ROMAI J., **2**, 2(2006),
45–48

ICA BY MINIMIZATION OF MUTUAL INFORMATION

Doru Constantin
Faculty of Mathematics-Informatics, University of Pitesti
cdomanid@yahoo.com

Abstract An important approach for independent component analysis (ICA) estimation, inspired by information theory, is minimization of mutual information. The motivaton of this approach is that it may be not very realistic in many cases to assume that the data follows the ICA model. Therefore, it was developed an approach that does not assume anything about the data. We intended to have a general-purpose measure of the dependence of the components on the random vector. Using such a measure, we define ICA as a linear decomposition that minimizes that dependence measure. Such an approach can be developed using mutual information, which is a information-theoretic measure of statistical dependence.

1. DEFINING BY MUTUAL INFORMATION

1.1. INFORMATION-THEORETIC CONCEPTS

In the following we present some concepts of information theory. The differential entropy H of a random vector y with density $p(y)$ is defined as

$$H(y) = -\int p(y) \log p(y) \tag{1}$$

A normalized version of entropy is given by negentropy J, which is defined as

$$J(y) = H(y_{gauss}) - H(y), \tag{2}$$

where y_{gauss} is a Gaussian random vector of the same covariance matrix as y. Negentropy is always nonnegative, and zero only for Gaussian random vectors. Mutual information I between m random variables, $y_i, i = 1, ..., m$ is defined as

$$I(y_1, y_2, ..., y_m) = \sum_{i=1}^{m} H(y_i) - H(y). \tag{3}$$

1.2. MUTUAL INFORMATION AS MEASURE OF DEPENDENCE

Mutual information is a natural measure of the dependence between random variables. It is always nonnegative, and zero if and only if the variables are statistically independent. We can use mutual information as the criterion for finding the ICA representation. This approach is an alternative to the model estimation approach. We define the ICA of a random vector x as an invertible transformation

$$s = Bx, \tag{4}$$

where the matrix B is determined so that the mutual information of the transformed components s_i is minimized. Minimization of mutual information can be interpreted as giving the maximally independent components.

2. MUTUAL INFORMATION AND NONGAUSSIANITY

Using the formula for the differential entropy we obtain the expression of mutual information for an invertible linear transformation $y = Bx$

$$I(y_1, y_2, ..., y_m) = \sum_{i=1}^{m} H(y_i) - H(x) - \log|\det B| \tag{5}$$

Next, we constraint that y_i to be uncorrelated and unit variance. This means $E\{yy^T\} = BE\{xx^T\}B^T = I$, which implies

$$\det I = 1 = \det(BE\{xx^T\}B^T = I) = (\det B)(\det E\{xx^T\})(\det B^T) \tag{6}$$

and this implies that $\det B$ must be constant since $\det E\{xx^T\}$ does not depend on B. Moreover, for y_i of unit variance, entropy and negentropy differ only by a constant and the sign. Thus we obtain

$$I(y_1, y_2, ..., y_n) = const - \sum_i J(y_i), \tag{7}$$

where the constant term does not depend on B. This shows the fundamental relation between negentropy and mutual information.

The relation (7) shows that finding an invertible linear transformation B that minimizes the mutual information is roughly equivalent to finding directions in which the negentropy is maximized. Negentropy is a measure of nonGaussianity. Thus, (7) shows that *ICA estimation by minimization of mutual information is equivalent to maximizing the sum of nonGaussianities of the*

estimates of the independent components, when the estimation are constrained to be uncorrelated.

It follows that the formulation of ICA as minimization of mutual information gives another rigorous justification of idea of finding maximally nonGaussian directions.

In practice there are some important differences between these two criteria:

1 : Negentropy, and other measure of nonGaussianity, enable the deflationary (one-by-one), estimation of the independent components, since we can look for the maxima of nonGaussianity of a single projection $b^T x$. This is not possible with mutual information or most other criteria, like the likelihood.

2 : A smaller difference is that in using nonGaussianity, we force the estimations of the independent components to be uncorrelated. This is not necessary when using mutual information, because we could use the form in (5) directly.

3. MUTUAL INFORMATION AND LIKELIHOOD

Mutual information and likelihood are intimately connected. To see the connection between likelihood and mutual information, consider the expectation of the log-likelihood in

$$\frac{1}{T} E\{\log L(B)\} = \sum_{i=1}^{n} E\{\log p_i(b_i^T x)\} + \log |\det B|. \tag{8}$$

If the p_i were equal to the actual density functions's of the $b_i^T x$, the first term would be equal to $-\sum_i H(b_i^T x)$. Thus the likelihood would be equal, up to an additive constant given by the total entropy of x, to the negative of mutual information as given in (5).

In practice, the connection may be just as strong, or even stronger. This is because in practice we do not know the distributions of the independent components that are needed in ML estimation. A reasonable approach would be to estimate the density of $b_i^T x$ as part of the ML estimation method, and use this approximation of the density of s_i. Then, the p_i in this approximation of likelihood are indeed equal to the actual density functions $b_i^T x$. Thus, the equivalency would really hold.

In order to approximate mutual information, we would take a fixed approximation of the densities y_i and plug this in the definition of entropy. Denoting the densities functions's by $G_i(y_i) = \log p_i(y_i)$, we could approximate (5) as

$$I(y_1, y_2, ..., y_m) = \sum_i E\{G_i(y_i)\} - \log |\det B| - H(x). \tag{9}$$

Concluding remarks A rigorous approach that is different from maximum likelihood approach is given by minimization of mutual information. Mutual

information is a natural information-theoretic measure of dependence, and therefore it is natural to estimate the independent components by minimizing the mutual information of their estimates. Mutual information gives a rigorous justification of the principles of searching for maximally nongaussian directions, and in the end turns out to be very similar to the likelihood as well. Mutual information can be approximated by the same methods that negentropy is approximated.

References

[1] Bishop, C. M. *Neural network for pattern recognition*, Clarendon, 1995.

[2] Cichocki, A., Unbehauen, R., *Neural networks for signal processing and optimization*, New York, Wiley, 1994.

[3] Cover, T. M., Thomas, J. A. *Elements of information theory*, Wiley, New York, 1991.

[4] Diamantaras, K. I., Kung, S. Y., *Principal component neural networks: theory and applications*, Wiley, New York, 1996.

[5] Gardner, W., *Introduction to random processes with applications to signal and systems*, Macmillan, 1986.

[6] Gonzalez, R., Wintz, P., *Digital image processing*, Addison-Wisley, 1987.

[7] Girolami, M., *Self-organising neural networks - independent component analysis and blind source separation*, Springer, Berlin, 1999.

[8] Hyvärinen, A., Karhunen, J., Oja, E., *Independent component analysis*, John Wiley , New York, 2001.

[9] Kay, S., *Modern spectral estimation: theory and application*, Prentince Hall, Englewood Hills, 1988.

[10] Kohonen, T., *Self-organizing maps*, Springer, Berlin, 1995.

[11] Lee, T. W., *Independent component analysis: theory and applications*, Kluwer, Dordrecht, 1998.

[12] Oppenheim, A., Schafer, R., *Discrete-time signal processing*, Prentince Hall, Englewood Hills, 1989.