TOWARDS COGNITIVE COMPONENT ANALYSIS

Lars Kai Hansen, Peter Ahrendt, and Jan Larsen

Intelligent Signal Processing,
Informatics and Mathematical Modelling,
Technical University of Denmark B321,
DK-2800 Kgs. Lyngby, Denmark

ABSTRACT

Cognitive component analysis (COCA) is here defined as the process of unsupervised grouping of data such that the ensuing group structure is well-aligned with that resulting from human cognitive activity. We have earlier demonstrated that independent components analysis is relevant for representing semantics, not only in text, but also in dynamic text (chat), images, and combinations of text and images. Here we further expand on the relevance of the ICA model for representing context, including two new analyzes of abstract data: social networks and musical features.

1. INTRODUCTION

In this paper our aim is to discuss the generality of the so-called *independent component hypothesis*. It is well documented that human perceptional systems can model complex multi-agent scenery. Human cognition uses a broad spectrum of cues for analyzing perceptual input and separate individual signal producing agents, such as speakers, gestures, affections etc. Unsupervised signal separation has also been achieved in computers using a variety of independent component analysis algorithms [1]. It is an intriguing fact that representations are found in human and animal perceptual systems which closely resembles the information theoretically optimal representations obtained by independent component analysis, see e.g., [2] on visual contrast detection, [3] on visual features involved in color and stereo processing, and [4] on representations of sound features. Here we go one step further and ask: *Are such optimal representation rooted in independence also relevant in higher cognitive functions*? Our presentation is largely qualitative and will mainly be based on simple visualizations of data and avoid unnecessary algebraic complication.

Brittanica online defines cognition as the 'act or process of knowing', and continues:

> Cognition includes every mental process that may be described as an experience of knowing (including perceiving, recognizing, conceiving, and reasoning), as distinguished from an experience of feeling or of willing.

Wagensberg has recently argued the importance of being able to recognize independence for successful 'life forms' [5]

> A living individual is part of the world with some identity that tends to become independent of the uncertainty of the rest of the world

Thus natural selection favors innovations that increase independence of the agent in the face of environmental uncertainty, while maximizing the gain from the predictable aspects of the niche. This view represents a precision of the classical Darwinian formulation that natural selection simply favors adaptation to given conditions. Wagensberg points out that recent biological innovations, such as nervous systems and brains are means to decrease the sensitivity to un-predictable fluctuations. Furthermore, by creating alliances, agents can in Wagensberg's picture give up independence for the benefit of a group, which in turns may increase independence for the group as an entity. Both in its simple one-agent form and in the more tentative analysis of the group model, Wagensberg's theory points to the crucial importance of *statistical independence* for evolution of perception, semantics and indeed cognition.

While cognition may be hard to quantify, its direct consequence, human behavior, has a rich phenomenology which is becoming increasingly accessible to modeling. The digitalization of everyday life as reflected, say, in telecommunication, commerce, and media usage allows quantification and modeling of human patterns of activity, often at the level of individuals.

Grouping of events or objects in categories is fundamental to human cognition. In machine learning, classification is a rather well-understood task when based on *labelled* examples [6]. In this case classification belongs to the class of *supervised* learning problems. Clustering is a closely related *unsupervised* learning problem, in which we use general statistical rules to group objects, without a priori providing a set of labelled examples. It is a fascinating finding in many real world data sets that the label structure discovered by unsupervised learning closely coincides with labels obtained by letting a human or a group of humans perform classification, labels derived from human cognition. *Here we will define cognitive component analysis (COCA) as the process of unsupervised group-*

ing of data such that the ensuing group structure is well-aligned with that resulting from human cognitive activity. Without directly using the phrase 'cognitive component analysis', the concept of cognitive components appears frequently in the context of Factor analysis of behavioral studies, see e.g., [7, 8].

We have pursued grouping by independent component analysis in several abstract data types including text, dynamic text (chat), images, and combinations hereof, see e.g., [9, 10, 11, 12, 13]. In this presentation we will briefly review our analysis of text data and add visualizations of two new types of abstract data, namely co-worker networks and music, that further underlines the broad relevance of the independent component hypothesis.

2. COGNITIVE COMPONENT ANALYSIS

In 1999 Lee and Seung introduced the method of non-negative matrix factorization (NMF) [14] as a scheme for parts-based object recognition. They argued that the factorization of an observation matrix in terms of a relatively small set of cognitive components, each consisting of a feature vector and a loading vector (both non-negative) lead to a parts based object representation. They demonstrated this for objects in images and in text representations. More recently, in 2002, it was shown that very similar parts-based decompositions were obtained in a latent variable model based on positive linear mixture of positive *independent* source signals [15]. Holistic, but parts-based, recognition of objects is frequently reported in perception studies across multiple modalities and increasingly in abstract data, where object recognition is a cognitive process. Together these findings are often referred to as instances of the more general *Gestalt laws*.

2.1. Latent semantic indexing (LSI)

Salton proposed the so-called vector space representation for statistical modeling of text data, for a review see [16]. A term set is chosen and a document is represented by the vector of term frequencies. A document database then forms a so-called term-document matrix. The vector space representation can be used for classification and retrieval by noting that similar documents are somehow expected to be 'close' in the vector space. A metric can be based on the simple Euclidean distance if document vectors are properly normalized, otherwise angular distance may be useful. This approach is principled, fast, and language independent. Deerwester and co-workers developed the concept of latent semantics based on principal component analysis of the term-document matrix [17]. The fundamental observation behind the latent semantic indexing (LSI) approach is that similar documents are using similar vocabularies, hence, the vectors of a given topic could appear as produced by a stochastic process with highly correlated term-entries. By projecting the term-frequency vectors on a relatively low dimensional subspace, say determined by the maximal amount of variance one would be able to filter out the inevitable 'noise'. Noise should here be thought of as individual document differences in term usage within a specific context. For well-defined topis, one could simply hope that a given context would have a stable core term set that would come out as a 'direction' in the term vector space. Below we will explain why this is likely not to happen in general document databases, and LSI is therefore often used as a dimensional reduction tool, which is then post-processed to reveal cognitive components, e.g., by interactive visualization schemes [18].

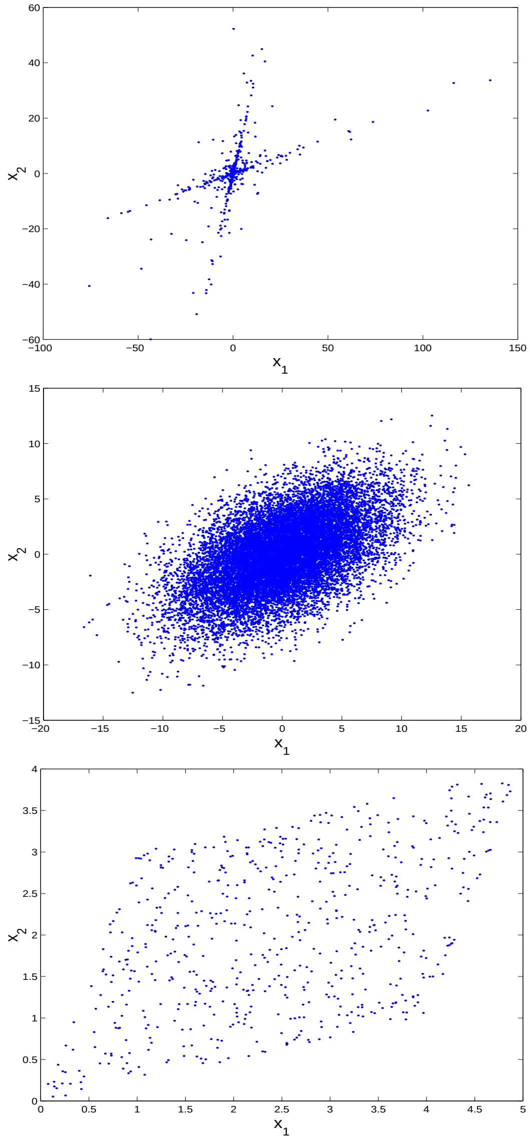

Figure 1. Prototypical feature distributions produced by a linear mixture, based on sparse (top), normal (middle), or dense source signals (bottom), respectively. The characteristic of the sparse signal is that it consists of relatively few large magnitude samples on a background of small signals.

2.2. Non-negative matrix factorization (NMF)

Noting that many basic feature sets are naturally positive and that a non-negative decomposition could lead to a parts-based decomposition, Lee and Seung analyzed several data sets using the NMF decomposition technique

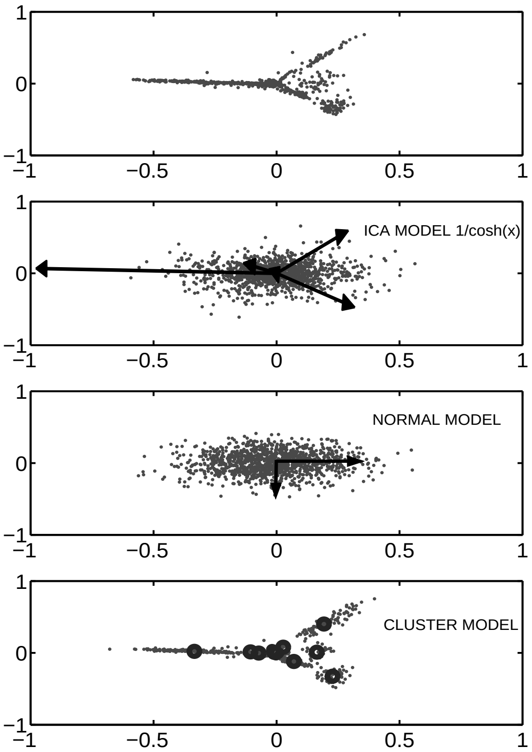

Figure 2. Latent semantic analysis of text, based on Salton's vector space representation reveals that the semantic components are very sparse, as indicated in the scatter plot of two components (top). We spot the signature of a sparse linear mixture: 'rays' emanating from $(0, 0)$. Performing a five component ICA on the subspace spanned by the five most significant latent vectors, provide the mixing matrix with column vector as shown in the second panel. Using a simple classification scheme (magnitude of the source signal) yields a classifier with less then 10% error rate using the document labels manually assigned by a human editor. Below, in the third plot, we indicate the corresponding normal model, with axis aligned latent vectors. Finally, we show in the bottom plot the results of an alternative unsupervised analysis, based on clustering, using a Gaussian mixture model. While the mixture model do capture the density well, the ensuing components are not related in a simple way to content.

[14]. A basic difficulty of the approach is the possible non-uniqueness of the components. This issue has been discussed in detail by Donoho and Stodden [19]. A possible route to more unique solutions, hence, potentially more interpretable and relevant components is to add a priori knowledge, e.g., in form of independence assumptions. An algorithm for decomposing independent positive components from a positive mixture is discussed in [15].

2.3. Independent component analysis (ICA)

Blind signal separation is the general problem of recovering source signals from an unknown mixture. This aim is in general not feasible without additional information. If we assume that the unknown mixture is linear, i.e., that the mixture is a linear combination of the sources, and furthermore assume that the sources are statistically independent processes it is often possible to recover sources and mixing, using a variety of independent component analysis techniques [1]. Here we will discuss some basic characteristics of mixtures and the possible recovery of sources.

First, we note that LSI/PCA can not do the job in general. Let the mixture be given as

$$\mathbf{X} = \mathbf{AS}, \;\; X_{j,t} = \sum_{k=1}^{K} A_{j,k} S_{k,t}, \tag{1}$$

where $X_{j,t}$ is the value of j'th feature in the t'th measurement, $A_{j,k}$ is the mixture coefficient linking feature j with the component k, while $S_{k,t}$ is the level of activity in the k'th source. In a text instance a feature is a term and the measurements are documents, the components are best thought as topical contexts. The k'th column $A_{j,k}$ holds the relative frequencies of term occurrence in documents within context k. The source matrix element $S_{k,t}$ quantifies the level of expression of context k in document t.

As a linear mixture is invariant to an invertible linear transformation we need define a normalization of one of the matrices $\mathbf{A}, \mathbf{S}$. We will do this by assuming that the sources are unit variance. As they are assumed independent the covariance will be trivial,

$$\Sigma_S = \lim_{T\to\infty} \frac{1}{T} \mathbf{SS}^\top = \mathbf{I}. \tag{2}$$

LSI, hence PCA, of the measurement matrix is based on analysis of the covariance

$$\Sigma_X = \lim_{T\to\infty} \frac{1}{T} \mathbf{XX}^\top = \mathbf{AA}^\top. \tag{3}$$

Clearly the information in $\mathbf{AA}^\top$ is not enough to uniquely identify $\mathbf{A}$, since if a solution $\mathbf{A}$ is found, any (row) rotated matrix $\tilde{\mathbf{A}} = \mathbf{AU}, \mathbf{UU}^\top = \mathbf{I}$ is also a solution, because $\tilde{\mathbf{A}}$ has the same outer product as $\mathbf{A}$.

This is a potential problem for LSI based analysis. If the document database can be modelled as in eq. (1) then the original characteristic context histograms will not be found by LSI. The field of independent component analysis has on the other hand devised many algorithms that use more informed statistics to locate $\mathbf{A}$ and thus $\mathbf{S}$, see [1] for a recent review.

The histogram of a source signal can roughly be described as sparse, normal, or dense. Scatter plots of projections of mixtures drawn from source distributions with one of these three characteristics are shown in Figure 1. In the upper panel of Figure 1 we show the typical appearance of a sparse source mixture. The sparse signal consists of relatively few large magnitude samples in a background of a large number of small signals. When mixing such independent sparse signals as in Eq. (1), we obtain a set of rays emanating from origo. The directions of the rays are directly given by the column vectors of the $\mathbf{A}$-matrix.

If the sources are truly normal distributed like in the middle panel of Figure 1, there is no additional information but the covariance matrix. Hence, in some sense this is a singular worst case for separation. Because we work from finite samples an ICA method, which assumes some non-normality, will in fact often find good approximations to the mixing matrix, simply because a finite normal sample will have non-normal oddities. But fortunately, many, many interesting real world data sets are not anywhere near normal, rather they are typically very sparse, hence, more similar to the upper panel of Figure 1.

3. COGNITIVE COMPONENTS FROM UNSUPERVISED DATA ANALYSIS

Having argued that tools are available for recovering, relatively uniquely, the underlying components in a mixture we now turn to some illustrative examples. In a text analysis example we show that an ICA based analysis indeed finds a small set of semantic components that very well aligned with human assigned labels that were not used in the analysis.

3.1. Text analysis

In Figure 2 (top) we indicate the corresponding scatter plots of a small text database. The database consists of documents with overlapping vocabulary but five different (high level cognitive) labels [20]. The 'ray'-structure is evident. In the second panel we show the directions identified by ICA. If we use a simple projection based classification rule, and associate a ray with a topic, the classification error rate is less than 10% [20]. If an ICA is performed with less components, the topics with close content are merged.

This rather striking alignment between human and machine classification in abstract features like those of vector space text analysis, is a primary motivation for the present work. In this example we also estimated an alternative unsupervised model based on document clustering using a gaussian mixture model. This model provides the representation shown in bottom panel of Figure 2, in this case the clusters are not specific enough to have a simple one-to-one correspondence, however, with a limited amount of supervision it will be possible to convert this cluster based representation into a classifier with similar performance as the ICA model.

3.2. Social networks

The ability to navigate social networks is a hallmark of successful cognition. Is it possible that the simple unsupervised scheme for identification of independent components, whose relevance we have established above for perceptual tasks, for context grouping in different media, could play a role in this human capacity? To investigate this issue we have initiated an analysis of a well-known social network of some practical importance. The so-called *actor network* is a quantitative representation of the co-participation of actors in movies, for a discussion of this network, see e.g., [21]. The observation model for the network is not too different from that of text. Each movie is represented by the *cast*, i.e., the list of actors. We have converted the table of the about $T = 128.000$ movies with a total of $J = 382.000$ individual actors, to a sparse $J \times T$ matrix $\mathbf{X}$. For visualization we have projected the data onto principal components (LSI) of the actor-actor co-variance matrix. The eigenvectors of this matrix are called 'eigen casts' and represent characteristic communities of actors that tend to co-appear in movies. The sparsity and magnitude of the network means that the components are dominated by communities with very small intersections, however, a closer look at such scatter plots reveals detail suggesting that a simple linear mixture model indeed provides a reasonable representation of the (small) coupling between these relative trivial disjunct subsets, see Figure 3.

Such insight may be used for computer assisted navigation of collaborative, peer-to-peer networks, for example in the context of search and retrieval.

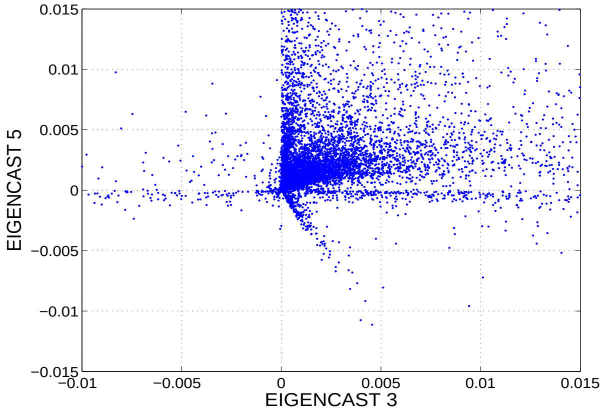

Figure 3. The so-called actor network quantifies the collaborative pattern of 382.000 actors participating in almost 128.000 movies. For visualization we have projected the data onto principal components (LSI) of the actor-actor co-variance matrix. The eigenvectors of this matrix are called 'eigencasts' and they represent characteristic communities of actors that tend to co-appear in movies. The network is extremely sparse, so the most prominent variance components are related to near-disjunct subcommunities of actors with many common movies. However, a close up of the coupling between two latent semantic components (the region $\sim (0, 0)$) reveals the ubiquitous signature of a sparse linear mixture: A pronounced 'ray' structure emanating from (0,0). We speculate that the cognitive machinery developed for handling of independent events can also be used to locate independent subcommunities, hence, navigate complex social networks, a hallmark of successful cognition.

3.3. Musical genre

The growing market for digital music and intelligent music services creates an increasing interest in modeling of music data. It is now feasible to estimate consensus musical genre by *supervised* from rather short music segments, say 10-30 seconds, see e.g., [22], thus enabling computerized handling of music request at a high cognitive complexity level. To understand the possibilities and limitations for unsupervised modeling of music data we here visualize a small music sample using the latent semantic analysis framework. The intended use is for a music search engine function, hence, we envision that a largely text based query has resulted in a few music entries, and the algorithm is going to find the group structure inherent in the retrieval for the user. We represent three tunes (with human labels: `heavy, jazz, classical`) by their spectral content in overlapping small time frames ($w = 30$msec, with an overlap of 10msec, see [22], for details). To make the visualization relatively independent of 'pitch', we use the so-called mel-cepstral representation (MFCC, $K = 13$ coefficients pr. frame). To reduce noise in the visualization we have 'sparsified' the amplitudes. This was achieved simply by retaining only coefficients that belonged to the upper 5% magnitude fractile. The total number of frames in the analysis was $F = 10^5$. PCA provided unsupervised latent semantic dimensions and a scatter plot of the data on the subspace spanned ny two such dimensions is shown in Figure 4. For interpretation we have coded the data points with signatures of the three genres involved. The ICA ray-structure is striking, however, we note that the situation is not one-to-one as in the small text databases. A component quantifies a characteristic 'theme' at the temporal level of a frame (30msec), it is an issue for further research whether genre *recognition* can be done from the salient themes, or we need to combine more than one theme to reach the classification performance obtained in [22] for $10-30$ second un-structured frame sets.

4. CONCLUSION

Cognitive component analysis (COCA) was defined as the process of unsupervised grouping of data such that the ensuing group structure is well-aligned with that resulting from human cognitive activity. It is well-established that information theoretically optimal representations, similar to those found by ICA, are in use in several information processing tasks in human and animal perception. By visualization of data using latent semantic analysis-like plots, we have shown that independent components analysis is also relevant for representing semantic structure, in text and also in other abstract data such as social networks, and musical features. We therefore speculate that the cognitive machinery developed for analyzing complex perceptual signals from multi-agent environments may also be used in higher brain function, such as understanding music or navigation of complex social networks, a hallmark of successful cognition. Hence, independent component analysis given the right representation may be a quite generic tool for COCA.

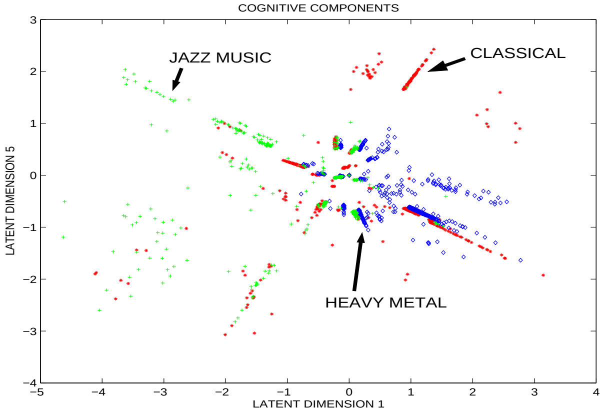

Figure 4. We represent three music tunes (with labels: `heavy metal, jazz, classical`) by their spectral content in overlapping small time frames ($w = 30$msec, with an overlap of 10msec, see [22], for details). To make the visualization relatively independent of 'pitch', we use the so-called mel-cepstral representation (MFCC, $K = 13$ coefficients pr. frame). To reduce noise in the visualization we have 'sparsified' the amplitudes. This was achieved simple by keeping coefficients that belonged to the upper 5% magnitude fractile. The total number of frames in the analysis was $F = 10^5$. Latent semantic analysis provided unsupervised subspaces with maximal variance for a given dimension. We show the scatter plot of the data on a 2D subspace within an original 5D PCA. For interpretation we have coded the data points with signatures of the three genres involved: classical ($*$), heavy metal (diamond), jazz ($+$). The ICA ray-structure is striking, however, note that the situation is not one-to-one (ray to genre) as in the small text databases. A component (ray) quantifies a characteristic musical 'theme' at the temporal level of a frame (30msec), i.e., an entity similar to the 'phoneme' in speech.

5. ACKNOWLEDGMENTS

This work is supported by the Danish Technical Research Council, through the framework project 'Intelligent Sound', www.intelligentsound.org (STVF No. 26-04-0092).

6. REFERENCES

[1] A. Hyvarinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, John Wiley & Sons, 2001.

[2] Anthony J. Bell and Terrence J. Sejnowski, "The 'independent components' of natural scenes are edge filters," *Vision Research*, vol. 37, no. 23, pp. 3327–3338, 1997.

[3] Patrik Hoyer and Aapo Hyvrinen, "Independent component analysis applied to feature extraction from colour and stereo images.," *Network: Comput. Neural Syst.*, vol. 11, no. 3, pp. 191–210, 2000.

[4] M.S. Lewicki, "Efficient coding of natural sounds,"

Nature Neuroscience, vol. 5, no. 4, pp. 356–363, 2002.

[5] Jorge Wagensberg, “Complexity versus uncertainty: The question of staying alife,” *Biology and philosophy*, vol. 15, pp. 493–508, 2000.

[6] C.M. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, Oxford, 1995.

[7] David Rawlings, Neus Barrantes-Vidal, Gordon Claridge, Charles McCreery, and Georgina Galanos, “A factor analytic study of the hypomanic personality scale in british, spanish and australian samples.,” *Personality and Individual Differences*, vol. 28, pp. 73–84, 2000.

[8] C.H. Waechtler, E.A. Zillmer, M.J. Chelder, and B. Holde, “Neuropsychological patterns of component factor scores from the positive and negative syndrome scale (panss) in schizophrenics.,” *Archives of Clinical Neuropsychology (Abstracts)*, vol. 10, pp. 400, 1995.

[9] L. K. Hansen, J. Larsen, and T. Kolenda, “On independent component analysis for multimedia signals,” in *Multimedia Image and Video Processing*, pp. 175–199. CRC Press, sep 2000.

[10] L. K. Hansen, J. Larsen, and T. Kolenda, “Blind detection of independent dynamic components,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing 2001*, 2001, vol. 5, pp. 3197–3200.

[11] T. Kolenda, L. K. Hansen, and J. Larsen, “Signal detection using ICA: Application to chat room topic spotting,” in *Third International Conference on Independent Component Analysis and Blind Source Separation*, 2001, pp. 540–545.

[12] T. Kolenda, L. K. Hansen, J. Larsen, and O. Winther, “Independent component analysis for understanding multimedia content,” in *Proceedings of IEEE Workshop on Neural Networks for Signal Processing XII*, H. Bourlard et al. Ed., Piscataway, New Jersey, 2002, pp. 757–766, IEEE Press, Martigny, Valais, Switzerland, Sept. 4-6, 2002.

[13] J. Larsen, L.K. Hansen, T. Kolenda, and F.AA. Nielsen, “Independent component analysis in multimedia modeling,” in *Fourth International Symposion on Independent Component Analysis and Blind Source Separation*, Shun ichi Amari et al. Ed., Nara, Japan, apr 2003, pp. 687–696, Invited Paper.

[14] D.D. Lee and H.S. Seung, “Learning the parts of objects by non-negative matrix factorization,” *Nature*, vol. 401, pp. 788–791, 1999.

[15] Pedro A. D. F. R. Højen-Sørensen, Ole Winther, and Lars Kai Hansen, “Mean-field approaches to independent component analysis,” *Neural Comput.*, vol. 14, no. 4, pp. 889–918, 2002.

[16] Gerard Salton, *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer.*, Addison-Wesley, 1989.

[17] Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman, “Indexing by latent semantic analysis.,” *JASIS*, vol. 41, no. 6, pp. 391–407, 1990.

[18] T.K̃. Landauer, D. Laham, and M. Derr, “From paragraph to graph: latent semantic analysis for information visualization.,” *Proc Natl Acad Sci*, vol. 101, no. Sup. 1, pp. 5214–5219, 2004.

[19] David L. Donoho and Victoria Stodden, “When does non-negative matrix factorization give a correct decomposition into parts?,” in *NIPS*, 2003.

[20] T. Kolenda, L. K. Hansen, and S. Sigurdsson, “Indepedent components in text,” in *Advances in Independent Component Analysis*, pp. 229–250. Springer-Verlag, 2000.

[21] A.-L. Barabasi and R. Albert, “Emergence of scaling in random networks.,” *Science*, vol. 286, pp. 509–512, 1999.

[22] P. Ahrendt, A. Meng, and J. Larsen, “Decision Time Horizon For Music Genre Classification Using Short Time Features,” in *EUSIPCO*, Vienna, Austria, sep 2004, pp. 1293–1296.