JOINT DIAGONALIZATION ON THE OBLIQUE MANIFOLD FOR INDEPENDENT COMPONENT ANALYSIS

P.-A. Absil[*]

Department of Mathematical Engineering
Université catholique de Louvain, Belgium
and Peterhouse, University of Cambridge, UK

K. A. Gallivan[†]

School of Computational Science
Florida State University
Tallahassee, FL 32306-4120, USA

ABSTRACT

Several blind source separation algorithms obtain a separating matrix by computing the congruence transformation that "best" diagonalizes a collection of covariance matrices. Recent methods avoid a pre-whitening phase and directly attempt to compute a non-orthogonal diagonalizing congruence. However, since the magnitude of the sources is unknown, there is a fundamental indeterminacy on the norm of the rows of the separating matrix. We show how this indeterminacy can be taken into account by restricting the separating matrix to the oblique manifold. The geometry of this manifold is developed and a trust-region-based algorithm for non-orthogonal joint diagonalization is proposed.

1. INTRODUCTION

Assume that n measured signals $x(t) = [x_1(t), \ldots, x_n(t)]^T$ are instantaneous linear mixtures of p underlying, statistically independent source signals $s(t) = [s_1(t), \ldots, s_p(t)]^T$; this can be compactly written as

$$x(t) = As(t),$$

where the matrix A is an unknown constant *mixing matrix* containing the mixture coefficients. We assume throughout that all vectors and matrices are real, and we let the superscript T denote the matrix transpose. The problem of independent component analysis (ICA) or blind source separation (BSS) is to identify the mixing matrix A or recover the source signals $s(t)$, using only the observed signals $x(t)$. This problem is usually translated to finding a *separating matrix* W such that the signals $y(t)$ given by

$$y(t) = W^T x(t)$$

are estimates of the signals $s(t)$. It is known that this problem has two basic indeterminacies: without any further information, it is impossible to recover the scaling and the order of the source signals. For non-Gaussian sources, these are the only indeterminacies; this also holds under mild conditions for sources that are not temporally white [1]. We will say that W is a *true* separating matrix if $W^T A$ can be expressed as the product of a permutation matrix (order indeterminacy) and a diagonal matrix (scaling indeterminacy).

In several ICA algorithms, the observed sources $x(t)$ are used to construct a set of "target matrices" $C_1, \ldots, C_N$ with the following property: all the matrices $W^T C_i W, i = 1, \ldots, N$ are diagonal if and only if W is a true separating matrix. In practice, due to the presence of noise and to the limited amount of samples of $x(t)$ available, the target set $\{C_1, \ldots, C_N\}$ does not admit exact joint diagonalization (JD), and one must resort to *approximate joint diagonalization*, that is, find the matrix W that "best diagonalizes" the target set. The various JD-based ICA algorithms differ in the choice of the target matrices and in the cost function used to define the "best diagonalization". Several possibilities for the choice of the target matrices are mentioned in [2], and [3] lists a few possible cost functions. A frequently encountered cost function is

$$f(W) = \sum_i \|\mathrm{off}(W^T C_i W)\|_F^2; \qquad (1)$$

here $\|M\|_F^2$ denotes the square Frobenius norm of M (that is, the sum of the squares of the elements of M) and $\mathrm{off}(M) := M - \mathrm{ddiag}(M)$, where $\mathrm{ddiag}(M)$ denotes the diagonal matrix whose diagonal elements are those of M. Following the notation in [4], we let $\mathrm{diag}(M)$ denote the vector of diagonal elements of M.

Most joint diagonalization algorithms, such as the SOBI algorithm [1], start with a *pre-whitening* step. First, in order to remove the scaling indeterminacy, it is assumed that $E[s(t)s^T(t)] = I$. (This is without loss of generality, since the scaling factors can be absorbed in the columns of A.) A *whitening matrix* $\tilde{W}$ is then sought such that one of the target matrices (say C_1, usually an estimation of the covariance matrix $E[x(t)x^T(t)]$) is reduced to the identity matrix; that is, $\tilde{W}^T C_1 \tilde{W} = I_p$. It follows that there exists an orthogonal matrix U such that $U^T \tilde{W}^T A = I$. In a second step, an orthogonal matrix U is sought that diagonalizes the new tar-

[*]This work was supported by Microsoft Research through a Research Fellowship at Peterhouse, Cambridge.

[†]This work was supported by the US NSF under Grant ACI0324944.

get set $\{\tilde{W}^T C_1 \tilde{W}, \ldots, \tilde{W}^T C_1 \tilde{W}\}$. This yields a separating matrix $W = \tilde{W}U$; see, for example, [1] for details.

Since U is constrained to be orthogonal, it is a solution of an optimization problem on a manifold—the orthogonal group, or more generally the compact Stiefel manifold of matrices with orthonormal columns. This calls for the use of differential-geometric optimization techniques. There has been interest for optimization on manifolds at least since the work of Luenberger and Gabay in the 1970s and 1980s; these and several other references are mentioned in [5]. Applications to the orthogonal joint diagonalization problem have been proposed by Rahbar and Reilly [6], Douglas [7], Joho and Mathis [8], Joho and Rahbar [4], Nikpour *et al.* [9], Nishimori and Akaho [10].

However, the pre-whitening step that yields the orthogonality constraint has the drawback that it singles out one of the target matrix C_1 of which it would attain exact diagonalization at the possible cost of poor diagonalization of the other target matrices [2]. Moreover, inaccuracies in the computation of $\tilde{W}$ cannot be compensated in the sequel. Therefore, a few algorithms have been proposed that directly compute a nonorthogonal separating matrix W without resorting to a pre-whitening process; see for example [11, 12, 2, 13].

Amari *et al.* [11] and Afsari and Krishnaprasad [13] use differential-geometric concepts for nonorthogonal JD. As was pointed out in [13], when the off-diagonal cost function (1) is allowed to take its argument W in the whole set $\mathbb{R}^{n\times p}$ of $n \times p$ matrices, it admits a global minimizer at the zero matrix. More generally, the cost function (1) is not scale invariant; that is, $f(WD)$ is in general different from $f(W)$ when D is a nonsingular diagonal matrix. This issue can be tackled by imposing constraints on the power of the separated signals $y(t)$. However, as argued by Amari *et al.* [11], this is impractical in frequently encountered applications where the amplitude of components may change suddenly; therefore, constraints should be placed instead on the separating matrix W. Amari *et al.* [11] (see also Afsari and Krishnaprasad [13]) propose to constrain the allowed variations of W to belong to a subspace *orthogonal* to the equivalence class $\{WD : D \text{ diagonal}\}$. The notion of orthogonality used in [11] has been shown to be a nonholonomic (or nonintegrable) constraint.

Finally, we point out that most algorithms for ICA do not achieve superlinear convergence, as they are based on steepest-descent or direct-search ideas. Exceptions are the conjugate gradient on the Stiefel manifold used by Rahbar and Reilly [6] and the Newton method on Stiefel of Joho and Rahbar [4] and Nikpour *et al.* [9]. Superlinear convergence is useful when a high precision is sought; this applies to situations where the noise level is low and the source signals (or the mixing matrix) want to be recovered accurately.

In this paper, we propose a superlinearly convergent algorithm for nonorthogonal joint diagonalization, based on a recently proposed trust-region method on Riemannian manifolds [14, 5]; we dub the algorithm RTR-ICA. In comparison with the Newton method, the trust-region approach offers better global convergence properties and similar local convergence properties at a lower computational cost [5]. Our approach also departs from previous work in the way constraints are imposed on the separating matrix W: we require that W be an *oblique rotation* [15], that is, all the columns of W have unit Euclidean norm. Instead of being nonholonomic, this constraint defines a submanifold of $\mathbb{R}^{n\times p}$ called the *oblique manifold*

$$\mathcal{OB}(n,p) = \{Y \in \mathbb{R}^{n\times p} : \mathrm{ddiag}(Y^T Y) = I_p\}. \quad (2)$$

Moreover, in contrast to the Stiefel manifold approach, a pre-whitening step is not required.

The rest of the paper is organized as follows. The geometry of the oblique manifold is described in Section 2. Formulas for the gradient and Hessian of the off-diagonal cost function (1) are obtained in Section 3. The workings of the RTR algorithm are briefly explained in Section 4 (we refer to [5] for details). Numerical experiments are presented in Section 5.

2. GEOMETRY OF THE OBLIQUE MANIFOLD

We refer the reader to [5] and references therein for the relevant notions of Riemannian geometry. The manifold $\mathcal{OB}(n,p)$ is the set of all $n \times p$ matrices with normalized columns. It is an embedded submanifold of $\mathbb{R}^{n\times p}$. We consider the canonical inner product

$$\langle Z_1, Z_2 \rangle := \mathrm{trace}(Z_1^T Z_2) \quad (3)$$

in $\mathbb{R}^{n\times p}$ and view $\mathcal{OB}$ as an embedded Riemannian submanifold of $\mathbb{R}^{n\times p}$. The tangent space (which is defined independently of the metric) is $T_Y\mathcal{OB} = \{Z : \mathrm{ddiag}(Y^T Z) = 0\}$ which means that $y_i^T z_i = 0$, $i = 1, \ldots, p$, where y_i denotes the ith column of Y. The dimension of $\mathcal{OB}$ is $\dim(\mathcal{OB}) = p(n-1)$. The normal space (which depends on the embedding in the Euclidean space $\mathbb{R}^{n\times p}$) is $N_Y\mathcal{OB} = \{YD : D \in \mathbb{R}^{p\times p} \text{ diagonal}\}$. Projections of $Z \in T_Y\mathbb{R}^{n\times p}$ into a normal and tangent component are $P_{N_Y}(Z) = Y\mathrm{ddiag}(Y^T Z)$ and $P_{T_Y}(Z) = Z - Y\mathrm{ddiag}(Y^T Z)$.

Finally, in order to apply the RTR schemes on $\mathcal{OB}$, we must define a *retraction*, which establishes a correspondence between tangent vectors and points on the manifolds. A natural choice that satisfies the required properties [5] is

$$R_Y(Z) = (Y+Z)(\mathrm{ddiag}((Y+Z)^T(Y+Z)))^{-1/2}, \quad (4)$$

which simply consists of adding Z to Y and scaling the columns of the result.

3. THE OFF-DIAGONAL COST FUNCTION ON THE OBLIQUE MANIFOLD

We compute the gradient and the Hessian of $f = \tilde{f}|_{\mathcal{OB}}$ where

$$\begin{aligned}\tilde{f}(Y) &= \sum_{i=1}^{N} \|Y^T C_i Y - \operatorname{ddiag}(Y^T C_i Y)\|_F^2 \\ &= \sum_{i=1}^{N} \operatorname{trace}(\operatorname{off}(Y^T C_i Y) Y^T C_i Y),\end{aligned}$$

C_i symmetric. Notice that f is a function on $\mathcal{OB}$, and its gradient and Hessian are thus defined in the sense of the manifold $\mathcal{OB}$ endowed with its Riemannian metric (3). Note also the identity $\operatorname{trace}(\operatorname{ddiag}(A)B) = \operatorname{trace}(A \operatorname{ddiag}(B))$. For the gradient of $\tilde{f}$, we get (see [16] for details)

$$\operatorname{grad} \tilde{f}(Y) = \sum_{i=1}^{N} 4 C_i Y \operatorname{off}(Y^T C_i Y).$$

We project onto the tangent space to obtain the gradient of f, which yields

$$\begin{aligned}&\operatorname{grad} f(Y) = P_{T_Y} \operatorname{grad} \tilde{f}(Y) \\ &= \sum_{i=1}^{N} 4 C_i Y \operatorname{off}(Y^T C_i Y) - 4 Y \operatorname{ddiag}(Y^T C_i Y \operatorname{off}(Y^T C_i Y)).\end{aligned}$$

Finally,

$$\begin{aligned}\operatorname{Hess} f(Y)[Z] &= P_{T_Y} \operatorname{Dgrad} f(Y)[Z] \\ &= P_{T_Y} \operatorname{Dgrad} \tilde{f}(Y)[Z] - Z \operatorname{ddiag}(Y^T \operatorname{grad} \tilde{f}(Y))\end{aligned}$$

with

$$\begin{aligned}\operatorname{Dgrad} \tilde{f}(Y)[Z] = \sum_{i=1}^{N} &4 C_i Z \operatorname{off}(Y^T C_i Y) + 4 C_i Y \operatorname{off}(Z^T C_i Y) \\ &+ 4 C_i Y \operatorname{off}(Y^T C_i Z).\end{aligned}$$

4. THE RTR-ICA APPROACH

We now have the necessary ingredients to apply the Riemannian trust-region (RTR) approach to the problem of minimizing the cost function (1) on the oblique manifold $\mathcal{OB}$ endowed with the Riemannian metric (3) and the retraction (4). Using the ingredients in the general RTR algorithm given in [5] is rather straightforward and will not be done in detail here. The general idea is as follows. The RTR scheme is an iterative process that, from a current iterate W on the oblique manifold, produces a next iterated W_+ on the oblique manifold. First, a model m_W of the cost function f is constructed around W; more precisely, $m_W(Z)$ approximates $f(R_W(Z))$. When not too expensive computationally (which is the case here), it is useful to choose m_W as the *Newton model*, i.e., second-order Taylor expansion of $f \circ R_W$:

$$m_W(Z) = f(W) + \langle \operatorname{grad} f(W), Z \rangle + \frac{1}{2} \langle \operatorname{Hess} f(W)[Z], Z \rangle.$$

Next, an (approximate) minimizer of the model is sought within a region where the model is "trusted" (hence the name of the method). The size of the trust-region has to be externally specified for the first iterate; it is subsequently automatically updated: if there is a good agreement between the cost function and the model at the proposed next iterate, then the proposed iterate is accepted and the size of the trust region is possibly increased. If the agreement is poor, the size of the trust region can be reduced and the proposed iterate can even be discarded. For more details on the trust-region concept, which originates from the work of Powell in the 1970s, we refer to [5] and references therein. Finally, we point out that there are several ways to approximately solve trust-region subproblems, i.e., to compute an approximate minimizer of a quadratic model within a trust region. Usually, finding a high-precision solution of each subproblem is not necessary and would constitute a waste of computational effort; on the other hand, the approximate solution has to be sufficiently precise so that strong local and global convergence properties hold. To handle this tradeoff, the use of Steihaug's truncated CG method is advocated in [5], and we use it in the numerical experiments reported on in the next section. We also refer to [5] for a convergence analysis of the general RTR schemes.

5. SIMULATION RESULTS

The following simulation follows closely the one in [6]; we simply increased the number of data points and modified some of the signals to improve the spatial uncorrelatedness of the sources. We use four synthetic signals with 10^6 sample points. The sources are mixed using a four-by-four randomly generated mixing matrix A. Three target matrices C_1, C_2, C_3 are chosen as lagged sample covariance matrices of $x(t)$. The iteration is initialized with W equal to the identity matrix. The performance of separation displayed in Figure 1 is measured using the formula

$$P_{\text{index}} = 20 \log_{10} \left(\frac{1}{n} \left(\sum_{i=1}^{n} \left(\sum_{j=1}^{n} \frac{|q_{ij}|}{\max_\ell |q_{i\ell}|} - 1 \right) \right) \right),$$

where q_{ij} is the (i,j)th element of $Q := W^T A$. To illustrate the benefit of taking the model m_W as the Newton model, we applied to the same problem the linearly convergent method obtained by defining m_W as the first-order Taylor expansion; the algorithm then took more than 6000 (inner) iterations to only reach a performance index of -100.

With a view towards a comparison with the simulation results in [6, 4], notice that one iteration of nonlinear CG involves a line-search process with possibly several evaluations

of the cost function, and that an iteration of Newton's method involves solving the Newton equation. In contrast, in the truncated CG process, the major work consists in one application of the Hessian operator per inner iteration.

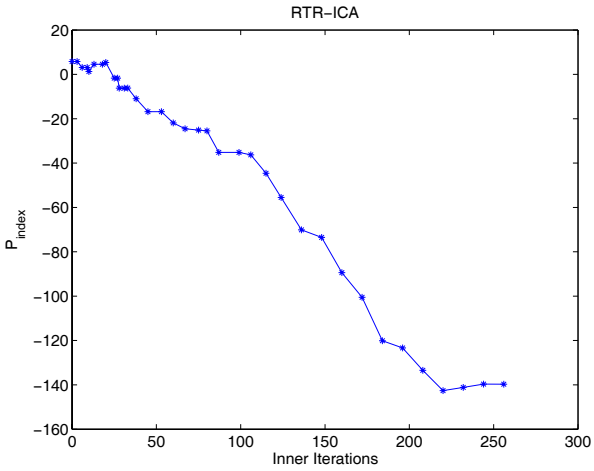

Fig. 1. Convergence of RTR-ICA.

Acknowledgement

Special thanks to Bijan Afsari for helpful discussions about ICA and joint diagonalization.

6. REFERENCES

[1] Adel Belouchrani, Karim Abed-Meraim, Jean-François Cardoso, and Eric Moulines, "A blind source separation technique using secton-order statistics," *IEEE Trans. Signal Process.*, vol. 45, no. 2, pp. 434–444, 1997.

[2] Arie Yeredor, "Non-orthogonal joint diagonalization in the least-squares sense with application in blind source separation," *IEEE Trans. Signal Process.*, vol. 50, no. 7, pp. 1545–1553, 2002.

[3] Wenwu Wang, Saeid Sanei, and Jonathon A. Chambers, "Penalty function-based joint diagonalization approach for convolutive blind separation of nonstationary sources," *IEEE Trans. Signal Process.*, vol. 53, no. 5, pp. 1654–1669, 2005.

[4] Marcel Joho and Kamran Rahbar, "Joint diagonalization of correlation matrices by using Newton methods with applications to blind signal separation," in *Proceedings of IEEE Sensor Array and Multichannel Signal Processing Workshop SAM*, 2002, pp. 403–407.

[5] P.-A. Absil, C. G. Baker, and K. A. Gallivan, "Trust-region methods on Riemannian manifolds," http://www.inma.ucl.ac.be/˜absil/, submitted, 2005.

[6] Kamran Rahbar and James P. Reilly, "Geometric optimization methods for blind source separation of signals," in *International Conference on Independent Component Analysis ICA2000,Helsinki, Finland*, June 2000.

[7] Scott C. Douglas, "Self-stabilized gradient algorithms for blind source separation with orthogonality constraints," *IEEE Trans. Neural Networks*, vol. 11, no. 6, pp. 1490–1497, 2000.

[8] Marcel Joho and Heinz Mathis, "Joint diagonalization of correlation matrices by using gradient methods with application to blind signal separation," in *Proceedings of IEEE Sensor Array and Multichannel Signal Processing Workshop SAM*, 2002, pp. 273–277.

[9] Maziar Nikpour, Jonathan H. Manton, and Gen Hori, "Algorithms on the Stiefel manifold for joint diagonalization," in *Proc. ICASSP*, 2002, pp. II–1481–1484.

[10] Yasunori Nishimori and Shotaro Akaho, "Learning algorithms utilizing quasi-geodesic flows on the stiefel manifold," *Neurocomputing*, vol. 67, pp. 106–135, 2005.

[11] Shun-ichi Amari, Tian-Ping Chen, and Andrzej Cichocki, "Nonholonomic orthogonal learning algorithms for blind source separation," *Neural Computation*, vol. 12, pp. 1463–1484, 2000.

[12] Dinh Tuan Pham, "Joint approximate diagonalization of positive definite Hermitian matrices," *SIAM J. Matrix Anal. Appl.*, vol. 22, no. 4, pp. 1136–1152, 2001.

[13] Bijan Afsari and P. S. Krishnaprasad, "Some gradient based joint diagonalization methods for ICA," in *Proceedings of the 5th International Conference on Independent Component Analysis and Blind Source Separation*, Springer LCNS Series, Ed., 2004.

[14] P.-A. Absil, C. G. Baker, and K. A. Gallivan, "Trust-region methods on Riemannian manifolds with applications in numerical linear algebra," in *Proceedings of the 16th International Symposium on Mathematical Theory of Networks and Systems (MTNS2004), Leuven, Belgium, 5–9 July 2004*, 2004.

[15] N. T. Trendafilov and R. A. Lippert, "The multimode Procrustes problem," *Linear Algebra Appl.*, vol. 349, pp. 245–264, 2002.

[16] P.-A. Absil and K. A. Gallivan, "Joint diagonalization on the oblique manifold for independent component analysis," Tech. Rep. NA2006/01, DAMTP, University of Cambridge, http://www.damtp.cam.ac.uk/user/na/reports.html, 2006.