Independent Components in Text

Thomas Kolenda and Lars Kai Hansen
Department of Mathematical Modelling
Technical University of Denmark
DK-2800 Lyngby, DENMARK
thko,lkhansen@imm.dtu.dk

Abstract

In this communication we analyze the feasibility of independent component analysis (ICA) for dimensional reduction and representation of word histograms. The analysis is carried out in a likelihood framework which allows estimates of the loadings (source signals), the mixing matrix and the noise level. In the face of noisy signals, the estimated sources are non-linear functionals of the observed signals, in contrast to the linear noise free case. We also discuss the generalizability of the estimated models and show that an empirical test error estimate may be used to optimize model dimensionality, in particular the optimal number of sources. When applied to word histograms ICA is shown to produce representations that are better aligned with the group structure in the text data than the LSA.

1 Background

Pattern recognition in text data is based on statistical representations of documents. In the face of limited sets of labeled data pattern recognition algorithms typically fail to generalize in high dimensions, and there is a need for efficient and robust means for data reduction. In [4] the Latent Semantic Analysis (LSA) approach was defined. LSA is based on summarization of the *term by document matrix*, i.e., a count of how often a given set of terms occur in the set of documents under analysis. The list of terms is adaptive and derived, e.g., by words that occur with a certain minimum frequency, in several documents, and possibly screened by a list of simple high-frequency *stop words*. In LSA term occurrence histograms are projected on a orthogonal set of "eigen-histograms" found by singular value decomposition. LSA can aid interpretation by visualizing group structure in the set of documents, typically by scatterplots of the term histograms on a few salient eigen-histograms, see figure 1. The first observation is that while there is notable grouping of the documents, this structure is spanned by the eigen-histograms, but not "explained" by these. The apparent cone structure also noted in [4], points to a representation based on a non-orthogonal basis set.

In this paper we show that independent component analysis (ICA) is a viable tool for identification of such a basis set. We apply a new ICA algorithm which is able to identify a generalizable low-dimensional basis set in the face of high-dimensional noisy data [5]. The term by document matrix is considered a linear mixture of a set of independent sources each activating its characteristic semantic network. The semantic networks take the form

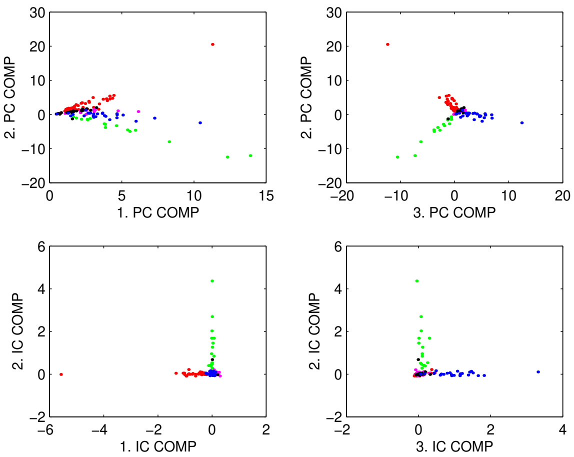

Figure 1: Analysis of the MED set of medical abstracts, labeled in five classes here coded in colors. The two upper panels show scatterplot of documents in the Latent Semantic or Principal Component basis. In the lower panels we show the document location as seen through the ICA representation. Note that while the group structure is clearly visible in the PCA plots, only in the ICA plots is the group structure aligned with independent components.

non-orthogonal term occurrence histograms. Translating to the more conventional use of ICA for speech separation, terms play the role of "microphones" and the document index corresponds to the time index, while the independent components (sources) corresponds to speakers.

1.1 Independent Component Analysis

Reconstruction of statistically independent components / sources from linear mixtures is relevant to many information processing contexts, see e.g. [1] for an introduction and a recent review. This contribution is to our knowledge the first application to the realm of textmining. We will derive a solution to the source separation based on the likelihood formulation, see e.g., [2, 8, 7]. The specific model investigated here is a special case of the general framework proposed by Belouchrani and Cardoso [2], however we formulate the parameter estimation problem in terms of the Boltzmann learning rule, which allows for a particular transparent derivation of the mixing matrix estimate. We focus on generalizability of the ICA representation, and use the generalization error as a means for optimizing the complexity of the representation.

Let the observed mixture signals be denoted X, a matrix of size $T \times N$, where T is the number terms in the word histogram and N is the number of documents. The noisy mixing model takes the form,

$$X = AS + U, \tag{1}$$

where S is the source signal matrix (size $M \times N$, M is the number of sources), A is the $T \times M$ mixing matrix, while U is a matrix of noise signals with a parameterized distribution. The properties of the source signals are introduced by a parameterized prior distribution $P(S|\psi)$. The likelihood of the parameters of the noise distribution, the parameters of the source distribution and of the mixing matrix is given by,

$$L(A,\theta,\psi) = P(X|A,\theta,\psi) = \int P(X - AS|\theta)P(S|\psi)dS, \tag{2}$$

where $P(.|\theta)$ is the parameterized noise distribution. We assumed i.i.d. sources in Eq. 2 and we will assume also that the noise can modeled by i.i.d. Gaussian variables with variance $\theta = \sigma^2$,

$$P(U|\sigma^2) = \frac{1}{(2\pi\sigma^2)^{TN/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{t,d} U_{t,d}^2\right). \tag{3}$$

where t, d index terms and document variables. Finally, we will assume the parameter free source distribution of [7],

$$P(S) = \frac{1}{\pi^{NM}} \exp\left(-\sum_{m,d} \log\cosh S_{m,d}\right). \tag{4}$$

The sum runs over all sources ($m = 1, ...M$) and documents ($d = 1, ..., N$).

Let us first address the problem of estimating the sources if the mixing parameters are known, i.e., for given A, σ^2. We use Bayes formula $P(S|X) \propto P(X|S)P(S)$ to obtain a the posterior distribution of the sources

$$P(S|X, A, \sigma^2) \propto \exp\left(-\frac{1}{2\sigma^2}\sum_{t,d}(X - AS)_{t,d}^2) - \sum_{m,d} \log\cosh S\right). \tag{5}$$

The *maximum a posteriori* (MAP) source estimate is found by maximizing this expression w.r.t. S, providing the following non-linear equation to solve iteratively for the MAP estimate $\widehat{S}$,

$$-A^\top A\widehat{S} + A^\top X - \sigma^2 \tanh \widehat{S} = 0. \tag{6}$$

Since the likelihood is of the hidden-Gibbs form we can use a generalized Boltzmann learning rule to find the gradients of the likelihood of the parameters A, σ^2, These averages can be estimated in a mean field approximation [9, 5] leading to recursive rules for A and σ^2,

$$\widehat{A} = X\widehat{S}^\top \left(\widehat{S}\widehat{S}^\top + \beta\mathbf{1}\right)^{-1} \tag{7}$$

$$\widehat{\sigma^2} = \frac{1}{TN}\mathrm{Tr}_t(X - \widehat{A}\widehat{S})^\top(X - \widehat{A}\widehat{S}) \tag{8}$$

β is a lumped effect of fluctuations neglected in the mean field approach. Fluctuation corrections (hence the magnitude of β) can derived in the low noise limit, based on a Gaussian approximation to the likelihood [5].

1.2 Generalization and the Bias-Variance Dilemma

The parameters of our blind separation model are estimated from a finite random sample, and therefore they too are random variables inheriting noise from the data set they were trained on. Within the likelihood formulation the generalization error of a specific set of parameters is given by the average negative log-likelihood,

$$\Gamma(A,\theta,\psi) = \int P_*(X)[-\log \int P(X - AS|\theta)P(S|\psi)dS]dX. \tag{9}$$

$P_*(X)$ is the true distribution of data. The generalization error is a principled tool for model selection. In the context of blind separation the optimal number of sources retained

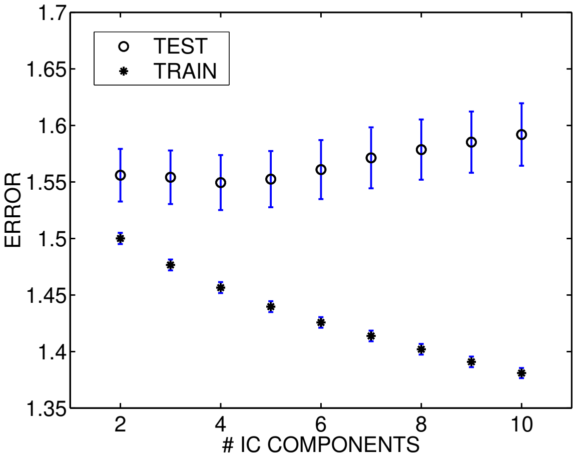

Figure 2: ICA analysis of the MED set of medical abstracts. Training and test errors as function of the dimensionality of the mixing matrix. The generalization error shows a shallow minimum for four independent components, reflecting a bias-variance tradeoff as function of the complexity of the estimated mixing matrix.

in the model is of crucial interest. We face a typical bias-variance dilemma [3]. If too few components are used, a structured part of the signal will be lumped with the noise, hence leading to a high generalization error because of "lack of fit". On the other hand, if too many sources are used we expect "overfit" since the model will use the additional degrees of freedom to fit non-generic details in the training data. The generalization error in Eq. (9) can be estimated using a test set of data independent of the training set.

1.3 Learning ICA text representations

Since we typically face problems with thousands of words in the terms list and a possibly much fewer documents, we face a so-called extremely ill-posed learning problem which can be "cured" without loss of generality by PCA projection. The PCA decompose the term by document matrix on eigen-histograms. These eigen-histograms are subject to an orthogonality constraint being eigenvectors to a symmetric real matrix. We are interested in a slightly more general separation of sources that are independent as sequences, but not necessarily orthogonal in the word histogram, i.e., we would like to be able to perform a more general decomposition of the data matrix, corresponding to the model in Eq. (1). Before performing the ICA we can make use of the PCA for simplification of the ICA problem. The approach taken here is similar to the so-called "cure for extremely ill-posed learning" [6] used to simplify supervised learning in short image sequences. We first note that the likelihood, considered a function of the columns of A (histograms) can be split in two parts. A part A_1 orthogonal to the subspace spanned by the N rows of X, and a part A_2 situated in the subspace spanned by the N columns of X. The first is part trivially minimized for any non-zero configuration of sources by putting $A_1 = 0$. It simply does not "couple" to data. The remaining part A_2 can be projected onto an N-dimensional hyperplane spanned by the documents. In this way we reduce the high-dimensional separation problem to the separation of a square (projected) datamatrix of size $N \times N$. We note that

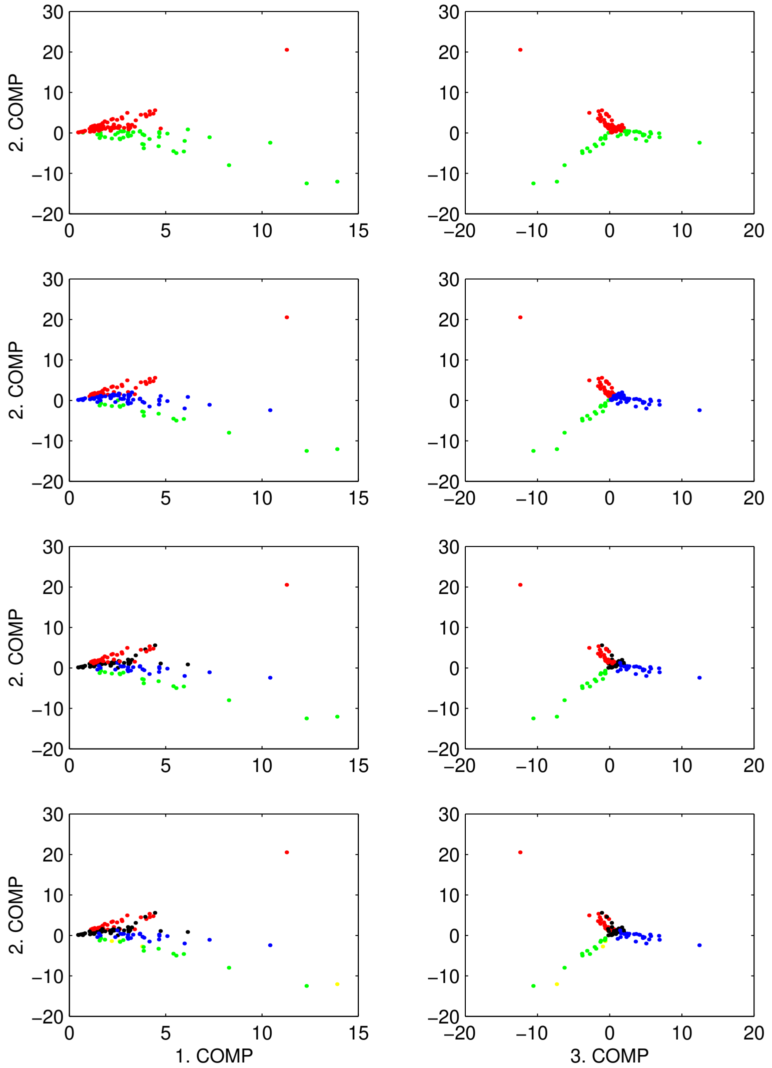

Figure 3: The MED dataset of medical abstracts. The dataset consists of 124 documents in five topics. The “source signals” recovered in the ICA has been converted to a simple classifier, and we have coded these classes by different colors. From top to bottom we show scatterplots in the principal component representation 1 vs 2 and 3 vs. 2., with colors signifying the classification proposed by the ICA with 2,3,4,5 independent components respectively.

it often may be possible to further limit the dimensionality of the PCA subspace, hence further reducing the histogram dimensionality T of the remaining problem.

2 Application to the MED dataset

The MED dataset is a commonly studied collection of medical abstracts [4]. In total it consists of 1033 abstracts, of which 30 group labels have been applied to 696 of the documents. We selected a subset consisting of 124 abstracts corresponding to the first 5 groups in the dataset. When constructing the histogram term by document matrix, words that occurred in two or more abstracts where chosen as a term word. Common words were removed by a stop list leaving a total of 1159 terms. The abstracts chosen are characterized briefly as: 1) The crystalline lens in vertebrates, including humans. 2) The relationship of blood and cerebrospinal fluid oxygen concentrations or partial pressures. A method of interest is polarography. 3) Electron microscopy of lung or bronchi. 4) Tissue culture of lung or bronchial neoplasms. 5) The crossing of fatty acids through the placental barrier. Normal fatty acid levels in placenta and fetus.

2.1 Results

We first represent the datamatrix using the “cure for extremely ill-posed learning” method, reducing the learning problem to a 124×124 problem without loss of generality. However, we expect that even fewer components are needed for creating a generalizable model. In Figure 2 we show the test and training set errors evaluated on training sets of 104 patterns randomly chosen among the set of 124. The test set consits of the remaining 20 documents in each resample (ten-fold crossvalidation). The generalization error shows a shallow minimum for four independent components, reflecting a bias-variance tradeoff as function of the complexity of the estimated mixing matrix. In Figure 1 we show scatterplots in the most prominent principal components and by the most variant independent components. While the distribution of documents forms rather well-defined group structure in the PCA scatterplots, clearly the ICA scatterplots are much better axis aligned. We conclude that the non-orthogonal basis found by ICA better “explains” the group structure. To further illustrate this finding we have converted the ICA solution to a pattern recognition device by a simple heuristic. We assign a group label based on the magnitude of the recovered source signal. In Table 1 and 2 we show that this device is quite successful in recognizing the group structure although the ICA training procedure is completely unsupervised. For an ICA with three independent components two are recognized perfectly, and three classes are lumped together. The four component ICA, which is the generalization optimal model, “recognizes” three of the five classes almost perfectly and confuses the two classes 3) and 4). Inspecting the groups we found that the two classes indeed are on very similar topics (they both concern medical documents on diseases of the human lungs), and investigating classifications for five or more ICA component did not resolve the ambiguity between them. The ability of the ICA-classifier to identify the topic structure is further illustrated in figure 3 where we show scatterplots color coded according to ICA classifications. This shows that the ICA is better than LSI in identifying relevant latent semantic structure. Finally, we inspect the histograms produced by ICA by backprojection using the PCA basis. Thresholding the ICA histograms we find the salient terms for the given component. These terms are keywords for the given topic as shown in tables 1 and 2, and follow nicely the behavour of the confusionmatrixes.

	$\mathbf{C_1}$	$\mathbf{C_2}$	$\mathbf{C_3}$	$\mathbf{C_4}$	$\mathbf{C_5}$	**keywords**
$\mathbf{IC_1}$	37	0	0	0	0	lens protein
$\mathbf{IC_2}$	0	16	1	1	0	arterial blood cerebral oxygen rise
$\mathbf{IC_3}$	0	0	21	22	26	acid blood cell fatty free glucose insulin

Table 1: Confusionmatrix for a simple classifier constructed from the three component ICA. Two of the five MED classes are recovered while the last independent component contains a mixture of the remaining three classes.

	$\mathbf{C_1}$	$\mathbf{C_2}$	$\mathbf{C_3}$	$\mathbf{C_4}$	$\mathbf{C_5}$	**keywords**
$\mathbf{IC_1}$	31	0	0	0	0	lens protein
$\mathbf{IC_2}$	0	16	0	1	0	arterial blood cerebral oxygen rise
$\mathbf{IC_3}$	6	0	22	21	2	alveolar cell lens lung
$\mathbf{IC_4}$	0	0	0	1	24	acid blood fatty free glucose insulin

Table 2: Confusionmatrix for a simple classifier constructed from the four component ICA. Three of the five MED classes are recovered, while the remaining two classes are mixed. The two unresolved classes are related by both making reference to the lung physiology.

3 Conclusion

A likelihood-MAP formulation of the noisy separation problem has been given. In the face of additive noise, the MAP estimate of the independent sources is a nonlinear functional of the observed signal, in contrast with the linear "classical" solution. Learning rules for the mixing matrix and the noise parameter are based on the Boltzmann learning. We proposed the generalization error – defined as the average negative log-likelihood and estimated on an independent test set– as a means for optimization of source model complexity. The ICA model can be used for representation of word histograms in textmining. We found that the non-orthogonal basis found by ICA "explains" the group structure better than laten semantic analysis (PCA). This was substantiated by the confusion matrix for a simple pattern recognition device based on the recovered sources.

Acknowledgment This work was funded by the Danish Research Councils through the Intermedia plan for multimedia research and the THOR Center for Neuroinformatics.

References

[1] T.-W. Lee, M. Girolami, A.J. Bell and T.J. Sejnowski: *A unifying Information-theoretic framework for Independent Component Analysis*. International Journal on Mathematical and Computer Modeling, in press (1998)

[2] A. Belouchrani and J.-F. Cardoso. *Maximum likelihood source separation by the expectation-maximization technique: deterministic and stochastic implementation* In Proc. NOLTA, 49-53 (1995).

[3] S. Geman, E. Bienenstock, and R. Doursat *Neural Networks and the Bias/Variance Dilemma*, Neural Computation, **4**, 1-58 (1992).

[4] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer and R. Harshman: Indexing by Latent Semantic Analysis. Journ. Amer. Soc. for Inf. Science. **41** 391-407 (1990).

[5] L.K. Hansen: *Blind Separation of noisy mixtures*. Department of Mathematical Modeling. Tech. Univ. Denmark (1998) http://eivind.imm.dtu.dk/pub.

[6] B. Lautrup, L.K. Hansen I. Law, N. Mørch, C. Svarer, S.C. Strother: *Massive weight sharing: A cure for extremely ill-posed problems*. In H.J. Hermanet al., eds. Supercomputing in Brain Research: From Tomography to Neural Networks. World Scientific Pub. Corp. 137-148 (1995).

[7] D. MacKay: *Maximum Likelihood and Covariant Algorithms for Independent Components Analysis*. "Draft 3.7" (1996).

[8] B.A. Pearlmutter and L.C. Parra: *A context-sensitive generalization of ICA*. In Proc. 1996 International Conference on Neural Information Processing. Hong Kong (1996).

[9] C. Peterson & J.R. Anderson: *A Mean Field Theory Learning Algorithm for Neural Networks* Complex Systems **1**, 995-1019 (1987).