# Independent Component Analysis Using Nonparametric Likelihood Ratio Criterion

Jen-Tzung Chien and Bo-Cheng Chen

Department of Computer Science and Information Engineering National Cheng Kung University, Tainan, Taiwan 70101, ROC jtchien@mail.ncku.edu.tw

## ABSTRACT

This paper presents a novel nonparametric likelihood ratio criterion for independent component analysis (ICA). This criterion is derived through statistical hypothesis test of independence of random variables. A likelihood ratio (LR) criterion is developed to measure the strength of independence. We accordingly estimate the unmixing matrix by maximizing the LR function and applied to transform data into independent component space. Conventionally, the test of independence was established assuming data distributions being Gaussian, which is improper to realize ICA. To prevent assuming Gaussianity in hypothesis testing, we propose a nonparametric approach where the distributions of random variables are calculated using kernel density functions and adopted for estimation of LR function. Finally, a new ICA is fulfilled using the nonparametric likelihood ratio (NLR) criterion. In the experiments, we apply the proposed ICA for blind source separation and speech recognition. The evaluation on using NLR criterion shows good performance for separation and recognition of speech signals.

## **1. INTRODUCTION**

Independent component analysis (ICA) has been increasingly important and widely applied for data analysis and blind signal separation (BSS). Many applications were developed including financial data analysis, telecommunication, speech signal processing and medical image processing [5]. In general, ICA aims to find an  $N \times N$  unmixing matrix W, which separates the mixed signal  $\mathbf{x} = [x_1, \dots, x_N]^T$  and recovers the original independent components  $\mathbf{s} = [s_1, \dots, s_N]^T$ . The signal  $\mathbf{x}$  was mixed via an unknown linear matrix A,  $\mathbf{x} = A\mathbf{s}$ . The key idea of estimating ICA model is to measure non-Gaussianity so as to attain the independence of sources [5]. Traditionally, the highorder statistics and information-theoretic criteria were exploited to measure the non-Gaussianity or independence. The high-order statistics using absolute value of kurtosis was maximized to find independent components. However, kurtosis is sensitive to outlier data. Also, the information-theoretic criteria using negentropy, likelihood function and mutual information are successful for ICA framework. We may minimize the mutual information between the transformed sources. A transformation (unmixing) matrix is accordingly estimated to separate the mixed signals. Such optimization was shown to be equivalent to maximum likelihood and maximum negentropy principles [5].

In this paper, we propose a statistical approach to establish ICA model using the *hypothesis test* principle. We are testing the hypothesis whether the transformed set of variates is independent or not. The hypothesis is verified when the test statistics in a form of *likelihood ratio* exceeds the specified significant level. The likelihood ratio of independence to dependence hypotheses serves as an objective function to find the unmixing matrix. We maximize the LR function, or equivalently the *confidence for* 

*independence*, to realize ICA. In traditional test of independence [1], the data distributions were assumed to be Gaussian, which is undesirable for ICA problem. Instead of assuming Gaussian, we present a nonparametric approach where the kernel density functions are adopted to approximate data distributions. A new nonparametric likelihood ratio criterion is derived to build the ICA model. In this study, we carry out the nonparametric likelihood ratio for blind separation of speech and music signals. The proposed approach is also applied to clustering of speech signals for hidden Markov modeling. We achieve good performance in terms of signal-to-interference ratios and speech recognition rates.

## 2. TEST OF INDEPENDENCE

We are interested in resolving ICA via test of independence for the transformed data  $\mathbf{y} = W\mathbf{x}$ . In [1], Anderson presented the criterion for testing independence of the components of  $\mathbf{y} = [y_1, \dots, y_N]^T$ . We are verifying the null hypothesis that the components  $y_1, y_2, \dots, y_N$  are mutually independent against the alternative hypothesis that components are dependent. The null and alternative hypotheses are stated as

$$H_0: y_1, y_2, \dots, y_N$$
 are mutually independent.  
 $H_1: y_1, y_2, \dots, y_N$  are not mutually independent.

Assuming that **y** is a *N* dimensional Gaussian density with unknown mean vector  $\boldsymbol{\mu} = [\mu_1, \dots, \mu_N]^T$  and covariance matrix  $\boldsymbol{\Sigma}$ , the null hypothesis is equivalent to see the covariance between vectors  $y_i$  and  $y_j$ 

$$H_0: \ \sigma_{ij}^2 = E\{(y_i - \mu_i)(y_j - \mu_j)\} = 0 \text{ for all } i \neq j.$$
(1)

Namely, when any two distinct components  $\{(y_i, y_j), i \neq j\}$  are uncorrelated,  $y_1, y_2, \dots, y_N$  are mutually independent. This property holds only for Gaussian distributions. Under  $H_0$ , the covariance matrix  $\Sigma$  becomes  $\Sigma_D$ , which is a diagonal matrix with diagonal elements  $\{\sigma_{ii}^2, i=1,\dots,N\}$ . Then, the optimal solution to this hypothesis testing is to determine the ratio of likelihoods of null hypothesis  $p(\mathbf{y}|H_0)$  to alternative hypothesis  $r(\mathbf{y}|H_0)$  as

$$p(\mathbf{y}|H_0)$$
 as

$$\lambda = \frac{p(\mathbf{y}|H_0)}{p(\mathbf{y}|H_1)} = \frac{\max_{\mu, \Sigma_D} p(\mathbf{y}|\mu, \Sigma_D)}{\max_{\mu, \Sigma} p(\mathbf{y}|\mu, \Sigma)}.$$
 (2)

By substituting Gaussian densities, the likelihood ratio turns out to be

$$\lambda = \left[\frac{\det \hat{\Sigma}}{\prod_{i=1}^{N} \hat{\sigma}_{ii}^2}\right]^{M/2}, \qquad (3)$$

where  $\hat{\Sigma}$  is the sample covariance matrix calculated from M samples  $\{\mathbf{y}^1, \mathbf{y}^2, \dots, \mathbf{y}^M\}$  and  $\hat{\sigma}_{ii}^2$  is the *i*th diagonal component of  $\hat{\Sigma}$ . With a level of significance  $\alpha$ , we can determine the threshold  $\lambda_{\alpha}$ . The null hypothesis  $H_0$  is verified when  $\lambda$  has the value in acceptance region  $\lambda \ge \lambda_{\alpha}$ . Basically, the likelihood ratio  $\lambda$  measures the confidence for null hypothesis, or equivalently the independence for  $y_1, y_2, \dots, y_N$ . The larger the likelihood ratio is, the more likely the components are mutually independent. In this study, we highlight on developing the likelihood ratio criterion for solving ICA problem. Our goal is to estimate the unmixing matrix W according to the theoretical evidence originated from the test of independence [1]. We are motivated to estimate the most likely unmixing matrix with the largest likelihood ratio

$$W_{\rm LR} = \underset{W}{\arg\max} \ \lambda(W) \ . \tag{4}$$

Matrix  $W_{LR}$  is able to optimally separate the observed signal **x** into **y** which is nearest to original independent signal **s** in likelihood ratio manner.

### **3. NONPARAMETRIC LIKELIHOOD RATIO**

As mentioned in [2][5], the key to estimating W for ICA is non-Gaussianity. According to Central Limit Theorem, the Gaussianity of random variables is increased by the linear transformation. The assumption of Gaussian distribution provides no additional information for finding W. It is forbidden to assume Gaussianity for ICA. Hence, we could not use likelihood ratio criterion in (3) to estimate unmixing matrix.

#### 3.1 Nonparametric Density Estimation

However, incorrect assumptions on distribution of unknown signals can result in poor estimation performance. As suggested in [2], Boscolo et al. presented the nonparametric density estimation for the source signals. They adopted the mutual information as the objective function to derive W while a kernel density estimation technique was applied. Differently, we are presenting a new *nonparametric likelihood ratio (NLR)* objective function. There is no assumption of parametric distributions in NLR criterion. To fulfill ICA, we investigate the distribution of the transformed signals  $\{\mathbf{y}^1, \mathbf{y}^2, \dots, \mathbf{y}^M\}$  from the observed samples  $\{\mathbf{x}^1, \mathbf{x}^2, \dots, \mathbf{x}^M\}$ . Each sample is transformed by  $\mathbf{y}^m = W\mathbf{x}^m$ . Using the Parzen windowing approach, the nonparametric density of component  $y_i$  is provided by

$$p(y_i) = \frac{1}{Mh} \sum_{m=1}^{M} \varphi \left( \frac{y_i - y_i^m}{h} \right), \quad i = 1, \dots, N , \quad (5)$$

where *h* is the kernel bandwidth and  $\varphi$  is the univariate Gaussian kernel

$$\varphi(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2} \,. \tag{6}$$

The kernel centroid  $y_i^m$  is the *i*th component of sample  $\mathbf{y}^m$ 

$$y_i^m = \mathbf{w}_i \mathbf{x}^m = \sum_{j=1}^N w_{ij} x_j^m .$$
<sup>(7)</sup>

Matrix W can be expressed by  $W = [w_{ij}]_{N \times N} = [\mathbf{w}_1^T \cdots \mathbf{w}_N^T]^T$ .

#### 3.2 Derivation of NLR Criterion for ICA

Interestingly, we employ the nonparametric density function in likelihood ratio criterion and develop the NLR criterion for ICA. Under the null hypothesis, or equivalently assuming independence between  $y_1, y_2, \dots, y_N$ , the joint distribution of **y** can be factored into the product of distributions of individual components. The NLR criterion accumulated from samples  $\{\mathbf{y}^1, \mathbf{y}^2, \dots, \mathbf{y}^M\}$  is established by

$$\lambda_{\text{NLR}}(W) = \frac{\prod_{k=1}^{M} \prod_{i=1}^{N} p(y_i^k)}{\prod_{k=1}^{M} p(\mathbf{y}^k)} = \frac{\prod_{k=1}^{M} \prod_{i=1}^{N} \left\lfloor \frac{1}{Mh} \sum_{m=1}^{M} \varphi\left(\frac{y_i^k - y_i^m}{h}\right) \right\rfloor}{\prod_{k=1}^{M} \left\lfloor \frac{1}{Mh^N} \sum_{m=1}^{M} \psi\left(\frac{\mathbf{y}^k - \mathbf{y}^m}{h}\right) \right\rfloor}.$$
 (8)

 $\psi$  is the multivariate Gaussian kernel for vectors **v** and given by

$$\psi(\mathbf{v}) = \frac{1}{\sqrt{(2\pi)^N}} e^{-\frac{1}{2}\mathbf{v}^T\mathbf{v}} \,. \tag{9}$$

The log likelihood ratio is consisted of the log likelihoods from null hypothesis  $L_0(W)$  and alternative hypothesis  $L_1(W)$ 

$$\log \lambda_{\rm NLR}(W) = L_0(W) - L_1(W) .$$
 (10)

Then, we maximize  $\log \lambda_{\text{NLR}}(W)$  with respect to W and find the optimal solution  $W_{\text{NLR}}$ . It is a natural way to apply the gradient descent algorithm. Using this algorithm, we calculate two gradients  $\nabla_W L_0(W)$  and  $\nabla_W L_1(W)$ . Specifically, we derive the components of  $\nabla_W L_0(W)$  as

$$\frac{\partial L_0(W)}{\partial w_{ij}} = \sum_{k=1}^{M} \frac{\frac{1}{Mh} \sum_{m=1}^{M} \frac{\partial}{\partial w_{ij}} \varphi\left(\frac{1}{h} \sum_{j=1}^{N} w_{ij} (x_j^k - x_j^m)\right)}{\frac{1}{Mh} \sum_{m=1}^{M} \varphi\left(\frac{\mathbf{w}_i (\mathbf{x}^k - \mathbf{x}^m)}{h}\right)}$$
$$= -\sum_{k=1}^{M} \frac{\sum_{m=1}^{M} \mathbf{w}_i (\mathbf{x}^k - \mathbf{x}^m) (x_j^k - x_j^m) \varphi\left(\frac{\mathbf{w}_i (\mathbf{x}^k - \mathbf{x}^m)}{h}\right)}{h^2 \sum_{m=1}^{M} \varphi\left(\frac{\mathbf{w}_i (\mathbf{x}^k - \mathbf{x}^m)}{h}\right)}.$$
(11)

Also, the gradient due to alternative hypothesis has the form

$$\nabla_{W} L_{1}(W) = -\sum_{k=1}^{M} \frac{\sum_{m=1}^{M} W(\mathbf{x}^{k} - \mathbf{x}^{m})(\mathbf{x}^{k} - \mathbf{x}^{m})^{T} \psi\left(\frac{W(\mathbf{x}^{k} - \mathbf{x}^{m})}{h}\right)}{h^{2} \sum_{m=1}^{M} \psi\left(\frac{W(\mathbf{x}^{k} - \mathbf{x}^{m})}{h}\right)}.$$
(12)

The iterative learning algorithm for  $W_{\rm NLR}$  is obtained by

$$W^{(n+1)} = W^{(n)} - \eta(\nabla_W L_0(W^{(n)}) - \nabla_W L_1(W^{(n)})), \qquad (13)$$

where *n* is iteration index and  $\eta$  is learning rate. In summary, the ICA procedure based on NLR criterion is shown as follows:

Parameter Initialization Initialize W & Set  $\eta$ , h Centering  $\mathbf{x}^{k} \leftarrow \mathbf{x}^{k} - E[\mathbf{x}]$ Whitening 1.  $E[\mathbf{x}\mathbf{x}^{T}] = \Phi D \Phi^{T}$ 2.  $\mathbf{x}^{k} \leftarrow \Phi D^{-1/2} \Phi^{T} \mathbf{x}^{k}$ Repeat 1. Compute  $\nabla_{W} L_{0}(W)$  and  $\nabla_{W} L_{1}(W)$ 2.  $W \leftarrow W - \eta (\nabla_{W} L_{0}(W) - \nabla_{W} L_{1}(W))$ 3.  $\mathbf{w}_{i} \leftarrow \mathbf{w}_{i} / ||\mathbf{w}_{i}||$ ,  $1 \le i \le N$ 4. Compute  $\lambda_{\text{NLR}}(W) = \prod_{k=1}^{M} \prod_{i=1}^{N} p(\mathbf{w}_{i}\mathbf{x}^{k}) / \prod_{k=1}^{M} p(W\mathbf{x}^{k})$ Until  $\lambda_{\text{NLR}}(W) \ge \lambda_{\alpha}$  (stopping criterion)

Following the instructions suggested in [5], we perform the centering and whitening processes before applying NLR based ICA algorithm. In whitening process, the eigenvalue matrix D and eigenvector matrix  $\Phi$  are computed. The learning process is terminated when  $\lambda_{\text{NLR}}(W) \ge \lambda_{\alpha}$ . Typically,  $\lambda_{\text{NLR}}(W)$  has an upper bound 1 happening in case of independence  $p(\mathbf{y}) = \prod_{i=1}^{N} y_i$ .

## 4. EXPERIMENTS

In this study, we carry out the proposed NLR based ICA algorithm for speech processing applications including the blind separation of speech and audio signals and the clustering of speech hidden Markov models (HMM's).

#### 4.1 Application for Blind Signal Separation

First of all, the proposed ICA algorithm is examined for BSS application. As shown in Figure 1, we sampled the speech and music signals from the ICA '99 BSS Test Sets [8] for evaluation. A mixing matrix *A* was randomly generated to mix the signals given in Figure 2. For comparison, we implement the ICA algorithm based on minimum mutual information (MMI) criterion [4]. The separated signals using MMI and NLR are illustrated in Figures 3 and 4, respectively. Compared to MMI, we can see the significant improvement when applying NLR criterion for ICA. During implementation, we searched the best kernel bandwidth  $\hat{h}$  within a predefined region.



Figure 1: Source speech and music signals.



Figure 3: Reconstructed signals using MMI based ICA.



Figure 4: Reconstructed signals using NLR based ICA.

Also, we evaluate the separation performance of speech and music signals by the measure of signal-to-interference ratio (SIR). Given original signal  $\{s_n\}$  and reconstructed signal  $\{y_n\}$ , SIR is calculated by SIR(dB) =  $10 \log_{10} \sum_{i=1}^{N} s_i^2 / \sum_{i=1}^{N} (y_i - s_i)^2$ . Table 1 shows the comparison of SIR for the cases of mixed signals without ICA and separated signals with MMI and NLR based ICA. The SIR of mixed signals is measured by -1.76 dB. Using MMI based ICA, SIR is increased to 5.8 dB. Further, using NLR based ICA, the SIR was improved to 17.93 dB.

	Without ICA	MMI-ICA	NLR-ICA
SIR (dB)	-1.76	5.80	17.93

Table 1: Comparison of SIR (dB) with and without ICA.

### 4.2 Application for HMM Clustering

Further, ICA is feasible to analyze and compensate the pronunciation variations among speakers. In [6], it was found that the first and the second independent components of speech

features provided sexual and accent information, respectively. ICA based speech features achieved better speech recognition performance. Nevertheless, the pronunciation variations could be compensated through clustering of speech HMM's [9]. Different clusters of HMM's corresponding to the same phonetic unit contain the variations of gender, accent, emotion etc. Having sophisticated HMM's covering different variations, we are able to elevate the speech recognition rates. Similarly, the speaker clustering was developed for speech recognition [7]. The clustered HMM's close to test speaker were used to perform speaker adaptation. In this study, we fulfill the NLR based ICA algorithm for clustering of HMM's. The speech recognition is performed using the extended HMM parameters. The HMM clustering procedure is described as follows: We first perform the forced alignment for all training data via a Viterbi decoder. The training frames corresponding to the same phonetic/HMM unit are collected. Then, we calculate the sample mean vectors  $\overline{\mathbf{x}}_{s_n}$  for the states  $\{s_n, 1 \le n \le N\}$  belonging to this HMM unit. This operation is done utterance by utterance. Accordingly, we can build utterance-level supervectors consisted of sample mean vectors of HMM states  $\mathbf{x} = [\overline{\mathbf{x}}_{s_1}^T, \overline{\mathbf{x}}_{s_2}^T, \cdots, \overline{\mathbf{x}}_{s_N}^T]^T$ . There are M samples  $\mathbf{X} = \{\mathbf{x}^1, \mathbf{x}^2, \dots, \mathbf{x}^M\}$  collected from M utterances. We subsequently perform ICA to find  $W_{\rm NLR}$  and project the sample data X to independent component subspace by  $Y = W_{NLR}X$ . As displayed in Figure 5, the male and female samples are well separated in the ICA space using NLR criterion. The clusters of HMM's are estimated using the clustered data through k-means algorithm. In general, each cluster represents a specific kind of pronunciation variation. Using clusters of HMM's, we conduct the experiments of continuous Mandarin speech recognition. The experimental setup and speech database were introduced in [3]. Training data contained forty males and forty females. Totally, there were 1000 test utterances spoken by ten males and ten females different from training speakers. Each HMM unit had at most four clusters.



Figure 5: Scattering diagrams of the first two dimensions of supervectors of Mandarin subsyllable "le" in original space (upper) and NLR-ICA space (down). "\*" and "o" represent male and female samples, respectively.

	Without Clustering	Clustering with no ICA	Clustering with MMI- ICA	Clustering with NLR- ICA
SER (%)	38.9	36.5	33.6	31.4

Table 2 Comparison of SER (%) using different methods.

As listed in Table 2, we find that clusters of HMM's do reduce the syllable error rate (SER). However, the HMM model size is increased as well. When applying ICA for clustering, the pronunciation variations can be properly compensated. Finally, SER is improved from 38.9% without clustering to 31.4% with NLR-ICA clustering.

## 5. CONCLUSION

We have presented a general objective criterion to evaluate the independence of variables. The underlying concept was initialized from the statistical test of hypotheses of independence over dependence. It turned out that a likelihood ratio criterion was determined to measure the independence. To establish ICA model, we avoided assuming Gaussian distribution and exploited a nonparametric likelihood ratio criterion using kernel density function. The resulting NLR was maximized to estimate the unmixing matrix and detect the independent sources. The experiments on speech processing applications showed that the proposed NLR based ICA algorithm did not only effectively separate the mixed speech and audio signals but also estimate the clusters of HMM's for improving speech recognition performance.

#### 6. REFERENCES

- T. W. Anderson, An Introduction to Multivariate Statistical Analysis, John Wiley & Sons, Inc, 1984.
- [2] R. Boscolo, H. Pan and V. P. Roychowdhury, "Independent component analysis based nonparametric density estimation", *IEEE Transactions on Neural Networks*, vol. 15, no. 1, pp. 55-65, January 2004.
- [3] J.-T. Chien and C.-H. Huang, "Bayesian learning of speech duration models", *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 558-567, November 2003.
- [4] P. Comon, "Independent component analysis, a new concept?", Signal Processing, vol. 36, pp. 287-314, 1994.
- [5] A. Hyvarinen and E. Oja, "Independent component analysis: algorithms and applications", *Neural Networks*, vol. 13, pp. 411-430, 2000.
- [6] C. Huang, T. Chen, S. Li, E. Chang and J. Zhou, "Analysis of speaker variability", Proc. European Conference on Speech Communication and Technology (EUROSPEECH), pp. 1377-1380, 2001.
- [7] M. Padmanabhan, L. R. Bahl, D. Nahamoo and M. A. Picheny, "Speaker clustering and transformation for speaker adaptation in speech recognition systems", *IEEE Transactions on Speech* and Audio Processing, vol. 6, no. 1, pp. 71-77, January 1998.
- [8] D. Schobben, K. Torkkola and P. Smaragdis, "Evaluation of blind signal separation methods", *Proc. Int. Workshop on Independent Component Analysis and Blind Signal Separation (ICA '99)*, pp. 261-266, 1999.
- [9] R. Singh, B. Raj and R. M. Stern, "Automatic generation of subword units for speech recognition systems," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 2, pp. 89-99, February 2002.