GEOMETRIC OPTIMIZATION METHODS FOR INDEPENDENT COMPONENT ANALYSIS APPLIED ON GENE EXPRESSION DATA

M. Journée[*], *A. E. Teschendorff*[†], *P.-A. Absil*[‡] *and R. Sepulchre*[*]

[*]Dept. of Electrical Engineering and Computer Science, University of Liège, Belgium.
[†]Dept. of Oncology, Hutchison-MRC Research Centre, University of Cambridge, UK.
[‡]Dept. of Mathematical Engineering, Université catholique de Louvain, Belgium.

ABSTRACT

DNA microarrays provide a huge amount of data and require therefore dimensionality reduction methods to extract meaningful biological information. Independent Component Analysis (ICA) was proposed by several authors as an interesting means. Unfortunately, experimental data are usually of poor quality because of noise, outliers and lack of samples. Robustness to these hurdles will thus be a key feature for an ICA algorithm. This paper identifies a robust contrast function and proposes a new ICA algorithm.

Index Terms— Independent Component Analysis (ICA), optimization on matrix manifolds, RADICAL algorithm, steepest-descent on the orthogonal group, gene expression data.

1. INTRODUCTION

The DNA microarray technology is intensively used by biomedical researchers for a systematic estimation of gene expression levels. Gene expression denotes the relevance of a specific gene for the biological functions to be fulfilled within the cell. Microarrays typically provide expression levels for several thousands of genes over a few hundreds of experiments. An important challenge is to extract some biological insight from these large databases. For more detail about microarrays and their analysis we refer to [1] and references therein.

Several authors have proposed Independent Component Analysis (ICA) as an interesting means to extract information from gene expression data [2, 3, 4]. The motivation behind this idea lies in the following intuition: gene expression results from several biological processes that take place independently. Each biological function relies on a subset of genes that are activated or inhibited and defines a so-called expression mode. These expression modes are expressed according to the biological tasks to complete within the cell. The latter are assumed to be independent. Gene expression is usually described as a linear function of the expression modes. The linear model is assumed for simplicity, but the application of ICA to gene expressions in yeast corroborates this model [2].

Let us define the gene expression matrix X such that its element (i, j) corresponds to the expression level of gene i in the jth experiment. X is thus a $n \times N$ matrix, where n is the number of analyzed genes and N is the number of experiments. The different experiments correspond, e.g., to different patients, tissues or environmental conditions. Note that n is usually much larger than N. ICA performs an approximate decomposition of the gene expression matrix into two smaller matrices A and S that are respectively $n \times p$ and $p \times N$ with $p < N$, i.e.,

$$X \approx AS. \tag{1}$$

The matrices A and S are selected to minimize the error between X and AS and to minimize a measure of dependence between the rows of S. The columns of matrix A define the expression modes. The element (i, j) of matrix S specifies the activity of the ith expression mode in the jth experiment. Since the expression modes are assumed to be involved independently, the lines of S may be interpreted as samples of independent sources.

Unfortunately, classical ICA algorithms are usually not well adapted to gene expression. These data present a huge number of observations while only few samples are available. They are furthermore highly subjected to noise and outliers (entries in the dataset that do not have any biological sense because of failures during the experimental process). The robustness of the ICA algorithm to these hurdles is an essential issue for the analysis of gene expression data.

Most ICA methods proceed by searching for a demixing matrix W such that the estimated sources $Z = W^T X$ are as independent as possible. Various contrast functions have been proposed that provide a quantitative measure of dependence between sampled signals. These contrasts are nonnegative and go to zero for statistically independent signals as the

This work was supported by the Belgian National Fund for Scientific Research (FNRS) through a Research Fellowship at the University of Liège and by Microsoft Research through a Research Fellowship at Peterhouse, Cambridge. This paper presents research results of the Belgian Programme on Interuniversity Attraction Poles, initiated by the Belgian Federal Science Policy Office. The scientific responsibility rests with its authors.

number of samples go to infinity. Hence, each ICA method consists in the minimization of a particular contrast by some optimization algorithm.

This work investigates in detail the measure of statistical independence proposed in the RADICAL algorithm [5]. This measure is based on an accurate and computationally efficient approximation of the mutual information by using spacings estimates of the differential entropy. It seems robust to the lack of samples as well as to noise and outliers. A gradient-descent algorithm based on that contrast function is next derived and compared to classical ICA algorithms on benchmark simulations.

This paper is organized as follows. In Section 2 the cost function of the RADICAL algorithm is recalled and its robustness is illustrated. In Section 3 a gradient-descent algorithm based on that cost function is derived on the orthogonal group. This algorithm is applied on simulated gene expression data in Section 4. The paper ends with conclusions in Section 5.

2. THE RADICAL CONTRAST

Like many measures of statistical independence, the RADICAL contrast [5] is derived from the mutual information, i.e., the Kullback-Leibler divergence between the joint distribution and the product of the marginal distributions,

$$J(Z) = \int p(z_1, \ldots, z_n) \log \frac{p(z_1, \ldots, z_n)}{p(z_1) \ldots p(z_n)} dz_1 \ldots dz_n,$$

where $Z = (z_1, \ldots, z_n)^T$ are the estimated sources. The mutual information can be expressed in terms of differential entropies as follows,

$$J(Z) = \sum_{i=1}^{n} H(z_i) - H(z_1, \ldots, z_n).$$

After introduction of the demixing model $Z = W^T X$, a function defined over the space of the demixing matrices is obtained,

$$J(W) = \sum_{i=1}^{n} H(e_i^T W^T X) - \log(|W|) - H(x_1, \ldots, x_n), \quad (2)$$

where e_i is the ith basis vector. The difficulty of function (2) lies in the evaluation of the differential entropies for one-dimensional variables. An efficient estimator of these quantities was derived by considering order statistics [5]. Given a one-dimensional variable z defined by its samples, the order statistics of z is the set of samples $\{z^1, \ldots, z^N\}$ rearranged in non-decreasing order, i.e., $z^1 \leq \ldots \leq z^N$. The differential entropy of a one-dimensional variable z defined by its order statistics $\{z^1, \ldots, z^N\}$ can be estimated by

$$\hat{H}(z) = \frac{1}{N-m} \sum_{j=1}^{N-m} \log \left(\frac{N+1}{m} (z^{(j+m)} - z^{(j)}) \right), \quad (3)$$

where m is typically set to $\sqrt{N}$. The RADICAL contrast is actually the function (2) where the differential entropies are evaluated with the estimator (3),

$$J_{\mathrm{RADICAL}}(W) = \sum_{i=1}^{n} \hat{H}^{(i)}(W) - \log(|W|), \quad (4)$$

with $\hat{H}^{(i)}(W) = \hat{H}(e_i^T W^T X)$.

Figure 1 gives a rough idea of the shape of the RADICAL contrast by representing its evolution along geodesic curves on the orthogonal group,

$$\mathcal{O}_n = \{W \in \mathbb{R}^{n \times n} : W^T W = I_n\}, \quad (5)$$

for several benchmark problems. Each of them considers a data set with an important number of observations (n = 15) but with rather few samples available (N=100). These observations result from a mixture of 15 independent sources and can be subjected to noise and outliers. Table 1 describes each problem in more detail. The plots on the left part of Figure 1 illustrate directly the contrast (4), while an empirical smoothing process was used for the plots on the right part. This smoothing process simply expands the dataset with noisy replicates of the original data [5]. The origin of each plot corresponds to the solution of the ICA problem.

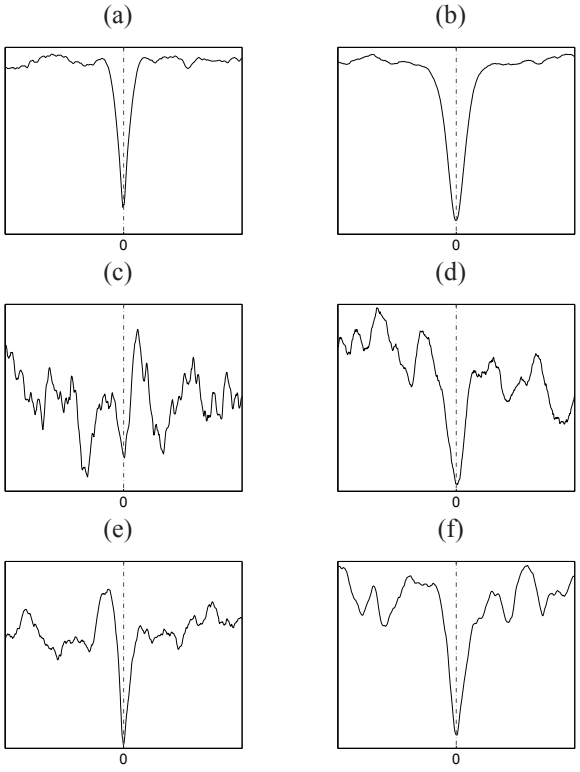

Fig. 1. Evolution of the RADICAL contrast along geodesics of the orthogonal group for six benchmark problems.

	Noise	Outliers	Smoothing
a	×	×	×
b	×	×	✓
c	✓	×	×
d	✓	×	✓
e	×	✓	×
f	×	✓	✓

Table 1. Configuration of the benchmark problems considered on Figure 1.

Figure 1 highlights some important features of the RADICAL contrast. First of all, the global minimum stays nearly unaffected by the presence of noise and outliers as well as by the few number of samples available. This illustrates the high robustness of the RADICAL contrast. Next, the smoothing method appears to be very efficient and useful for noisy datasets. Finally, the RADICAL contrast seems to be a very hilly function and is likely to present many local minima, which complicates the optimization process.

Figure 2 illustrates the favorable robustness of the RADICAL contrast with respect to outliers by applying several ICA algorithms on a simple benchmark problem. The abscissa indicates the proportion of entries in the dataset that are corrupted. α measures the quality of the identification of the independent sources. A value close to zero stands for a good performance. ICA algorithms based on the RADICAL contrast, i.e., the gradient-descent algorithm introduced in Section 3 as well as the original implementation of RADICAL [5], tolerate up to 1.5% of corrupted values in the dataset, while classical algorithms such as FastICA [6] and JADE [7] collapse as soon as there are outliers.

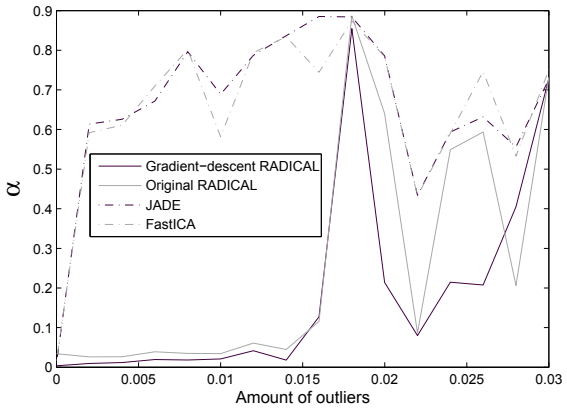

Fig. 2. Evolution of the performance of some ICA algorithms with the amount of outliers (n=4, N=200).

3. A GRADIENT-DESCENT ALGORITHM

Each ICA algorithm consists in the minimization of a particular contrast by some optimization algorithm. Since the contrast is a function defined over the space of the demixing matrices, the search space of the optimization problem is a matrix manifold. Because of the inherent scale symmetry of ICA, some constraints on the entries of the demixing matrix have to be added to get an efficient optimization. In particular, the columns of the demixing matrix should have a unit-norm. Most often, the data X is prewhitened and the demixing matrix is restricted to be orthogonal. Hence, ICA is a direct application of the theory of optimization over nonlinear matrix manifolds [8].

In its original implementation, the RADICAL contrast is optimized by means of Jacobi rotations [5]. Only one parameter is varying at each iteration and global minimization over that parameter is accomplished by exhaustive search.

This section is dedicated to the derivation of a gradient-descent algorithm over the orthogonal group (5). Generalization to non-orthogonal manifolds, e.g., the oblique manifold [9], will be the topic of future research. It should be first noted that, because of the rearranging process required by the order statistics, the RADICAL contrast function is only piecewise differentiable. Nevertheless, the evaluation of an analytical expression to the gradient of the estimator (3) is rather straightforward. All derivatives are performed in the embedding Euclidian space $\mathbb{R}^{n\times n}$, while the gradient is obtained after projection onto the tangent space to the orthogonal group,

$$\mathrm{grad}\hat{H}^{(i)}(W) = P_{T_W}\mathrm{grad}\tilde{H}^{(i)}(W),$$

where $\tilde{H}^{(i)}$ is the extension of $\hat{H}^{(i)}$ over $\mathbb{R}^{n\times n}$, i.e., $\tilde{H}^{(i)} = \hat{H}^{(i)}|_{\mathcal{O}_n}$, and $P_{T_W}(Z)$ is the projection operator, namely, in case of the orthogonal group, $P_{T_W}(Z) = \frac{1}{2}W(W^TZ - Z^TW)$. The evaluation of the gradient in the embedding manifold is performed by means of the identity,

$$\mathrm{D}\tilde{H}^{(i)}(W)[Z] = \langle \mathrm{grad}\tilde{H}^{(i)}(W), Z\rangle,$$

with the metric $\langle Z_1, Z_2\rangle = \mathrm{tr}(Z_1^TZ_2)$. The directional derivative is given by

$$\mathrm{D}\tilde{H}^{(i)}(W)[Z] = \mathrm{tr}\left(\frac{1}{N-m}\sum_{j=1}^{N-m}\frac{e_i(x^{(k_{j+m})} - x^{(k_j)})^T}{e_i^TW^T(x^{(k_{j+m})} - x^{(k_j)})}Z\right),$$

where $x^{(k)}$ denotes the kth column of the data matrix X. The indices k_{j+m} and k_j point to the samples of the estimated source z_i, which are respectively at positions $j+m$ and j in the order statistics of z_i.[1]

[1]More details about the calculations can be found in a forthcoming extended version of the present paper.

We propose a gradient-descent algorithm based on an exact line-search method, i.e., the algorithm is searching at each iteration for the minimum in the direction opposed to the gradient. This one-dimensional optimization is performed by means of a golden section search [10]. The gradient-descent algorithm inherits of all the local convergence properties of line-search optimization methods [8], but it is unable to perform the global optimization of the RADICAL contrast. Nevertheless, it can be extended to global optimization by adding a stochastic component to the gradient. This is the topic of ongoing research.

4. SIMULATION RESULTS

The following simulations are based on a benchmark setup that simulates the analysis of gene expression data. Expression modes are generated artificially by building a vector of 7114 genes with entries around one for the genes that are part of the expression mode and entries close to zero otherwise. An artificial gene expression database of 7114 genes and 200 experiments is then obtained by multiplying the matrix of the expression modes A with the matrix of the independent sources S, according to equation (1). Some ICA algorithms are applied on this dataset and the quality of the identification of the expression modes is evaluated afterwards by a measure α, which stands for a good performance once it is close to zero. All algorithms are identically initialized with a matrix that is close to the solution of the ICA problem. Hence, Figure 3 analyzes the local behavior of the ICA algorithms for several problem setups.

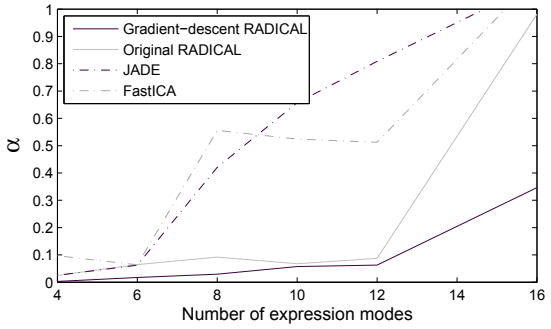

Fig. 3. Performance of the ICA algorithms on a gene expression benchmark involving an increasing number of expression modes.

This figure indicates that algorithms based on the RADICAL contrast seem well-adapted to the analysis of gene expression data. Furthermore, the new gradient-descent algorithm appears to be more accurate than the original implementation of RADICAL. Its global performance is not addressed in the present paper.

5. CONCLUSION

ICA is expected to become a method of choice in many application areas, and in particular for the analysis of gene expression data. Unfortunately, current ICA algorithms are usually not well adapted for experimental datasets. Improvement in term of robustness to lack of samples, noise and outliers is an important issue for future ICA approaches. This paper sets a first step in that direction.

6. REFERENCES

[1] A. Riva, A.-S. Carpentier, B. Torrésani, and A. Hénaut, "Comments on selected fundamental aspects of microarray analysis," *Computational Biology and Chemistry*, vol. 29, no. 5, pp. 319–336, 2005.

[2] W. Liebermeister, "Linear modes of gene expression determined by independent component analysis," *Bioinformatics*, vol. 18, pp. 51–60, 2002.

[3] A.-M. Martoglio, J. W. Miskin, S. K. Smith, and D. J. C. MacKay, "A decomposition model to track gene expression signatures: preview on observer-independent classification of ovarian cancer," *Bioinformatics*, vol. 18, no. 12, pp. 1617–1624, 2002.

[4] S. A. Saidi, C. M. Holland, D. P. Kreil, D. J. C. MacKay, D. S. Charnock-Jones, C. G. Print, and S. K. Smith, "Independent component analysis of microarray data in the study of endometrial cancer," *Oncogene*, vol. 23, no. 39, pp. 6677–6683, 2003.

[5] E.G. Learned-Miller and J. W. Fisher III, "ICA using spacings estimates of entropy," *Journal of Machine Learning Research*, vol. 4, pp. 1271–1295, 2003.

[6] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, John Wiley & Sons, 2001.

[7] J.-F. Cardoso, "High-order contrasts for independent component analysis," *Neural Computation*, vol. 11, no. 1, pp. 157–192, 1999.

[8] P.-A. Absil, R. Mahony, and R. Sepulchre, *Optimization Algorithms on Matrix Manifolds*, Princeton University Press, To appear.

[9] P.-A. Absil and K.A. Gallivan, "Joint diagonalization on the oblique manifold for independent component analysis," in *Proceedings of ICASSP2006*, 2006.

[10] W.H. Press, S.A.Teukolsky, W.T. Vetterling, and B.P. Flannery, *Numerical Recipes in C, The Art of Scientific Computing, 2nd edition*, Cambridge University Press, 1999.