships between parts. NMF assumes that the hidden variables are non-negative, but makes no further assumptions about their statistical dependencies.

This is in contrast to independent components analysis (ICA), a variant of PCA that assumes that the hidden variables are statistically independent and non-gaussian[16,12]. Applying ICA to the facial images 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

$$W_{ia} \leftarrow W_{ia} \sum_{\mu} \frac{V_{i\mu}}{(WH)_{i\mu}} H_{a\mu}$$

$$W_{ia} \leftarrow \frac{W_{ia}}{\sum_j W_{ja}}$$

$$H_{a\mu} \leftarrow H_{a\mu} \sum_i W_{ia} \frac{V_{i\mu}}{(WH)_{i\mu}}$$

Figure 2 Iterative algorithm for non-negative matrix factorization. Starting from non-negative initial conditions for W and H, iteration of these update rules for non-negative V finds an approximate factorization $V \approx WH$ by converging to a local maximum of the objective function given in equation (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[22,23]. The update rules preserve the non-negativity 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 $W \rightarrow W\Lambda$, $H \rightarrow \Lambda^{-1}H$, where Λ is a diagonal matrix.
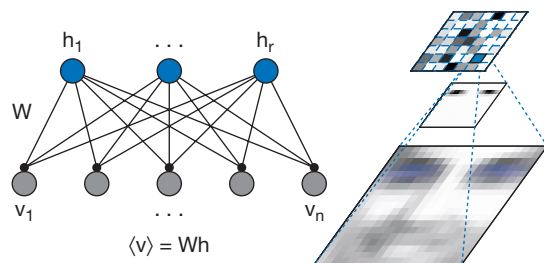

Figure 3 Probabilistic hidden variables model underlying non-negative matrix factorization. The model is diagrammed as a network depicting how the visible variables $v_1,\ldots,v_n$ in the bottom layer of nodes are generated from the hidden variables $h_1,\ldots,h_r$ in the top layer of nodes. According 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, whereas 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) which is combined with other basis images to represent a whole facial image (right bottom).

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, to make the images rather than the encodings as statistically independent as possible[18]. This results in a basis that is non-global; however, in this representation all the basis images 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 infers values for the hidden variables from the visible variables. Although the generative model of Fig. 3 is linear, the inference computation is nonlinear owing to the non-negativity constraints. The computation is similar to maximum likelihood reconstruction in emission tomography[19], and deconvolution of blurred astronomical images[20,21].

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 beyond the scope of this letter. Here we only point out the

consequence of the non-negativity constraints, which is that synapses are either excitatory or inhibitory, but do not change sign. Furthermore, the non-negativity of the hidden and visible variables corresponds to the physiological fact that the firing rates of neurons cannot be negative. We propose 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 greyscale 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. The 49 eigenvectors with the largest eigenvalues are displayed. 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 15,276 most frequent words in the database of Grolier encyclopedia articles, after removal of the 430 most common words, such as 'the' and 'and'. Because 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.

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. Field, D. J. What is the goal of sensory coding? *Neural Comput.* **6,** 559–601 (1994).
8. 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).
9. 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).
10. Lee, D. D. & Seung, H. S. Unsupervised learning by convex and conic coding. *Adv. Neural Info. Proc. Syst.* **9,** 515–521 (1997).
11. Paatero, P. Least squares formulation of robust non-negative factor analysis. *Chemometr. Intell. Lab.* **37,** 23–35 (1997).
12. Nakayama, K. & Shimojo, S. Experiencing and perceiving visual surfaces. *Science* **257,** 1357–1363 (1992).
13. Hinton, G. E., Dayan, P., Frey, B. J. & Neal, R. M. The "wake-sleep" algorithm for unsupervised neural networks. *Science* **268,** 1158–1161 (1995).
14. Salton, G. & McGill, M. J. *Introduction to Modern Information Retrieval* (McGraw-Hill, New York, 1983).
15. Landauer, T. K. & Dumais, S. T. The latent semantic analysis theory of knowledge. *Psychol. Rev.* **104,** 211–240 (1997).
16. Jutten, C. & Herault, J. Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture. *Signal Proc.* **24,** 1–10 (1991).
17. Bell, A. J. & Sejnowski, T. J. An information maximization approach to blind separation and blind deconvolution. *Neural Comput.* **7,** 1129–1159 (1995).
18. Bartlett, M. S., Lades, H. M. & Sejnowski, T. J. Independent component representations for face recognition. *Proc. SPIE* **3299,** 528–539 (1998).
19. Shepp, L. A. & Vardi, Y. Maximum likelihood reconstruction for emission tomography. *IEEE Trans. Med. Imaging.* **2,** 113–122 (1982).
20. Richardson, W. H. Bayesian-based iterative method of image restoration. *J. Opt. Soc. Am.* **62,** 55–59 (1972).
21. Lucy, L. B. An iterative technique for the rectification of observed distributions. *Astron. J.* **74,** 745–754 (1974).
22. Dempster, A. P., Laired, N. M. & Rubin, D. B. Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Stat. Soc.* **39,** 1–38 (1977).
23. Saul, L. & Pereira, F. *Proceedings of the Second Conference on Empirical Methods n Natural Language Processing* (eds Cardie, C. & Weischedel, R.) 81–89 (Morgan Kaufmann, San Francisco, 1997).
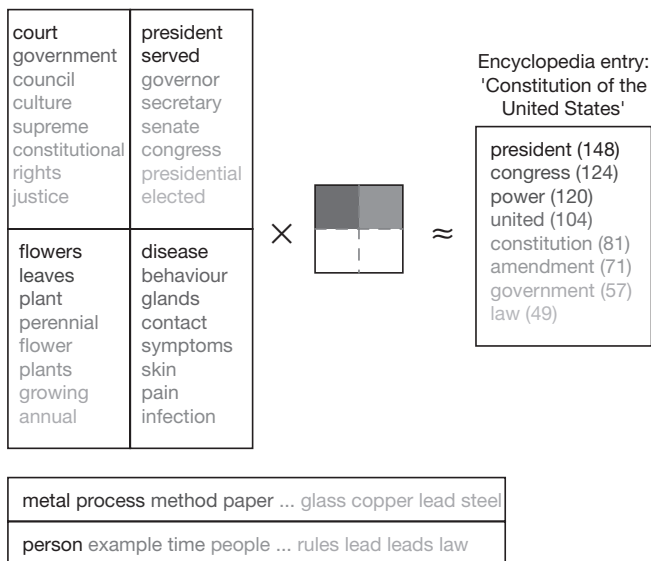
Figure 4 Non-negative matrix factorization (NMF) discovers semantic features of $m = 30,991$ articles from the Grolier encyclopedia. For each word in a vocabulary of size $n = 15,276$, the number of occurrences was counted in each article and used to form the $15,276 \times 30,991$ matrix V. Each column of V contained the word counts for a particular article, whereas 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 algorithm described in Fig. 2. Upper left, four of the $r = 200$ semantic features (columns of W). As 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. Right, 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 NMF.