Survey Paper on Fixed-Point Algorithms for Independent Component Analysis

Dr. Rakesh Shrivastava[1], Dr. Sanjay Choudhary[2], Dr. Shyam Patkar[3]

[1] *Professor, SIRT, Bhopal*

[2]*Professor & HOD, Govt. NMV, Hoshangabad*

[3]*Register, Bhabha University, Bhopal*

Abstract: - Independent component analysis (ICA) is a statistical method for transforming an observed multidimensional random vector into components that are statistically as independent from each other as possible. In this paper, we use a combination of two different approaches for linear ICA: Comon's information-theoretic approach and the projection pursuit approach. Using maximum entropy approximations of differential entropy, we introduce a family of new contrast (objective) functions for ICA. These contrast functions enable both the estimation of the whole decomposition by minimizing mutual information, and estimation of individual independent components as projection pursuit directions. The statistical properties of the estimators based on such contrast functions are analyzed under the assumption of the linear mixture model, and it is shown how to choose contrast functions that are robust and/or of minimum variance. Finally, we intro- duce simple fixed-point algorithms for practical optimization of the contrast functions. These algorithms optimize the contrast functions very fast and reliably.

I. INTRODUCTION

A central problem in neural network research, as well as in statistics and signal processing, is finding a suitable representation or transformation of the data. For computational and conceptual simplicity, the representation is often sought as a *linear* transformation of the original data. Let us denote by $\mathbf{x} = (x_1, x_2 ... x_m)^T$ a zero-mean *m*-dimensional random variable that can be observed, and by $\mathbf{s} = (s_1, s_2, ..., s_n)^T$ its *n*-dimensional transform. Then the problem is to determine a constant (weight) matrix W so that the linear transformation of the observed variables has some suitable properties. Several principles and methods have been developed to find such a linear representation, including principal component analysis [1], factor analysis [2, 3], projection pursuit [4], independent component analysis [5], etc.

$$s = Wx \qquad (1)$$

The transformation may be defined using such criteria as optimal dimension reduction, statistical 'interestingness' of the resulting components s_i, simplicity of the transformation, or other criteria, including application-oriented ones. We treat in this paper the problem of estimating the transformation given by (linear) independent component analysis (ICA) [7]. As the name implies, the basic goal in determining the transformation is to find a representation in which the transformed components s_i are statistically as independent from each other as possible. Thus this method is a special case of redundancy reduction [2].

Two promising applications of ICA are blind source separation and feature extraction. In *blind source separation* [8], the observed values of **x** correspond to a realization of an *m*-dimensional discrete-time signal $\mathbf{x}(t)$, t= 1, 2, Then the components $s_i(t)$ are called source signals, which are usually original, uncorrupted signals or noise sources. Often such sources are statistically independent from each other, and thus the signals can be recovered from linear mixtures x_i by finding a transformation in which the transformed signals are as independent as possible, as in ICA. In *feature extraction* [9], s_i is the coefficient of the *i*-th feature in the observed data vector **x**. The use of ICA for feature extraction is motivated by results in neurosciences that suggest that the similar principle of redundancy reduction [10] explains some aspects of the early processing of sensory data by the brain. ICA has also applications in *exploratory data analysis* in the same way as the closely related method of projection pursuit [16, 12].

II. CONTRAST FUNCTIONS FOR ICA

ICA data model, minimization of mutual information, and projection pursuit

One popular way of formulating the ICA problem is to consider the estimation of the following generative model for the data [1, 3, 5, 6]:

$$x = As \qquad (2)$$

Where **x** is an observed *m*-dimensional vector, **s** is an *n*-dimensional (latent) random vector whose components are assumed mutually independent, and **A** is a constant $m \times n$ matrix to be estimated. It is usually further assumed that the dimensions of **x** and **s** are equal, i.e., $m = n$; we make this assumption in the rest of the paper. A noise vector may also be present. The matrix **W** defining the transformation as in (1) is then obtained as the (pseudo)inverse of the

estimate of the matrix **A**. Non-Gaussianity of the independent components is necessary for the identifiability of the model (2), see [7].

Comon [7] showed how to obtain a more general formulation for ICA that does not need to assume an underlying data model. This definition is based on the concept of mutual information. First, we define the differential entropy H of a random vector $\mathbf{y} = (y_1 \ldots y_n)^T$ with density $f(.)$ as follows [8]:

$$H(y) = -\int f(y) \log f(y)\, dy \qquad (3)$$

Differential entropy can be normalized to give rise to the definition of negentropy, which has the appealing property of being invariant for linear transformations. The definition of negentropy J is given by

$$J(y) = H(Y_{gauss}) - H(y) \qquad (4)$$

Where $\mathbf{y}_{gauss}$ is a Gaussian random vector of the same covariance matrix as **y**. Negentropy can also be interpreted as a measure of non gaussianity [7]. Using the concept of differential entropy, one can define the mutual information I between the n (scalar) random variables y_i, $i = 1 \ldots n$ [8, 7]. Mutual information is a natural measure of the dependence between random variables. It is particularly interesting to express mutual information using negentropy, constraining the variables to be *uncorrelated*. In this case, we have [7]

$$I(y_1, y_2 \ldots\ldots y_n) = J(y) - \sum_i J(y_i) \qquad (5)$$

Since mutual information is the information-theoretic measure of the independence of random variables, it is natural to use it as the criterion for finding the ICA transform. Thus we define in this paper, following [7], the ICA of a random vector **x** as an invertible transformation **s** = **Wx** as in (1) where the matrix **W** is determined so that the *mutual information of the transformed components* s_i *is minimized*. Note that mutual information (or the independence of the components) is not affected by multiplication of the components by

scalar constants. Therefore, this definition only defines the independent components up to some multiplicative constants. Moreover, the constraint of uncorrelated ness of the s_i is adopted in this paper. This constraint is not strictly necessary, but simplifies the computations considerably.

Because negentropy is invariant for invertible linear transformations [7], it is now obvious from (5) that finding an invertible transformation **W** that minimizes the mutual information is roughly equivalent to *finding directions in which the negentropy is maximized*. This formulation of ICA also shows explicitly the connection between ICA and projection pursuit [11, 12, 16]. In fact, finding a single direction that maximizes negentropy is a form of projection pursuit, and could also be interpreted as estimation of a single independent component [2].

Contrast Functions through Approximations of Negentropy

To use the definition of ICA given above, a simple estimate of the negentropy (or of differential entropy) is needed. We use here the new approximations developed in [19], based on the maximum entropy principle. In [19] it was shown that these approximations are often considerably more accurate than the conventional, cumulate based approximations in [7, 1]. In the simplest case, these new approximations are of the form:

$$J(y_i) = c[E\{G(y_i) - E\{G(V)\}]^2 \qquad (6)$$

Where G is practically any non-quadratic function, c is an irrelevant constant, and $\square$ is a Gaussian variable of zero mean and unit variance (i.e., standardized). The random variable y_i is assumed to be of zero mean and unit variance. For symmetric variables, this is a generalization of the cumulate based approximation in [7], which is obtained by taking $G(y_i) = y^4$.

III. FIXED POINT ALGORITHM FOR ICA

In the preceding sections, we introduced new contrast (or objective) functions for ICA based on minimization of mutual information (and projection pursuit), analyzed some of their properties, and gave guidelines for the practical choice of the function G used in the contrast functions. In practice, one also needs an algorithm for maximizing the contrast functions in (7) or (8).

The advantage of neural on-line learning rules is that the inputs $\mathbf{x}(t)$ can be used in the algorithm at once, thus enabling faster adaptation in a non-stationary environment. A resulting trade-off, however, is that the convergence is slow, and depends on a good choice of the learning rate sequence, i.e. the step size at each iteration. A bad choice of the learning rate can, in practice, destroy convergence. Therefore, it would important in practice to make the learning faster and more reliable. This can be achieved by the fixed-point iteration algorithms that we introduce here. In the fixed-point algorithms, the computations are made in batch (or block) mode, i.e., a large number of data points are used in a single step of the algorithm. In other respects, however, the algorithms may be considered neural. In particular, they are parallel, distributed, computationally simple, and require little memory space. We will show below that the fixed-point algorithms have very appealing convergence properties, making them a very interesting alternative to adaptive learning rules in environments where fast real-time adaptation is not necessary.

Properties of the Fixed-Point Algorithm

The fixed-point algorithm and the underlying contrast functions have a number of desirable properties when

compared with existing methods for ICA.

- The convergence is cubic (or at least quadratic), under the assumption of the ICA data model (for a proof, see the convergence proof in the Appendix). This is in contrast to gradient descent methods, where the convergence is only linear. This means a very fast convergence, as has been confirmed by simulations and experiments on real data.
- Contrary to gradient-based algorithms, there are no step size parameters to choose (in the original fixed-point algorithm). This means that the algorithm is easy to use. Even in the stabilized version, reasonable values for the step size parameter are very easy to choose.
- The algorithm finds directly independent components of (practically) any non-Gaussian distribution using any nonlinearity g. This is in contrast to many algorithms, where some estimate of the probability distribution function has to be first available, and the nonlinearity must be chosen accordingly.
- The performance of the method can be optimized by choosing a suitable nonlinearity g. In particular, one can obtain algorithms that are robust and/or of minimum variance.
- The fixed-point algorithm inherits most of the advantages of neural algorithms: It is parallel, distributed, computationally simple, and requires little memory space. Stochastic gradient methods seem to be preferable only if fast adaptively in a changing environment is required.

IV. SIMULATION AND EXPERIMENTAL RESULT

First, we investigated the robustness of the contrast functions. We generated four artificial source signals, two of which were sub-Gaussian, and two were super-Gaussian. The source signals were mixed using several different random matrices, whose elements were drawn from a standardized Gaussian distribution. To test the robustness of our algorithms, *four outliers* whose values were ±10 were added in random locations. The fixed-point algorithm for sphered data was used with the three different contrast functions in eq. (14–16), and symmetric orthogonalization. Since the robust estimation of the covariance matrix is a classical problem independent of the robustness of our contrast functions, we used in this simulation a hypothetical robust estimator of covariance, which was simulated by estimating the covariance matrix from the original data without outliers. In all the runs, it was observed that the estimates based on kurtosis (16) were essentially worse than the others, and estimates using G_2 in (15) were slightly better than those using G_1 in (14). These results confirm the theoretical predictions on robustness in Section 3.

To investigate the *asymptotic variance*, i.e., efficiency, of the estimators, we performed simulations in which the 3 different contrast functions were used to estimate one independent component from a mixture of 4 identically distributed independent components. We also used three different distributions of the independent components: uniform, double exponential (or Laplace), and the distribution of the third power of a Gaussian variable. The asymptotic mean absolute deviations (which are a robustified measure of error) between the components of the obtained vectors and the correct solutions were estimated and averaged over 1000 runs for each combination of non-linearity and distribution of independent component. The results in the basic, noiseless case are depicted in Fig. 1. As one can see, the estimates using kurtosis were essentially worse for super-Gaussian independent components. Especially the strongly super-Gaussian independent component (cube of Gaussian) was estimated considerably worse using kurtosis. Only for the sub-Gaussian independent component, kurtosis was better than the other contrast functions. There was no clear difference between the performances of the contrast functions G_1 and G_2. Next, the experiments were repeated with added Gaussian noise whose energy was 10% of the energy of the independent components. The results are shown in Fig. 2. This time, kurtosis did not perform better even in the case of the sub-Gaussian density. The robust contrast functions seem to be somewhat robust against Gaussian noise as well.

We also studied the *speed of convergence* of the fixed-point algorithms. Four independent components of different distributions (two sub Gaussian and two super Gaussian) were artificially generated, and the symmetric version of the fixed-point algorithm for sphered data was used. The data consisted of 1000 points, and the whole data was used at every iteration. We observed that for all three contrast functions, only *three* iterations were necessary, on the average, to achieve the maximum accuracy allowed by the data. This illustrates the fast convergence of the fixed-point algorithm. In fact, a comparison of our algorithm with other algorithms was performed in [13], showing that the fixed-point algorithm gives approximately the same statistical efficiency as other algorithms, but with a fraction of the computational cost.

V. CONCLUSION

The problem of linear independent component analysis (ICA), which is a form of redundancy reduction, was addressed. Following Comon [7], the ICA problem was formulated as the search for a linear transformation that minimizes the mutual information of the resulting components. This is roughly equivalent to finding

directions in which negentropy is maximized and which can also be considered projection pursuit directions [16]. The novel approximations of negentropy introduced in [19] were then used for constructing novel contrast (objective) functions for ICA. This resulted in a generalization of the kurtosis-based approach in [7, 9], and also enabled estimation of the independent components one by one. The statistical properties of these contrast functions were analyzed in the framework of the linear mixture model, and it was shown that for suitable choices of the contrast functions, the statistical properties were superior to those of the kurtosis- based approach. This was the family of fixed-point algorithms that are not neural in the sense that they are non-adaptive, but share the other benefits of neural learning rules. The main advantage of the fixed-point algorithms is that their convergence can be shown to be very fast (cubic or at least quadratic). Combining the good statistical properties (e.g. robustness) of the new contrast functions, and the good algorithmic properties of the fixed- point algorithm, a very appealing method for ICA was obtained. Simulations as well as applications on real-life data have validated the novel contrast functions and algorithms introduced. Some extensions of the methods introduced in this paper are presented in [20], in which the problem of noisy data is addressed, and in [22], which deals with the situation where there are more independent components than observed variables.

REFERENCES

[1] S.-I. Amari, A. Cichocki, and H.H. Yang. A new learning algorithm for blind source separation. In *Advances in Neural Information Processing Systems 8*, pages 757–763. MIT Press, 1996.

[2] H. B. Barlow. Possible principles underlying the transformations of sensory messages. In W. A. Rosen- blith, editor, *Sensory Communication*, pages 217–234. MIT Press, 1961.

[3] A.J. Bell and T.J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7:1129–1159, 1995.

[4] A.J. Bell and T.J. Sejnowski. The 'independent components' of natural scenes are edge filters. *Vision Research*, 37:3327–3338, 1997.

[5] J.-F. Cardoso and B. Hvam Laheld. Equivariant adaptive source separation. *IEEE Trans. on Signal Processing*, 44(12):3017–3030, 1996.

[6] A. Cichocki and R. Unbehauen. *Neural Networks for Signal Processing and Optimization*. Wiley, 1994.

[7] P. Comon. Independent component analysis—a new concept? *Signal Processing*, 36:287–314,1994.

[8] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley,1991.

[9] N. Delfosse and P. Loubaton. Adaptive blind separation of independent sources: a deflation approach. *Signal Processing*, 45:59–83, 1995.

[10] The FastICA MATLAB package. Available at http://www.cis.hut.fi/projects/ica/fastica/.

[11] J. H. Friedman and J. W. Tukey. A projection pursuit algorithm for exploratory data analysis. *IEEE Trans. of Computers*, c-23(9):881–890, 1974.

[12] J.H. Friedman. Exploratory projection pursuit. *J. of the American Statistical Association*, 82(397):249– 266, 1987.

[13] H. H. Harman. *Modern Factor Analysis*. University of Chicago Press, 2nd edition, 1967.

[14] P.J. Huber. Projection pursuit. *The Annals of Statistics*, 13(2):435–475,1985.

[15] A. Hyvärinen. A family of fixed-point algorithms for independent component analysis. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP'97)*, pages 3917–3920, Munich, Ger- many, 1997.

[16] A. Hyvärinen. One-unit contrast functions for independent component analysis: A statistical analysis. In *Neural Networks for Signal Processing VII (Proc. IEEE Workshop on Neural Networks for Signal Processing)*, pages 388–397, Amelia Island, Florida, 1997.

[17] A. Hyvärinen. New approximations of differential entropy for independent component analysis and projection pursuit. In *Advances in Neural Information Processing Systems*, volume 10, pages 273–279. MIT Press, 1998.

[18] A. Hyvärinen. Fast independent component analysis with noisy data using gaussian moments. In *Proc. Int. Symp. on Circuits and Systems*, pages V57–V61, Orlando, Florida,1999.

[19] A. Hyvärinen. The fixed-point algorithm and maximum likelihood estimation for independent compo- nent analysis. *Neural Processing Letters*, 10(1):1–5, 1999.

[20] A. Hyvärinen, R. Cristescu, and E. Oja. A fast algorithm for estimating overcomplete ICA bases for image windows. In *Proc. Int. Joint Conf. on Neural Networks*, pages 894–899, Washington, D.C., 1999.

[21] A. Hyvärinen and E. Oja. A fast fixed-point algorithm for independent component analysis. *Neural Computation*, 9(7):1483–1492, 1997.

[22] A. Hyvärinen and E. Oja. Independent component analysis by general nonlinear Hebbian-like learning rules. *Signal Processing*, 64(3):301–313, 1998.

[23] A. Hyvärinen, E. Oja, P. O. Hoyer, and J. Hurri. Image feature extraction by sparse coding and inde- pendent component analysis. In *Proc. Int. Conf. on Pattern Recognition (ICPR'98)*, pages 1268–1273, Brisbane, Australia, 1998.

[24] M.C. Jones and R. Sibson. What is projection pursuit ? *J. of the Royal Statistical Society, Ser. A*, 150:1–36, 1987.

[25] C. Jutten and J. Hérault. Blind separation of sources, part I: An adaptive algorithm based on neu- romimetic architecture. *Signal Processing*, 24:1–10, 1991.