Optimization on the Orthogonal Group for Independent Component Analysis

Michel Journée[1], Pierre-Antoine Absil[2], and Rodolphe Sepulchre[1]

[1] Dept. of Electrical Engineering and Computer Science, University of Liège, Belgium
[2] Dept. of Mathematical Engineering, Université catholique de Louvain, Belgium

Abstract. This paper derives a new algorithm that performs independent component analysis (ICA) by optimizing the contrast function of the RADICAL algorithm. The core idea of the proposed optimization method is to combine the global search of a good initial condition with a gradient-descent algorithm. This new ICA algorithm performs faster than the RADICAL algorithm (based on Jacobi rotations) while still preserving, and even enhancing, the strong robustness properties that result from its contrast.

Keywords: Independent Component Analysis, RADICAL algorithm, optimization on matrix manifolds, line-search on the orthogonal group.

1 Introduction

Independent Component Analysis (ICA) was originally developed for the blind source separation problem. It aims at recovering independent source signals from linear mixtures of these. As in the seminal paper of Comon [1], a linear instantaneous mixture model will be considered in this paper,

$$X = AS, \tag{1}$$

where X, A and S are matrices in $\mathbb{R}^{n\times N}$, $\mathbb{R}^{n\times p}$ and $\mathbb{R}^{p\times N}$ respectively, with p less or equal to n. The rows of S are assumed to be samples of independent random variables. Thus, ICA provides a linear representation of the data X in terms of components S that are statistically independent.

ICA algorithms are based on the inverse of the mixing model (1),

$$Z = W^T X,$$

where Z and W are matrices in $\mathbb{R}^{p\times N}$ and $\mathbb{R}^{n\times p}$, respectively. The aim of ICA algorithms is to optimize over W the statistical independence of the p random variables, whose samples are given in the p rows of Z. The statistical independence is measured by a cost function

$$\gamma : \mathbb{R}^{n\times p} \to \mathbb{R} : W \mapsto \gamma(W),$$

termed the *contrast function*.

M.E. Davies et al. (Eds.): ICA 2007, LNCS 4666, pp. 57–64, 2007.

In the remainder of this paper, we assume that the data matrix X has been preprocessed by means of prewhitening and its dimensions have been reduced by retaining the dominant p-dimensional subspace. Consequently, the contrast function γ is defined on a set of *square* matrices, i.e,

$$\gamma : \mathbb{R}^{p \times p} \to \mathbb{R} : W \mapsto \gamma(W).$$

Several contrast functions for ICA can be found in the literature. In this paper, we consider the RADICAL contrast function proposed in [2]. Advantages of this contrast are a strong robustness to outliers as well as to the lack of samples.

A good contrast for γ is not enough to make an efficient ICA algorithm. The other ingredient is a suitable numerical method to compute an optimizer of γ. This is the topic of the present paper. The authors of [2] optimize their contrast by means of Jacobi rotations combined with an exhaustive search. This yields the complete *Robust Accurate Direct ICA aLgorithm* (RADICAL). We propose a new steepest-descent-based optimization method that reduces the computational load of RADICAL.

The paper is organized as follows. The contrast function of RADICAL is detailed in Section 2. Section 3 describes a gradient-descent optimization algorithm. In Section 4, this local optimization is integrated within a global optimization framework. The performance of this new ICA algorithm is briefly illustrated in Section 5.

2 A Robust Contrast Function

Like many other measures of statistical independence, the contrast of RADICAL [2] is derived from the mutual information [3]. The mutual information $I(Z)$ of a multivariate random variable $Z = (z_1 \ldots, z_p)$ is defined as the Kullback-Leibler divergence between the joint distribution and the product of the marginal distributions,

$$I(Z) = \int p(z_1, \ldots, z_p) \log \frac{p(z_1, \ldots, z_p)}{p(z_1) \ldots p(z_p)} dz_1 \ldots dz_p. \tag{2}$$

This quantity presents all the required properties for a contrast function: it is nonnegative and equals zero if and only if the variables Z are statistically independent. Hence, its global minimum corresponds to the solution of the ICA problem.

The challenge is to get a good estimator of $I(Z)$. A possible approach is to express the mutual information in terms of the differential entropy of a univariate random variable z,

$$S(z) = \int p(z) \log(p(z)) dz, \tag{3}$$

for which efficient statistical estimators are available.

According to definitions (2) and (3), the following holds,

$$I(Z) = \sum_{i=1}^{p} S(z_i) - S(z_1, \ldots, z_p). \tag{4}$$

The introduction of the demixing model $Z = W^T X$ within (4) results in

$$\gamma(W) = \sum_{i=1}^{p} S^{(i)}(W) - \log(|W|) - S(x_1, \ldots, x_p), \tag{5}$$

where $S^{(i)}(W) = S(e_i^T W^T X)$ and e_i is the ith basis vector. The last term of (5) is constant and its evaluation can be skipped by the ICA algorithm. An estimator for the differential entropy of univariate variables was derived in [2] by considering order statistics. Given a univariate random 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 z can be estimated by the simple formula

$$S(z) = \frac{1}{N-m} \sum_{j=1}^{N-m} \log\left(\frac{N+1}{m}(z^{(j+m)} - z^{(j)})\right), \tag{6}$$

where m is typically set to $\sqrt{N}$. Function (5) with the differential entropies being estimated by (6) is the contrast of the RADICAL algorithm [2].

This contrast presents several assets in terms of robustness. Its robustness to outliers was underlined in the original paper [2]. Robustness to outliers means that the presence of some corrupted entries in the observations data set X has little influence on the position of the global minimizer of that contrast. This is a key feature in many applications, especially for the analysis of gene expression data [4], where each entry in the observation matrix results from individual experiments that are likely to sometimes fail. The RADICAL contrast brings also advances in terms of robustness to the lack of samples. This will be illustrated in Section 5.

3 A Line-Search Optimization Algorithm

In accordance with the fact that the independence between random variables is not altered by scaling, the contrast function (5) presents the scale invariance property

$$\gamma(W) = \gamma(W\Lambda),$$

for all invertible diagonal matrices Λ. Optimizing a function with such an invariance property is a degenerate problem, which entails difficulties of theoretical (convergence analysis) and practical nature unless some constraints are introduced. In the case of prewhitening-based ICA, it is common practice to restrict the matrix W to be orthonormal [1], i.e., $W^T W = I$. Classical constrained optimization methods could be used. We favor the alternative to incorporate the

constraints directly into the search space and to perform unconstrained optimization over the orthogonal group, i.e.,

$$\min_{W \in \mathcal{O}_p} \gamma(W) \quad \text{with } \mathcal{O}_p = \{W \in \mathbb{R}^{p \times p} | W^T W = I\}. \tag{7}$$

Most classical unconstrained optimization methods — such as gradient-descent, Newton, trust-region and conjugate gradient methods — have been generalized to the optimization over matrix manifolds (see [5] and references therein).

The remainder of this section deals with the derivation of a line-search optimization method on the orthogonal group for the RADICAL contrast function (5). Line-search on a nonlinear manifold is based on the update formula

$$W_+ = R_W(t\eta), \tag{8}$$

which consists in moving from the current iterate $W \in \mathcal{O}_p$ in the search direction η with a certain step size t to identify the next iterate $W_+ \in \mathcal{O}_p$. t is a scalar and η belongs to $T_W\mathcal{O}_p = \{W\Omega | \Omega \in \mathbb{R}^{p \times p}, \Omega^T = -\Omega\}$, the tangent space to $\mathcal{O}_p$ at W. The retraction R_W is a mapping from the tangent space to the manifold. More details about this notion can be found in [5]. Our algorithm selects the Armijo point t^A as step size and the opposite of the gradient of the cost function γ at the current iterate as search direction.

The Armijo step size is defined by $t^A = \beta^m \alpha$, with the scalars $\alpha > 0$, $\beta \in (0, 1)$ and m being the first nonnegative integer such that

$$\gamma(W) - \gamma(R_W(\beta^m \alpha)) \geq -\sigma \langle \text{grad}\gamma(W), \beta^m \alpha \eta \rangle_W,$$

where W is the current iterate on $\mathcal{O}_p$ and $\sigma \in (0, 1)$. This step size ensures a sufficient decrease of the cost function at each iteration. The resulting line-search algorithm converges to the set of points where the gradient of γ vanishes [5].

An analytical expression of the gradient of the RADICAL contrast (5) has been derived in [6]. Let us just sketch the main points of this computation. First, because of the orthonormality condition, the second term of (5) vanishes. Furthermore, since the last term is constant, we have

$$\text{grad}\gamma(W) = \sum_{i=1}^{p} \text{grad}S^{(i)}(W).$$

The gradient of $S^{(i)}$ is given by

$$\text{grad}S^{(i)}(W) = P_{T_W}\left(\text{grad}\tilde{S}^{(i)}(W)\right),$$

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

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

with the metric $\langle Z_1, Z_2 \rangle = \mathrm{tr}(Z_1^T Z_2)$ and where

$$\mathrm{D}\tilde{S}^{(i)}(W)[Z] = \lim_{t \to 0} \frac{\tilde{S}^{(i)}(W + tZ) - \tilde{S}^{(i)}(W)}{t}$$

is the standard directional derivative of $\tilde{S}^{(i)}$ at W in the direction Z. Since one wants to compute the gradient on the orthogonal group, the direction Z can be restricted to the tangent plane at the current iterate, i.e., $Z \in T_W \mathcal{O}_p$.

As we have shown in [6], the gradient of the differential entropy estimator on the orthogonal group $\mathcal{O}_p$ is finally given by

$$\mathrm{grad} S^{(i)}(W) = P_{T_W} \left(\frac{1}{N-m} \sum_{j=1}^{N-m} \frac{(x^{(k_{j+m})} - x^{(k_j)}) e_i^T}{e_i^T W (x^{(k_{j+m})} - x^{(k_j)})} \right),$$

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. The computational cost for the gradient is of the same order as for the contrast, namely $\mathcal{O}(pN \log N)$.

More details about the Armijo point, the computation of gradients and, more generally, about line-search algorithms on manifolds can be found in [5].

4 Towards a Global Optimization Scheme

The algorithm described in the previous section inherits all the local convergence properties of line-search optimization methods [5]. Nevertheless, the contrast of RADICAL presents many spurious local minima that do not properly separate the observations X into independent sources. The line-search algorithm may thus fail in the context of ICA. Nevertheless, it leads to an efficient ICA algorithm when it is initialized within the basin of attraction of the global minimizer W_*. It is therefore essential to find good initial candidates for the line-search algorithm. The procedure proposed in this paper rests on empirical observations about the shape of the contrast function $\gamma(W)$. Figure 1 represents the evolution of this function as well as of the norm of its gradient along geodesic curves on the orthogonal group $\mathcal{O}_p$ for a particular benchmark setup (p=6, N=1000).

Figure 1 and extensive simulations not included in the present paper incite us to view the contrast function of RADICAL as possessing a very deep global minimum surrounded by many small local minima. Furthermore, the norm of the gradient tends to be much larger within the basin of attraction of the global minimizer. The norm of the gradient thus provides a criterion to discriminate between points that are inside this basin of attraction and those that are outside.

Our algorithm precedes the gradient optimization with the global search of a point where the gradient has a large magnitude. The search is performed along particular geodesics of the orthogonal group, exploiting the low numerical cost of Jacobi rotations. All geodesics on the orthogonal group $\mathcal{O}_p$ have the form $\Gamma(t) = We^{tB}$, where $W \in \mathcal{O}_p$ and B is a skew-symmetric matrix of the same

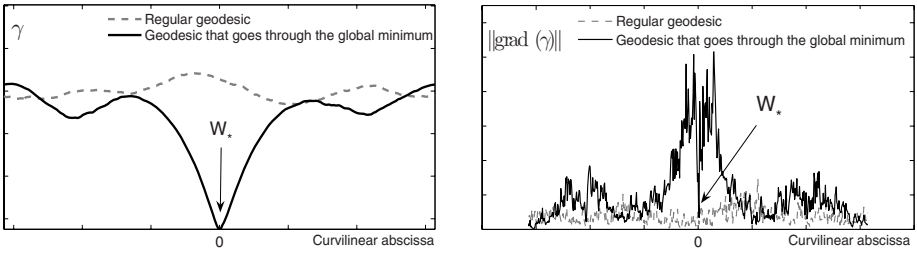

Fig. 1. Evolution of the contrast and the norm of its gradient along geodesics of $\mathcal{O}_p$

size as W. Jacobi rotations correspond to B having only zero elements except one element in the upper triangle and its symmetric counterpart, i.e., $B(i,j) = 1$ and $B(j,i) = -1$ with $i < j$. The contrast function γ evaluated along such geodesics has a periodicity of $\frac{\pi}{2}$, i.e.,

$$\gamma(We^{tB}) = \gamma(We^{(t+\frac{\pi}{2})B})$$

Such a geodesic is in fact a Jacobi rotation on the two-dimensional subspace spanned by the directions i and j. This periodicity is an interesting feature for an exhaustive search over the curvilinear abscissa t since it allows to define upper and lower bounds for t.

Our algorithm evaluates the gradient at a fixed number of points that are uniformly distributed on randomly selected geodesics of periodicity $\frac{\pi}{2}$. This process is pursued until a point with sufficient steepness is found. The steepness is simply evaluated by the Frobenius norm of the gradient of γ. Such a point is expected to belong to the basin of attraction of the global minimum and serves as initialization for the line-search algorithm of the previous section.

5 Some Benchmark Simulations

This section evaluates the performance of the new algorithm against the performance of the RADICAL algorithm. All results are obtained on benchmark setups that artificially generate observations X by linear transformation of known statistically independent sources S.

Figure 2 illustrates that the new algorithm reaches the global minimum of the contrast with less than half the computational effort required by the RADICAL algorithm. These results are based on a benchmark with $N = 1000$ samples while the dimension p of the problem varies from 2 to 8. For each p, five different data matrices X are obtained by randomly mixing p sources chosen as sinusoids of random frequencies and random phases. The indicated computational time is an average over these five ICA runs.

Figure 3 highlights the robustness properties of the contrast discussed in Section 2. The left graph results from a benchmark with $p = 6$ sources and

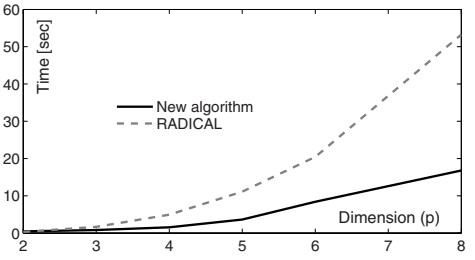

Fig. 2. Reduced computational time of the new ICA algorithm

$N = 1000$ samples. A given percent of the entries of the data set have been artificially corrupted to simulate outliers. The right graph considers a benchmark with $p = 6$ sources, no outliers and a varying number of samples. The quality of the ICA separation is measured by an index α[1], which stands for a good performance once it is close to zero. The left graph indicates that both the new algorithm and the RADICAL algorithm are robust to these outliers while classical ICA algorithms such as JADE [7] or FastICA [8] collapse immediately. It should be noted that the new algorithm supports up to 3% of outliers on the present benchmark and is thus more robust than RADICAL. Similarly, the right graph of Figure 3 suggests that the new algorithm is more robust to the lack of samples than RADICAL.

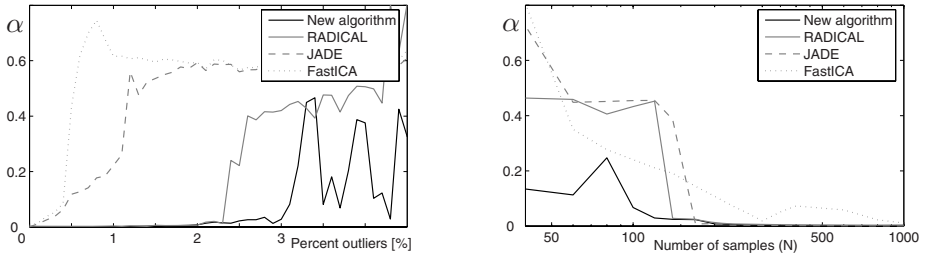

Fig. 3. Robustness properties of the new ICA algorithm

6 Conclusions

The RADICAL algorithm [2] presents very desirable robustness properties: robustness to outliers and robustness to the lack of samples. These are essential

[1] Given the demixing matrix W^* and the matrix W identified by the ICA algorithm,

$$\alpha(W, W^*) = \min_{\Lambda, P} \frac{\|W\Lambda P - W^*\|_F}{\|W^*\|_F},$$

where Λ is a non-singular diagonal matrix and P a permutation matrix.

for some applications, in particular for the analysis of biological data that are usually of poor quality because of the few number of samples available and the presence of corrupted entries resulting from failed experiments [4]. The RADICAL algorithm inherits these robustness properties from its contrast function. In this paper, we have shown that the computation of the demixing matrix by optimization of the RADICAL contrast function can be performed in a more efficient manner than with the Jacobi rotation approach considered in [2]. Our new optimization process works in two stages. It first identifies a point that supposedly belongs to the basin of attraction of the global minimum and performs afterwards the local optimization of the contrast by gradient-descent from this point. This new ICA algorithm requires less computational effort and seems to enhance the robustness margins.

Acknowledgments

M. Journée is a research fellow of the Belgian National Fund for Scientific Research (FNRS). This paper presents research results of the Belgian Network DYSCO (Dynamical Systems, Control, and Optimization), funded by the Interuniversity Attraction Poles Programme, initiated by the Belgian State, Science Policy Office. The scientific responsibility rests with its authors.

References

1. Comon, P.: Independent Component Analysis, a new concept. In: Signal Processing, vol. 36(3), pp. 287–314. Elsevier, Amsterdam (1994) (Special issue on Higher-Order Statistics)
2. Learned-Miller, E.G., Fisher III, J.W.: ICA using spacings estimates of entropy. Journal of Machine Learning Research 4, 1271–1295 (2003)
3. Cover, T.M., Thomas, J.A.: Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing). Wiley-Interscience, Chichester (2006)
4. Liebermeister, W.: Linear modes of gene expression determined by independent component analysis. Bioinformatics 18, 51–60 (2002)
5. Absil, P.-A., Mahony, R., Sepulchre, R.: Optimization Algorithms on Matrix Manifolds. Princeton University Press (to appear)
6. Journée, M., Teschendorff, A.E., Absil, P.-A., Sepulchre, R.: Geometric optim. methods for ICA applied on gene expression data. In: Proc. of ICASSP (2007)
7. Cardoso, J.-F.: High-order contrasts for independent component analysis. Neural Computation 11(1), 157–192 (1999)
8. Hyvärinen, A., Karhunen, J., Oja, E.: Independent Component Analysis. John Wiley & Sons, Chichester (2001)