AN ELECTRIC FIELD APPROACH TO INDEPENDENT COMPONENT ANALYSIS

Sepp Hochreiter and Michael C. Mozer

Department of Computer Science
University of Colorado, Boulder, CO 80309
{hochreit,mozer}@cs.colorado.edu

ABSTRACT

We propose a novel algorithm for Independent Component Analysis (ICA) that is based on an electric field metaphor. As with all ICA techniques, the algorithm searches for a demixing model that produces components whose joint distribution matches the factorial distribution (i.e., the product of the marginal distributions). The joint and factorial distributions are represented as positively and negatively charged particles, respectively, and the dynamics of the search are based on the interactions among particles. The algorithm can deal with arbitrary distributions for the sources, nonlinear mixing functions, noisy observations, and an unequal number of source and mixture components. The limitation of the algorithm is that it does not scale with the number of sources. Nonetheless, we demonstrate that the algorithm can solve challenging ICA problems that are beyond the capabilities of other ICA methods.

1. INTRODUCTION

Independent component analysis (*ICA*) is a method that discovers a representation of multivariate data in which the statistical dependence among the components is minimized. The observed data, a D-dimensional random variable x, is assumed to have been generated by a *mixing* process that operates on a set of D' independent *source* variables. The task of ICA is to determine a demixing transformation that recovers the source variables.

The primary approaches to ICA fall into two categories: model-based methods and global density-approximation methods. *Model-based methods* assume a parameterized model of the mixing function, e.g., a linear transformation, and the distribution of the source variables, e.g., Gaussian. Prominent model-based algorithms include maximum entropy approaches [1, 21] and maximum posteriori approaches [17, 27]. Maximum likelihood estimation is used to determine the model parameters. *Global density-approximation methods* involve characterizing data distributions in terms of a relatively small number of parameters. These methods rely on an *indicator function* that can compute the degree of independence of the components in a straightforward manner from parameters of the global density approximation. These methods search for a transformation of the observed data that yields a good measure of independence. Most global density-approximation methods rely on cumulants and Cramer-Edgeworth or Gram-Charlier expansion for approximating densities [6, 7, 26, 25], and/or utilize as a measure of independence cumulant tensors [4], or other nonlinear functions of the parameters [3, 5, 13]. Model-based and global density-approximation methods can be unified in a *general contrast function* framework [10], which permits a fixed-point algorithm for minimizing the contrast function [11], thereby enforcing independence of the demixing components.

The goal of ICA research is to discover methods that: (1) allow for arbitrary source distributions, (2) can accommodate nonlinear mixing functions, (3) perform well when the number of mixing components is smaller than the number of source components, (4) can handle noise in the observations, (5) operate in a computationally efficient manner, and (5) scale to high dimensional data. All existing algorithms fail on at least several of these criteria.

Model-based methods generally assume linear mixing functions and unimodal distributed sources (although some work has been done to overcome the source restriction, e.g., [15]). The strength of model-based methods is that they are fast and in most cases robust to variation in initial conditions and learning parameters. Global density-approximation methods face a trade off: describing the joint distribution of the observed data with a large number of parameters allows the method to represent arbitrary distributions, but is computation intensive [7]; with a small number of parameters, the method can represent only limited distributions [6, 25] but is computationally efficient.

Existing ICA methods attempt to address the first two criteria by allowing for various source distributions and mixing functions. However, they are limited to *specific* distributions and mixing functions. E.g., methods have been designed for uniform [20, 19, 18] and binary [23, 9] distributions, and for specific mixing nonlinearities [2, 26, 16, 24]. Deco [7] has developed a method that can be applied to arbitrary source distributions and nonlinearities, but it is slow and requires an equal number of sources and mixtures, and prior knowledge about the mixture function cannot be exploited.

We propose an ICA method that allows for arbitrary source distributions and arbitrary mixing nonlinearities. Our method is computationally efficient, performs well even if the number of mixing components is smaller than the number of source components, and can handle observation noise. The limitation of our method is that it is tractable only for low dimensional data (roughly, $D < 6$).

Our method does not satisfy *all* of the criteria listed above, but should one expect otherwise? It seems highly unlikely that a method will exist that is extremely general *and* fast. Each method in the literature handles special cases of the general ICA problem. The special case we have focused on—low dimensional observations—seems particularly interesting and useful, as many domains of application involve low dimensional data. Examples include identification of users entering a building [8], separating a signal from noise [14], and symbol separation in code-division multiple access (CDMA) telecommunications systems [22].

2. AN ELECTRIC-FIELD METHOD

Our approach involves hypothesizing an arbitrary nonlinear function approximator, such as a neural network, as a demixing model. After the observed data has been passed through this model, we construct a joint density estimate of the transformed data. We can also use the transformed data to construct a factorial density estimate: the product of marginal density estimates. The quality of the demixing model is evaluated by determining the discrepancy between the joint and factorial density estimates, and a search is performed to identify the model that minimizes the discrepancy.

A gradient-based method is applied to optimize the parameters of the demixing model so as to minimize the discrepancy (i.e., to achieve independence of the recovered source components). Given an objective function that quantifies the discrepancy, one can compute the derivative of this function with respect to each transformed data value. If one can also compute the derivative of the demixing model output with respect to the demixing model parameters, the two derivates can be chained to obtain a derivative of the objective function with respect to the demixing model parameters.

We call our approach an *electric-field method*, because the ICA problem is characterized as an electric field. We treat the joint density estimate as a distribution of positive charges, treat the factorial density estimate as a distribution of negative charges, and treat the search as arising from the electric field generated by superimposing the two distributions of charges.

Rather than describing the algorithm in terms of continuous charge distributions, we instead describe the algorithm in terms of samples from the distributions, i.e., particles. To understand the algorithm, consider each data point, which has been transformed by the demixing model, as a *positively charged particle* in the demixed space (Figure 1). Under the assumption of independence of the demixed components, we could randomly recombine components (i.e., values along each dimension of the space) of the positively charged particles to generate *negatively charged particles*. The positive and negative particles represent samples from the joint and factorial distributions, respectively. The interactions between these particles result in a solution to the ICA problem. Specifically, the positive particles must repel one another or a degenerate solution will be obtained in which the demixing function collapses the data to a single point. The positive and negative particles must attract one another to bring the joint and the factorial distributions into alignment. The repelling and attractive forces between particles are proportional to $\frac{1}{d^{D-1}}$, where d is the distance between particles. The electric forces on the particles are translated into forces on the parameters of the demixing function, and a solution corresponds to the state in which the forces on particles cancel. Technically, the integral of the electric field, called the *potential function*, serves as the objective function for the search. Consequently, the electric field serves as the gradient of the objective function with respect to each data point, and gradient-based methods can be applied to shift the data points—via the demixing model parameters—downhill. If a good solution is found for the demixing model parameters, the positive charges will end up superimposed on the negative charges, and the net charge distribution will be uniformly zero throughout the mixing space.

Our approach is related to the idea of an information force [18], where each data point is seen as repelling each other, leading to a uniform joint distribution. However, our method is not restricted to finding uniform distributions.

Although our approach seems elegant, the compu-

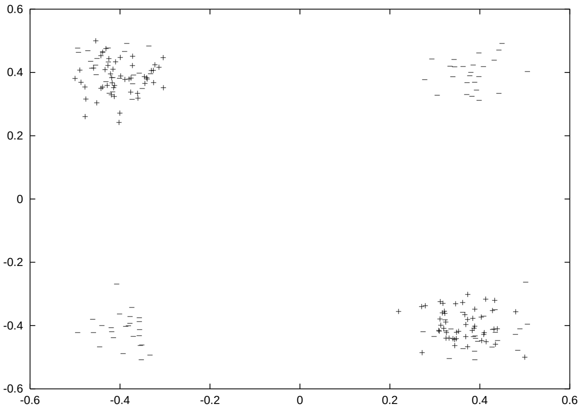

Figure 1: Points sampled from the joint density are positive particles and points sampled from the marginal product are negative particles.

tational cost is high. Given M data points corresponding to positive particles, and assuming we generate an equal number of negative particles, the number of interactions between particles is $O(M^2)$.

To make the computation tractable, we introduce several approximations to decrease the number of interactions. The approximations are based on a finite-difference method in which we define N equal sized intervals on each dimensions of the demixed space, which divides the space into D^N hypercubes. The approximations we make are as follows. First, we reduce the number of particles by canceling out positive and negative particles (randomly) within each hypercube. The result is that each hypercube contains only positive or negative particles, or none at all. Second, rather than considering the interaction between pairs of individual particles, we consider the interaction between an individual particle and the mass effect of all particles in a hypercube. Third, rather than considering the interaction between a particle and the particles in every hypercube, we consider only the interactions with the neighboring hypercubes. This approximation is justified because the electric field decreases hyperbolically with the distance; consequently, the neighboring hypercubes have a much greater effect on the field than more distant hypercubes. We only compute the force of the 2 D^2 nearest hypercubes on a particle.

Another Perspective on the Method

The electric-field method we are proposing could also be viewed from a somewhat different perspective as constructing local density approximations for the joint and factorial distributions. The local density approximation is a histogram, or a uniform bounded kernel, estimate. That is, we estimate the joint density by determining relative frequency of the demixed data in each hypercube. By comparing the joint density (the positive particles) with the factorial density (the negative particles), we can bring the two distributions into agreement. Using a local-density approximation has several potential problems. We briefly discuss these problems and how our approach minimizes the problems.

The Curse of Dimensionality

The memory, computational, and/or sample requirements of a local-density approximation grow exponentially in D, the number of mixture components. The number of hypercubes required in our approach is N^D, and the amount of data required to populate the hypercubes will also grow with N^D.[1]

Although we cannot altogether avoid the curse of dimensionality, we can alleviate it to some degree based on two observations. First, we have found empirically that small N tend to work surprisingly well. Second, we adopt a multiresolution approach in which we start with a small N and then split into a larger number of intervals as necessary.

Local Optima

Local-density approximations have two characteristics that often result in local optima in a search space. First, local density approximations often produce regions of uniform density; for example, with the histogram approximation, all points within a hypercube are assigned the same density. Second, each data point affects the density estimate in only a limited region of the space. As a result of these two characteristics, gradient-based techniques tend to become stuck in local optima, because shifting a data point a small amount (resulting from a small adjustment the demixing model parameters) will not affect the density approximation, either locally—due to the first characteristic—or nonlocally—due to the second characteristic. We mitigate these problems to a degree with the electric-field approach because the interactions are not hypercube-to-hypercube, but rather, hypercube-to-particle.

3. DETAILS OF THE APPROACH

The parameter N, the number of intervals into which each dimension is partitioned, controls a bias-variance

[1]Our method is not the only one that could suffer from the curse of dimensionality. For example, in [7], an Nth order approximation of a D-dimensional distribution requires that N^D cumulants be calculated, resulting in an exponential slowing of the algorithm with the dimensionality of the mixture space.

trade off. If N is large, the variance will be high and the ICA solution will depend on the particular data sample available; if N is small, the bias will be high, and the algorithm will have difficulty accurately modeling the probabilities densities. Therefore, we begin the search for the demixing model using a small N, $N = 2$, and double N after a fixed number of iterations until N reaches a fixed upper limit specified by the optimal bandwidth h_o of the histogram estimator, which minimizes the mean integrated squared error. h_o is obtained from a bias-variance analysis for kernel density estimators.

For each iteration of the electric-field method, the D-dimensional grid that defines the hypercubes is shifted on each dimension i by a random value in $[-\frac{h_i}{2}, \frac{h_i}{2}]$, where h_i is the hypercube edge length on dimension i. This randomization has the important effect of smoothing estimated values of neighboring hypercubes.

In computing the distance between an individual particle α and the set of particles in a hypercube, we treat all particles in the hypercube as being located in the same position—along the edge of the hypercube at the point nearest to particle α.

As a demixing model we used a one-hidden-layer neural network trained by back propagation. To prevent the degenerate solution in which the demixing function maps all data to a single point, and to make the full grid, we rescale the grid to the range of mixing model outputs.

The computational complexity of our algorithm is $O(\log(M) + W)$ per data point, where $O(W) \geq O(D^2)$ is the number of parameters in our demixing function.

4. EXPERIMENTS

4.1. Supergaussian Source Distribution

We begin with an experiment using five supergaussian sources. Denoting the Gaussian density with mean μ and standard deviation σ by $G(x; \mu, \sigma)$, the source densities are given by

$$\frac{1}{3x^{\frac{2}{3}}}G(x^{\frac{1}{3}}; 0.5, 0.8),\ \frac{1}{3x^{\frac{2}{3}}}G(x^{\frac{1}{3}}; 0.1, 1),\ \frac{1}{3x^{\frac{2}{3}}}G(x^{\frac{1}{3}}; -0.5, 0.5),$$
$$\frac{1}{3x^{\frac{2}{3}}}G(x^{\frac{1}{3}}; -0.8, 0.6),\ \frac{1}{3x^{\frac{2}{3}}}G(x^{\frac{1}{3}}; 0.5, 0.4).$$

We used as a demixing model a linear neural net trained with 1000 fixed examples and a learning rate of 0.00001. N was fixed at 2; there was no doubling of N. The network was trained for 100000 epochs, which took about an hour on a PC. With a larger learning rate, results are obtained faster but they are not as exact.

The mixing matrix multiplied by the demixing is:

-0.0062	0.0054	-0.0005	**0.2109**	-0.0104
-0.1617	0.0339	0.0216	-0.0063	0.0093
-0.0079	-0.0108	**0.2477**	0.0018	0.0095
-0.0675	**-0.1355**	-0.0327	0.0067	-0.0631
-0.0273	-0.0361	-0.0055	-0.0033	**0.2889**

Perfect inversion of the mixing process is indicated by an identity matrix, subject to a permutation and scaling transformation. In such a matrix, each row will contain one nonzero value and the remainder of values will all be zero. As one can see, our result indicates near ideal performance.

4.2. Subgaussian Source Distribution with Supergaussian Modes

In this experiment, we used three sources, two of which had two supergaussian modes, and one of which had one supergaussian mode (top row of Figure 2). The modes were defined as $\frac{1}{3\ x^{\frac{2}{3}}}G(x^{\frac{1}{3}}; \mu, \sigma)$. The parameters were the same as in the previous experiment but we started with $N = 8$, which was doubled every 50000 epochs during the 200000 training epochs. Figure 2 shows the results.

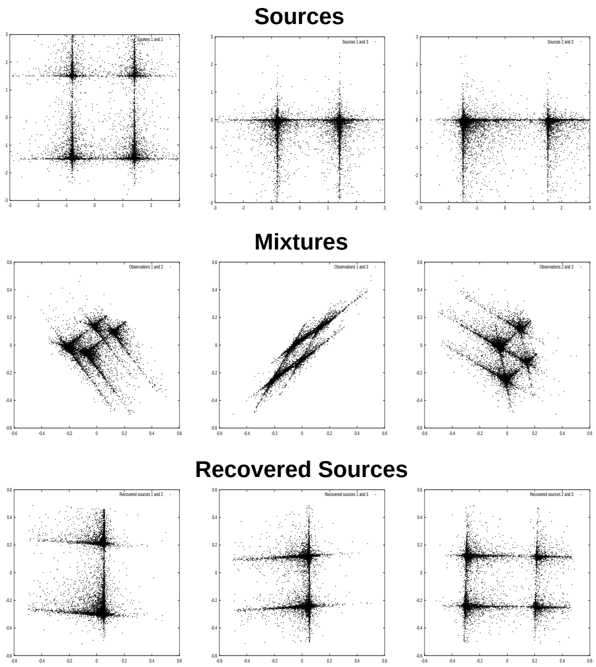

Figure 2: For a three-dimensional linear mixture projections of sources (1st line), mixtures (2nd line) and sources recovered by our approach (3rd line) on a two-dimensional plane are shown.

4.3. Subgaussian Source Distribution with Gaussian Modes

In this experiment, we used four sources, three with two modes and one with one mode, where each mode was Gaussian. The source densities were:

$$\frac{1}{2}G(x;0.4,0.2)+\frac{1}{2}G(x;-0.8,0.2),$$
$$\frac{1}{3}G(x;0.4,0.1)+\frac{2}{3}G(x;-0.3,0.1),$$
$$\frac{1}{4}G(x;1.5,0.3)+\frac{3}{4}G(x;0,1),\ \frac{1}{3x^{\frac{2}{3}}}G(x^{\frac{1}{3}};0,1).$$

All parameters were as in the previous experiment but the learning rate used was 0.01, and the simulation was started with $N=2$ and N was doubled every 3000 epochs. We trained for 10000 epochs. The product of the mixing and demixing matrices is:

0.4435	0.0066	0.0015	0.0024
-0.0099	-0.0098	**-0.1497**	-0.0027
0.0033	**0.6142**	0.0021	0.0035
-0.0031	0.0012	-0.0007	**0.1503**

4.4. Nonlinear Mixing

In this final experiment, we tried to recover sources from two nonlinear mixtures. The two-dimensional mixing function we used was: $f_1(z)=(z+a)^2$ and $f_2(z)=\sqrt{z+a}$, where z is a complex variable denoting the source datum, and a is the complex constant $3+3i$. The same density was used for both sources:

$$\frac{1}{4x^{\frac{2}{3}}}G(x^{\frac{1}{3}};0,0.4)+\frac{1}{8}G(x;2,0.03)+\frac{1}{8}G(x;-2,0.03)$$

We used a one-hidden-layer sigmoidal neural net with 50 hidden units as the demixing model. A set of 100000 training points was generated, of which 1000 was used on each epoch of training. The learning rate was 0.01. The simulation was started with $N=2$ and N was doubled every 100000 epochs. We trained for 500000 epochs. Figure 3 shows the sources, mixtures, and recovered sources. One cannot expect an exact inversion of the mixing function, because nonlinear independent component analysis has no unique solution unless the form of the demixing function is restricted [12].

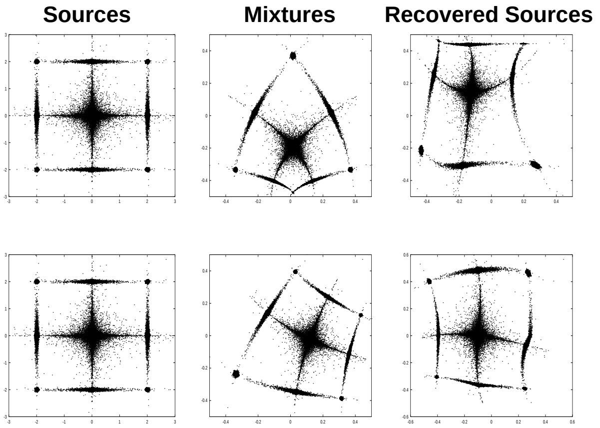

Figure 3: For two two-dimensional nonlinear mixing functions (first line $(z+a)^2$ second line $\sqrt{z+a}$) the sources, mixtures and recovered sources are shown. The mixing function is not completely inverted but the sources are recovered recognizable.

5. CONCLUSION

In this paper, we introduced a novel algorithm for ICA that is based on an electric field analogy. The algorithm can in principle handle arbitrary distributions for the sources, nonlinear mixing functions, noisy observations, and an unequal number of source and mixture components. The experiments showed that our algorithm is not limited to special distributions and can recover nonlinear mixed sources.

Our algorithm does not scale well with the dimensionality of the source space, and hence it is most useful for low-dimensional problems. If the separation can be done stepwise, i.e., by treating all but one source as noise or as a single second source, this limitation is not a serious problem. Moreover, many real world applications are in fact low dimensional, e.g., separation of a signal from noise, user/symbol/channel identification in telecommunications, or multi-tag radio-frequency authorization for entering a building.

6. REFERENCES

[1] A. J. Bell and T. J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159, 1995.

[2] G. Burel. Blind separation of sources: A nonlinear neural algorithm. *Neural Networks*, 5(6):937–947, 1992.

[3] J.-F. Cardoso and B. Laheld. Equivariant adaptive source separation. *IEEE Transactions on Signal Processing*, 44:3017–3030, 1996.

[4] J.-F. Cardoso and A. Souloumiac. Blind beamforming for non Gaussian signals. *IEE Proceedings-F*, 140(6):362–370, 1993.

[5] A. Cichocki, R. Unbehauen, L. Moszczynski, and E. Rummert. A new on-line adaptive algorithm for blind separation of source signals. In *Proc. Int. Symposium on Artificial Neural Networks, ISANN-94*, pages 406–411, 1994.

[6] P. Comon. Independent component analysis – a new concept? *Sig. Proc.*, 36(3):287–314, 1994.

[7] G. Deco and W. Brauer. Nonlinear higher-order statistical decorrelation by volume-conserving neural architectures. *Neural Networks*, 8(4):525–535, 1995.

[8] Y. Deville. Improved multi-tag radio-frequency identification systems based on new source separation neural networks. In ICA'99: 19–24, 1999.

[9] S. Hochreiter and J. Schmidhuber. Feature extraction through LOCOCODE. *Neural Computation*, 11(3):679–714, 1999.

[10] A. Hyvärinen. Survey on independent component analysis. *Neural Comp. Surveys*, 2:94–128, 1999.

[11] A. Hyvärinen and E. Oja. A fast fixed-point algorithm for independent component analysis. *Neural Computation*, 9(7):1483–1492, 1999.

[12] A. Hyvärinen and P. Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Netw.*, 12(3):429–439, 1999.

[13] C. Jutten and J. Herault. Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture. *Signal Processing*, 24(1):1–10, 1991.

[14] T.-W. Lee, A. J. Bell, and R. H. Lambert. Blind separation of delayed and convolved sources. In NIPS'9, pages 758–764, 1997.

[15] T.-W. Lee, M. Girolami, and T. J. Sejnowski. Independent component analysis using an extended infomax algorithm for mixed subgaussian and supergaussian sources. *Neural Computation*, 11(2):417–441, 1999.

[16] T.-W. Lee, B.-U. Köhler, and R. Orglmeister. Blind source separation of nonlinear mixing models. In *IEEE Proc. ICNN, Houston*, pages 406–415, 1997.

[17] T.-W. Lee, M. S. Lewicki, M. Girolami, and T. J. Sejnowski. Blind source separation of more sources than mixtures using overcomplete representations. *IEEE Signal Processing Letters*, 1998. in press.

[18] G. C. Marques and L. B. Almeida. Separation of nonlinear mixtures using pattern repulsion. In ICA'99: 277–282, 1999.

[19] P. Pajunen, A. Hyvärinen, and J. Karhunen. Nonlinear blind source separation by self-organizing maps. In S. Amari et al., editor, *Progress in Neural Information Processing*, volume 2, pages 1207–1210. Springer, 1997.

[20] P. Pajunen and J. Karhunen. A maximum likelihood approach to nonlinear blind source separation. In W. Gerstner, A. Germond, M. Hasler, and J.-D. Nicoud, editors, *Proceedings of the International Conference on Artificial Neural Networks, Lausanne, Switzerland*, pages 541–546. 1997.

[21] B. A. Pearlmutter and L. C. Parra. Maximum likelihood blind source separation: A context-sensitive generalization of ICA. In NIPS'9, pages 613–619, 1997.

[22] T. Ristaniemi and J. Joutsensalo. On the performance of blind source separation in CDMA downlink. In ICA'99: 437–442, 1999.

[23] J. Schmidhuber. Learning factorial codes by predictability minimization. *Neural Computation*, 4(6):863–879, 1992.

[24] A. Taleb and C. Jutten. Batch algoritm for source separation in postnonlinear mixtures. In ICA'99: 155–160, 1999.

[25] G. Yang and S. Amari. Adaptive online learning algorithms for blind source separation: Maximum entropy and minimum mutual information. *Neural Computation*, 9(7):1457–1482, 1997.

[26] H. H. Yang, S. Amari, and A. Cichocki. Information back-propagation for blind separation of sources from non-linear mixtures. In *Proceedings of the International Conference on Neural Networks*, pages 2141–2146. Houston, USA, 1996.

[27] M. Zibulevsky and B. A. Pearlmutter. Blind source separation by sparse decomposition. Technical Report CS99-1, Computer Science Department, Univ. of New Mexico, Albuquerque, 1999.

Abbreviations:
(1): ICA'99: J.-F. Cardoso, C. Jutten, and P. Loubaton, editors, *Proceedings of the First International Workshop on Independent Component Analysis and Signal Separation, Aussois, France.*
(2): NIPS'9: M. C. Mozer, M. I. Jordan, and T. Petsche, ed., *Advances in Neural Information Processing Systems 9*, MIT Press, Cambridge MA.