INDEPENDENT COMPONENT ANALYSIS

Submitted to

Dr. Joseph Picone
ECE8990 — Special Topics in ECE
Pattern Recognition

By
Issac Alphonso
March 19, 2001

INDEPENDENT COMPONENT ANALYSIS

Issac Alphonso

Critical Review Paper
ECE 8990 - Special Topics in ECE
Pattern Recognition
email: {alphonso}@isip.msstate.edu

ABSTRACT

Independent component analysis (ICA) is a technique that projects the data in the direction of maximum independence. ICA has been around since the seventies, however, it is not until recently that people have started using it in feature extraction. The main reason people have stayed away from ICA thus far is because it does not work if the independent components we seek are Gaussian in nature (although we may be able to get bye with just one Gaussian independent component). Unlike most traditional feature extraction techniques, which work on the assumption that the features being extracted come from a Gaussian distribution, ICA aims at maximizing the nongaussianity of the features. This review will focus on two things, the theory behind the ICA pursuit and a critique of the published experimental results.

1. INTRODUCTION

"The fundamental restriction in ICA is that the independent components must be nongaussian for ICA to be possible." (Hyvarinen, Erkki 1999) This restriction may come as a surprise to some people, but the fact remains that the initial mixing matrix cannot be estimated for Gaussian independent components. The interesting aspect of ICA is that it can be shown that by maximizing the nongaussianity of the data we can obtain the features that contribute the most information. The paper in review, "Independent component analysis applied to feature extraction for robust automatic speech recognition," was written by L. Potamitis, N. Fakotakis and G. Kokkinakis. In the paper the authors show an improvement in the word accuracy using ICA to select spectral and cepstral coefficients for training. The paper does not provide an apples to apples comparison of ICA with other more traditional techniques like PCA and LDA. However, it does provide an insight into the application of ICA to area of speech recognition.

This paper has been organized as follows: section 2 will be a review of PCA and its advantages. Similarly, section 3 will be a review of LDA. Section 4 will describe the theory behind the ICA pursuit. Section 5 will describe the principles if ICA estimation and then finally section 6 will be a critique of the published experimental results.

2. KARHUNEN LOEVE TRANSFORM

The Karhunen Loeve transform (PCA) is a classic dimensionality reduction technique which works by linearly combining features. PCA is a linear transformation that seeks a projection that best represents the data in a least-square sense. The linear transform can be obtained by considering the problem of representing a set of n d-dimensional samples $\mathbf{x}1,..., \mathbf{x}n$ by a single vector $\mathbf{x}0$. An interesting one-dimensional representation can be obtained by projecting the data onto a line running through the sample mean which is given by

$$\boldsymbol{x} = \boldsymbol{m} + a\boldsymbol{e}$$

where $\mathbf{m}$ is the sample mean, $\mathbf{e}$ is a unit vector in the direction of the line and the coefficient *a* is a scalar which correspond to the distance of any point **x** from the mean **m**. Using the above representation we can find an optimal set of coefficients by minimizing the squared error.

$$J_0(\boldsymbol{x}_0) = \sum_{k=1}^{n} \|(\boldsymbol{m} + a_k\boldsymbol{e}) - \boldsymbol{x}_k\|^2$$

This brings us to the more interesting problem of finding the best direction for **e.** The solution to this involves a scatter matrix **S** which is nothing more than *(n-1)* times the sample covariance. It can be shown that the vector **e** that maximizes **J**o(.) also maximizes $\mathbf{e}^t\mathbf{Se}$. We can use Lagrange multipliers and differentiate the above equation w.r.t **e** subject to the constraint that **e** is a unit vector to obtain

$$\boldsymbol{Se} = \lambda\boldsymbol{e}$$

Clearly the above equation suggests that the solution vector **e** is the eigen vector of the scatter matrix **S**.

3. FISHER'S LINEAR DISCRIMINANT

Linear discrimination analysis considers the problem of classifying *n d*-dimensional samples by reducing it into a more manageable *p*-dimension space ($p < n$) [6]. In two-dimensions LDA can be thought of as the projection of the samples onto a line. The goal of linear discrimination is to move the line around and find an orientation for which the projected samples are well separated.

If we have a set of *n d*-dimensional samples $\mathbf{x}_1 ... \mathbf{x}_n$ and if we use the samples in the set to form a linear combination of the components of **x**, we obtain the scalar dot product $y = \boldsymbol{W}^t\boldsymbol{x}$ As you can see the direction of the vector **W** is importance in discriminating between the classes. Hence, our goal is simply the matter of finding the best possible direction for **W**.

In order to determine the best possible direction for **W** we define the *scatter matrices* $\mathbf{S_i}$ and $\mathbf{S_w}$. $\mathbf{S_i}$ is defined as a measure of the variability or scatter of the samples within the class

$$\boldsymbol{S}_i = \sum_{\boldsymbol{x} \in D_i} (\boldsymbol{x} - \boldsymbol{m}_i)(\boldsymbol{x} - \boldsymbol{m}_i)^{\boldsymbol{t}}$$

and $\mathbf{S_w}$ is a measure of the total within-class variability or scatter and is given by

$$\boldsymbol{S}_w = \sum_{i=1}^{c} \boldsymbol{S}_i$$

Apart from the within-class scatter we define another *scatter matrix* called the between-class scatter. The between-class scatter $\mathbf{S_b}$ is a measure of the variability of the various class means w.r.t to the global mean and is given by

$$\boldsymbol{S}_b = \sum_{i=1}^{c} n_i(\boldsymbol{m}_i - \boldsymbol{m})(\boldsymbol{m}_i - \boldsymbol{m})^t$$

Using the *scatter matrices* we define an objective function ***J*(.)** such that maximizing the objective function leads to the optimal value for **W**. Intuitively it can be seen that in order to maximize ***J*(.)** the class means need to be as far apart as possible and the samples within the classes need to be tightly clustered. It can be shown that a vector **W** that maximizes ***J*(.)** must satisfy

$$\boldsymbol{S}_b\boldsymbol{W} = \lambda\boldsymbol{S}_w\,\boldsymbol{W}$$

Where λ represents the eigen values and **W** represents the eigen vectors of the between-class to the within-class ratio. In order to reduce the dimensionality we select the eigen vectors with the *p* largest eigen values ($p < n$).

4.THE THEORY BEHIND ICA

Assume that we observe *n* linear mixtures of the form $x_1,..., x_n$. The mixtures are a linear combination of *n* independent components.

$$x_j = a_{j1}s_1 + a_{j2}s_2 + \ldots + a_{jn}s_n$$

Where s_k represents the independent components we are trying to find. Without loss of generality we can assume that both the mixture variables and the independent components have zero mean. If the variables do not have a zero mean they can always be centered by subtracting the sample mean from them.

Using vector notation we can let **x** denote the mixtures x1,..., xn, and likewise we can let **s** denote the independent components. Let us denote the mixing matrix **A** as the matrix of all a_{ij} elements. The mixing model is written as

$$x = As$$

The statistical model above is know as the independent component analysis. The mixing matrix **A** is assumed to be unknown and needs to be estimated using the mixtures **x**. There are two main conditions upon which ICA depends. The first condition is that the components s_k are statistically independent. The second condition is that the independent components must have a nongaussian distribution. Then, after estimating the matrix **A** we can compute its inverse **W** and obtain the independent components by

$$s = Wx$$

The second condition upon which ICA depends is also the fundamental restriction in ICA. To see why this is a problem, assume that the mixing matrix **A** is orthogonal and the independent components s_k are Gaussian. Then the mixture components x1,..., xn are Gaussian since s_k are Gaussian, uncorrected since s_k are independent and of unit variance.The joint density of the mixture components x1,..., xn is completely symmetric. Therefore, it does not contain any information on the direction of the column vectors in the mixing matrix **A.** Hence, the mixing matrix **A** cannot be estimated.

5. PRINCIPLES OF ICA ESTIMATION

Intuitively speaking, the key to estimating the ICA model is nongaussianity. The Central Limit Theorem tells us that a sum of independent random variables tends towards a Gaussian distribution, under certain conditions. Thus, a sum of independent random variables usually has a distribution that is more Gaussian that any of the two original random variables. Let **x** be a vector of observations where each observation is a linear mixture of independent components. To estimate one of the independent components, we consider a linear combination of the x_i

$$y = w^T x = \sum_i w_i x_i$$

where **w** is a vector to be determined. If **w** were one of the columns of the inverse of **A**, this linear combination would actually equal one of the independent components. The main question now is how can we use the Central Limit Theorem to determine **w**. In practice we cannot determine such a **w** exactly because we have no knowledge of the mixing matrix **A**. However, we can find an estimator that gives us a good approximation.

5.1 Kurtosis

To use nongaussianity in ICA estimation, we must have a quantitative measure of nongaussianity of the output random variable, say *y*. To simplify things we can assume that *y* is centered and has a zero mean. A classical measure of nongaussianity is kurtosis or the forth-order cumulant. The kurtosis of *y* is denoted by

$$kurt(y) = E\{y^4\} - 3\left(E\{y^2\}\right)^2$$

Also, since we assumed that *y* is of unit variance, the right hand simplifies to $E\{y^4\}$ - 3. For a Gaussian random variable *y* the forth moment equals $3(E\{y^2\})^2$. Thus, the kurtosis for a gaussian random variable is zero and non-zero for most nongaussian random variables.

In practice, determining the independent components is simply a matter of maximizing the contrast function via some variation of the gradient decent search. Computationally it is just a matter of finding the forth moment, however, a drawback of using the kurtosis function is its sensitivity to outliers in the data.

5.2 Negentropy

A second measure of nongaussianity is given by negentropy. Negentropy is based on the information-theoretic quantity of differential entropy. The entropy of a continuous random variable **H** which is also called the differential entropy is given by

$$\boldsymbol{H}(y) = -\int f(y) \log f(y) dy$$

A fundamental result in information theory is that a Gaussian variable has the largest entropy among all random variables of equal variance. This means that the entropy could be used as a measure of nongaussianity. To obtain such a measure of nongaussianity we can use a slightly modified version of the differential entropy, called negentropy. The negentropy $\boldsymbol{J}$ is defined as

$$\boldsymbol{J}(y) = \boldsymbol{H}(y_{gauss}) - \boldsymbol{H}(y)$$

where y_{gauss} is a Gaussian random variable with the same covariance as *y*. In practice, finding a suitable estimator for negentropy is difficult and therefore this contrast function remains a theoretical one. A classical method of approximating negentropy is by using higher-order moments

$$\boldsymbol{J}(y) = \frac{1}{12} E\{y^3\}^2 + \frac{1}{48} kurt(y)^2$$

Note that the validity of such approximations may be rather limited. In particular, these approximations suffer from the non-robustness encountered with kurtosis. To overcome such problems new approximations were developed in [2] based on the maximum-entropy principle.

$$\boldsymbol{J}(y) = [\boldsymbol{E}\{\boldsymbol{G}(y)\} - \boldsymbol{E}\{\boldsymbol{G}(v)\}]^2$$

In the equation above *v* is a standardized Gaussian random variable and the function $\boldsymbol{G}(.)$ are a set of non-quadratic functions [2]. This clearly is a generalization of the moment-based approach if *y* is symmetric. In practice, taking $\boldsymbol{G}(u)$ to be $1/a_1 \log[\cosh[a_1 u]]$ or $-\exp[-u^2/2]$ has proven to be useful where a_1 is some suitable constant in the range [1, 2].

CONCLUSION

As mentioned in the introduction, the main aim of the paper in review is to give a description of the work done by the authors in extracting features using ICA. The paper does not introduce any new techniques or approximations. The paper is simply an application of the work done by other people in the *Neural Information Processing* community to speech recognition. Perhaps my biggest criticism of the paper are the comparisons in the experimental results.

Table 1: Word recognition Accuracy (%)

MFCC	56.57
MFCC + CMN	70.68
ICA from MFCC + CMN	74.21
ICA from log FBANKS	76.22

In the table above the authors do not provide an apples to apples comparison between a system that uses say log filter banks without ICA and a system that uses log filter banks with ICA. Also, a useful point of comparison would have been if the authors had provided word recognition accuracy results for PCA and LDA for the systems.

REFERENCES

[1] P. Common, “Independent component analysis - a new concept?.”, Signal Processing, pp. 36:287-314, 1994.

[2] A. Hyvarinen. “New approximations of differential entropy for independent component analysis,” *Advances in Neural Information Processing Systems*, pp. 10:273-279, MIT Press, 1998.

[3] R. O. Duda, P. E. Hart, D. G. Stork, *Pattern Classification*, Wiley-Interscience Publishing, New York, New York, USA, 2000.