UNSUPERVISED LEARNING FOR BLIND SOURCE SEPARATION: AN INFORMATION-THEORETIC APPROACH

Dragan Obradovic and Gustavo Deco

Siemens AG, Corporate Research and Development, ZT IK 4
Otto-Hahn-Ring 6, 81739 Munich, Germany
Dragan.Obradovic@mchp.siemens.de

Abstract

This paper provides a detailed and rigorous analysis of the two commonly used methods for redundancy reduction: Linear Independent Component Analysis (ICA) and Information Maximization (InfoMax). The paper shows analytically that ICA based on the Kullback-Leibler information as a mutual information measure and InfoMax lead to the same solution if the parameterization of the output nonlinear functions in the latter method is sufficiently rich. Furthermore, this work briefly discusses the alternative redundancy measures not based on the Kullback-Leibler information distance and Nonlinear ICA. The practical issues of applying ICA and InfoMax are also discussed.

1. INTRODUCTION

The pioneer work of Zipf [1] and the ideas of Attneave [2] about information processing in visual perception have led to the idea that nervous system and brain may be regulated by an economy principle. In the neural network society these ideas were introduced by the important paper of Barlow [3]. In this work the author presented the connectionist model of unsupervised learning under the perspective of redundancy reduction. The minimum entropy coding method was introduced for the generation of factorial codes [4]. Atick and Redlich [5] demonstrated that statistically salient input features can be optimally extracted from a noisy input by maximizing mutual information. Simultaneously, Atick and Redlich [6] and specially the works of Redlich ([7], [8]) concentrate on the original idea of feature extraction by redundancy reduction. Several neural network learning algorithms for PCA are presented, among others, in [9] and [10].

The problem of Linear Independent Component Analysis as linear feature extraction was introduced by Comon [11] and further extended in linear and defined in nonlinear case by the works of the authors ([12]-[19]). In parallel, Bell and Sejnowski [20] have demonstrated that their InfoMax method can also achieve linear feature extraction. This paper provides a detailed and rigorous analysis of the two methods and derives conditions under which these methods lead to identical solution. In addition, the paper briefly addresses the cumulant based criteria for ICA as well as Nonlinear ICA.

2. LINEAR INDEPENDENT COMPONENT ANALYSIS AND INFORMATION MAXIMIZATION

Let $\hat{x}$ be random vector of dimension n with the joint probability density function $p(\hat{x})$ whose covariance matrix is nonsingular. Furthermore, let M be a linear square map which maps $\hat{x}$ into the random vector $\hat{y}$ whose probability density function is $p(\hat{y})$.

Definition 1: ICA

Linear Independent Component Analysis (ICA) is an input/output linear transformation M from $\hat{x}$ to $\hat{y}$ such that the output components with joint probability:

$$p(\hat{y}) = p(y_1 \ldots y_n); \quad \hat{y} = M\hat{x} \tag{1}$$

are "as independent as possible" according to the appropriate measure. In the special case where the complete independence of the output components is achieved, the following holds:

$$p(y_1 \ldots y_n) = p(y_1) \ldots p(y_n) \tag{2}$$

If the input vector $\hat{x}$ is jointly Gaussian, ICA is equivalent to the problem of diagonalizing the output covariance matrix Q_y which is the standard PCA problem. In order to guarantee the existence of the solution for the ICA problem, we assume that the input signal $\hat{x}$ was originally obtained by the invertible linear mixture of the statistically independent signals $z_1 \ldots z_n$.

Definition 2: Information maximization

Let the above defined random vector $\vec{x}$ be transmitted through a combination of a matrix M and n nonlinear functions f_i; $i = 1 \div n$ such that the resulting components of the output vector $\vec{w}$ are defined as:

$$w_i = f_i(y_i) \qquad \vec{y} = M\vec{x} \tag{3}$$

Under the assumption that the every nonlinear function f_i is differentiable and that its derivative f_i' satisfies

$$\int_{-\infty}^{\infty} f_i' dy_i = 1 \tag{4}$$

the information maximization problem is defined as maximization of the entropy

$$H(\vec{w}) = \int d\vec{w}\ p(\vec{w})\ \log(p(\vec{w})) \tag{5}$$

over the elements of matrix M and, possibly, the free parameters in the parameterization of f_i. Typical choices for f_i are single or normalized sums of sigmoid functions.

At first glance the ICA and InfoMax problems seem to be substantially different. Nevertheless, it is known that the information maximization leads to the statistical factorization of the output components w_i, i.e. that it essentially performs the same task as ICA [20]. In the remaining part of the paper we give a rigorous proof that these two problems are identical when the Kullback-Leibler information is used as a measure of the statistical independence in ICA and when the derivatives f_i' are capable of approximating output marginal distributions with the infinite precision.

The Kullback-Leibler distance between the joint and the marginal probabilities is defined as:

$$K\{p(\vec{y}), \prod_i p(y_i)\} = \int d\vec{y}\ p(\vec{y})\ \log\left(\frac{p(\vec{y})}{\prod_i p(y_i)}\right) \geq 0 \tag{6}$$

or equivalently:

$$K\{p(\vec{y}), \prod_i p(y_i)\} = \sum_{i=1}^{n} H(y_i) - H(\vec{y}) \tag{7}$$

Equation (7) indicates that the Kullback-Leibler distance is the mutual information between the output components y_i.

The relationship between the input and output joint probabilities of a differentiable map g is equal to:

$$p(\overrightarrow{out}) = \frac{p(\vec{in})}{|\det(J)|} \tag{8}$$

where J is the Jacobian matrix of g. Consequently, the relationship between the corresponding entropies is:

$$H(p(\overrightarrow{out})) = H(p(\vec{in})) + \int d\vec{in}\ p(\vec{in})\ \ln(|\det(J)|) \tag{9}$$

Combining equations (6) and (7) with (8), it follows that:

$$K\{p(\vec{y}), \prod_i p(y_i)\} = \int d\vec{x}\ p(\vec{x})\ \log\left(\frac{p(\vec{x})}{\prod_i p(y_i)\cdot|\det(M)|}\right) = -H(p(\vec{x})) - \int d\vec{x}\ p(\vec{x})\ \log\left(\prod_i p(y_i)\cdot|\det(M)|\right) \geq 0 \tag{10}$$

Since the input entropy $H(p(\vec{x}))$ is independent of the input-output transformation, the minimization of $K\{p(\vec{y}), \prod_i p(y_i)\}$ is equivalent to maximization of $\int d\vec{x}\ p(\vec{x})\ \log\left(\prod_i p(y_i)\cdot|\det(M)|\right)$, i.e. to the Maximum Likelihood Expectation (MLE) of $\log\left(\prod_i p(y_i)\cdot|\det(M)|\right)$. In general, the analytical expression for the marginal probabilities $p(y_i)$ are not known, and their estimates $\hat{p}(y_i)$ have to be obtained from the data for every change of the matrix M.

Similarly, in the information maximization problem the output joint entropy $H(\vec{w})$ is equal to:

$$H(\vec{w}) = -\int d\vec{y}\ p(\vec{y})\ \log\left(\frac{p(\vec{y})}{\prod_i f_i'(y_i)}\right) = -K\{p(\vec{y}), \prod_i f_i'(y_i)\} = H(p(\vec{x})) + \int d\vec{x}\ p(\vec{x})\ \log\left(\prod_i f_i'(y_i)\cdot|\det(M)|\right) \geq 0 \tag{11}$$

or, equivalently, to the MLE of

$$\log\left(\prod_i f_i'(y_i)\cdot|\det(M)|\right)\cdot$$

Hence, the ICA with the Kullback-Leibler information measure and the maximum information transfer as defined in this paper are posed as:

$$\begin{aligned}\text{ICA} &\Rightarrow \min\ K\{p(\hat{y}),\prod_i p(y_i)\}\\ \text{INFOMAX} &\Rightarrow \min\ K\{p(\hat{y}),\prod_i f_i'(y_i)\}\end{aligned}\qquad(12)$$

or, equivalently,

$$\begin{aligned}\text{ICA} &\Rightarrow \text{MLE}\{\log\left(\prod_i \hat{p}(y_i)\cdot|\det(M)|\right)\}\\ \text{INFOMAX} &\Rightarrow \text{MLE}\{\log\left(\prod_i f_i'(y_i)\cdot|\det(M)|\right)\}\end{aligned}\qquad(13)$$

The initial parameterization of the derivatives $f_i'(y_i)$ in (13) has a possible interpretation as the prior on the estimation of the actual marginal densities $p(y_i)$. As mentioned earlier, both methods require parameterization of $\hat{p}(y_i)$ and $f_i'(y_i)$. Hence, the problem statements in (12) and (13) can be used to derive conditions for the equivalence of solutions of ICA and InfoMax.

Lemma:

For a given input distribution $p(\hat{x})$, the ICA and InfoMax problems achieve the same degree of statistical independence if the derivatives $f_i'(y_i)$ can be parameterized in the form of the marginal distribution estimates $\hat{p}(y_i)$.

The proof is straightforward since it requires that the parameterization of $p(\hat{x})$ and $f_i'(y_i)$ are identical. This can be illustrated on an example.

Example:

The marginal probabilities $p(y_i)$ have to be estimated from the data. A typical way of doing that is to estimate elements of a probability density function expansion up to the desired order. Let us use the first element of the Edgeworth expansion [19], i.e. let $\hat{p}(y_i)$ have the form of a Gaussian whose mean and standard deviation σ_i is equal to the those of the actual marginal distribution $p(y_i)$. Without a loss of generality let us assume that the input distribution $p(\hat{x})$ is zero-mean. In addition, let us parameterize the derivatives $f_i'(y_i)$ as zero-mean Gaussian distributions whose standard deviations r_i are optimization parameters. Hence, it is easy to see that the MLE problems in (13) become

$$\text{ICA} \Rightarrow \text{MLE}\{-\sum[\log(\sigma_i)] + \log(|\det(M)|)\}$$

$$\text{where}\quad \sigma_i = \langle y_i^2\rangle \qquad(14)$$

$$\text{INFOMAX} \Rightarrow \text{MLE}\left\{-\sum\left[\log(r_i)+\frac{y_i^2}{2r_i^2}\right]+\log(|\det(M)|)\right\}$$

The resulting ICA problem is nothing more than the covariance matrix diagonalization [19] where the optimization is performed over the elements of the matrix M. In the case of InfoMax, the unknown parameters are not only the elements of M but also the Gaussian parameters r_i. It is easy to see that the optimal value of r_i for every fixed matrix M is the actual standard deviation σ_i and, therefore, that the solution of InfoMax problem will also result in the covariance matrix diagonalization.

In practice, it is required that the solutions of both methods are unique modulo transformations that preserve statistical independence such as the component order permutation and diagonal scaling. The uniqueness is achieved if the number of Gaussian components of $p(\hat{x})$ does not exceed one. In the case of multiple Gaussian distributions, it is well known that there is an infinite number of matrix transformations that diagonalize the covariance matrix. Hence, the ICA and InfoMax algorithms will have unique solutions only if the original signal $\hat{x}$ did not have more than one Gaussian components. In addition, there can be problems concerning the scaling of the elements of the matrix M. Hence, it is the experience of the authors that imposing the condition

$$\det(M) = 1 \qquad(15)$$

makes the optimization numerically stable and avoids possible scaling problems. Different parameterizations of M such that the condition in (15) holds can be found in [19].

3. ALTERNATIVE REDUNDANCY MEASURES AND NONLINEAR ICA

The previous section has demonstrated that ICA and InfoMax are identical when the redundancy measure in ICA is the Kullback-Leibler information distance and when sufficient freedom is given to the marginal output probability modelling and estimation. Nevertheless, there are other measures that are easy to implement, especially in the case of a linear mixing with a matrix M. The following part of the paper briefly reviews ICA based on the properties of cumulant expansion of the joint

probability density function $p(\hat{y})$. The detailed derivation and analysis of the cumulant based ICA can be found in [17] and [19].

The cumulant based criterion for ICA is derived by comparison of the cumulant expansion of the joint probability density $p(\hat{y})$ and of the product of the marginal output probabilities $p(y_i)$. The complete factorization is achieved if the both expansions are the same, i.e. if the non-diagonal coefficients in the higher order cumulants of $p(\hat{y})$ take desired values (usually zero) imposed by the statistical independence of $p(y_i)$. Since the cumulant expansion of an arbitrary distribution has infinite number of elements, for practical purposes only cumulants up to the order four are considered. Hence, the resulting ICA cumulants based criterion has the following form:

$$J(M) = \sum_{i=1}^{4} \sum_{nondiag} \left[C^{(i)}_{nondiag} - C^{(i)}_{nondiag-desired} \right]^2 \quad (16)$$

where i defines the cumulant order, and where $C^{(i)}_{nondiag}$ and $C^{(i)}_{nondiag-desired}$ are the non-diagonal cumulant coefficients and their desired values for a given cumulant order i of the joint probability density function $p(\hat{y})$. In general, the desired coefficients $C^{(i)}_{nondiag-desired}$ are equal to zero. For every change of the matrix M, the non-diagonal coefficients are estimated and the cost function $J(M)$ further minimized. The cumulant based ICA criterion can be further simplified by using the properties of cumulant expansion when M is a rotation matrix [17]. It is the experience of the authors that the cumulant based ICA criteria are numerically superior to the Kullback-Leibler distance based ICA.

As the last point of this section, the authors would like to mention that the ICA problem can be formulated also in the case where the input-output map is not a matrix but an invertible nonlinear function F. A parameterization of such functions with the so called "triangular volume preserving network" is presented in [12] and [19]. The reference [19] presents several applications of the Nonlinear ICA.

4. REFERENCES

[1] G. Zipf: Human Behavior and the Principle of Least Effort. Addison-Wesley, Cambridge, Massachusetts, 1949.

[2] F. Attneave: Informational Aspects of Visual Perception. Psychological Review, 61, 183-193, 1954.

[3] H. Barlow: Unsupervised Learning. Neural Computation, 1, 295-311, 1989.

[4] H. Barlow, T. Kaushal and G. Mitchison: Finding Minimum Entropy Codes. Neural Computation, 1, 412-423, 1989.

[5] J. Atick and A. Redlich: Towards a theory of early visual processing. Neural Computation, 2, 308-320, 1990.

[6] J. Atick and A. Redlich: What Does the Retina Know about Natural Scenes. Neural Computation, 4, 196-210, 1992.

[7] A.N. Redlich: Redundancy Reduction as a Strategy for Unsupervised Learning. Neural Computation, 5, 289-304, 1993.

[8] A.N. Redlich: Supervised Factorial Learning. Neural Computation, 5, 750-766, 1993.

[9] G. Deco and D. Obradovic, Principal Component Analysis: A Factorial Learning Approach. International Conference on Artificial Neural Networks, Springer Verlag, vol. 2, 1059-1062, Sorrento, Italy, 1994.

[10] D. Obradovic and G. Deco: Linear Feature Extraction in Networks with Lateral Connections. IEEE World Congress on Computational Intelligence, vol. 2, 686-691, Florida, USA, 1994.

[11] P. Comon: Independent Component Analysis, A new concept? Signal Processing, 36, 287-314, 1994.

[12] G. Deco and W. Brauer: Higher Order Statistics with Neural Networks. Advances in Neural Information Processing 7, Eds.: G. Tesauro, D. Touretzky and T. Leen, MIT Press (Cambridge) 247-54, 1994.

[13] G. Deco and W. Brauer: Nonlinear Higher Order Statistical Decorrelation by Volume-Conserving Neural Networks. Neural Networks, 8, 525-535, 1995.

[14] G. Deco and B. Schürmann: Learning Time Series Evolution by Unsupervised Extraction of Correlations. Physical Review E, 51, 1780-1790, 1995.

[15] L. Parra, G. Deco and S. Miesbach: Redundancy Reduction with Information Preserving Nonlinear Maps. Networks: Computation in Neural Systems, 6, 61-72, 1995.

[16] G. Deco and D. Obradovic: Rotation Based Redundancy Reduction Learning. Neural Networks, 8, 751-755, 1995.

[17] D. Obradovic and G. Deco: Linear Feature Extraction in non-Gaussian Networks. World Congress on Neural Networks, vol 2, 523-527, Washington DC, 1995.

[18] D. Obradovic and G. Deco: An information theory based learning paradigm for linear feature extraction. Neurocomputing, 12, 203-221, 1996.

[19] G. Deco and D. Obradovic: An Information-Theoretic Approach to Neural Computing. Springer Verlag, New York, February 1996.

[20] A.J. Bell and T.J. Sejnowski: An Information-Maximization Approach to Blind Separation and Blind Deconvolution.Neural Computation, 7, Number 6, 1129-1160, November 1995.