in proceedings of *ICANN 98* , 2th September 1998, pp. 675-680

TDSEP – an efficient algorithm for blind separation using time structure

Andreas Ziehe and Klaus-Robert Müller
GMD FIRST Rudower Chaussee 5, 12489 Berlin, Germany
e-mail:{ziehe, klaus}@first.gmd.de

Abstract

An algorithm for blind source separation based on *several* time-delayed second-order correlation matrices is proposed. The technique to construct the unmixing matrix employs first a whitening step and then an approximate simultaneous diagonalisation of several time-delayed second-order correlation matrices. Its efficiency and stability are demonstrated for linear artificial mixtures with 17 sources.

1 Introduction

Blind source separation is an increasingly popular data analysis technique. It has been applied successfully to the so called cocktail party problem (e.g. [9, 3, 2, 5, 7, 12, 1]) and to various problems in biomedical data processing (e.g. [10, 13, 14]).
Usually it is assumed that the observed signals $\mathbf{x}$ are constituted of linearly mixed sources $\mathbf{s}$, which are unknown, but mutually statistically independent.

$$x_i(t) = \sum_{j=1}^{n} a_{ij} s_j(t) \qquad 1 \leq i, j \leq n \quad \text{i.e.} \quad \mathbf{x} = \mathbf{A}\mathbf{s}. \tag{1}$$

Since neither $\mathbf{s}$ nor the mixing process $\mathbf{A}$ are known and we have to estimate the inverse $\mathbf{C}$ of the mixing matrix blindly $\mathbf{y}(t) = \mathbf{C}\mathbf{x}(t) = \mathbf{C}\mathbf{A}\mathbf{s}(t) = \mathbf{\Lambda}\mathbf{P}\mathbf{s}(t) \propto \mathbf{s}$, only driven by the known (mixed) measurements $\mathbf{x}$. The unmixing is only possible up to scaling $\mathbf{\Lambda}$ and permutation $\mathbf{P}$. Ways to achieve this unmixing usually rely on minimizing a cost function that enforces statistical independence[1]. Mostly this involves an explicit or implicit calculation of higher-order moments, which can be difficult and in the case of scarce data or outliers error prone and furthermore computationally expensive.
Since many natural signals, like speech signals or biomedical signals (EEG, MEG) have a significant temporal structure, it is obvious to use the time-delayed second-order correlations for source separation [11, 3]. So in the following we will always assume a time structure (a pronounced autocorrelation)

[1] In a way the statistical independence assumption is the price we have to pay for *blind* separation, i.e if knowledge about the sensor array would be taken into account we could relax this assumption.

of the sources $\mathbf{s}$. In this work we propose to use several (> 2) time delayed correlation matrices for an exploitation of the temporal signal structure and discuss the implications of the choice of the delay parameters (section 2). Furthermore we illustrate the efficiency of our approach for a large example of mixed acoustic and synthetic signals (section 3).

2 Methods

Along the lines of Molgedey and Schuster [11] we define the following cost function

$$l_1(C_{ij}) = \sum_{i \neq j} \langle y_i(t) y_j(t) \rangle^2 + \sum_{i \neq j} \langle y_i(t) y_j(t+\tau) \rangle^2, \tag{2}$$

where τ is a certain time lag and $\langle\ \rangle$ denotes a time average. Inspecting Eq.(2) closer, we see that a minimum is reached, if the equal time and time lagged cross-correlations vanish simultaneously, i.e. Eq.(2) enforces *decorrelation over time*. The parameter τ must be chosen very carefully, so that the two correlation matrices carry maximally different information. To circumvey this selection problem we propose to use *several* time lagged correlation matrices (TDSEP - Temporal Decorrelation source SEParation), i.e. to minimize the generalized cost function

$$l_2(C_{ij}) = \sum_{i \neq j} \langle y_i(t) y_j(t) \rangle^2 + \sum_{k=1}^{N} \sum_{i \neq j} \langle y_i(t) y_j(t+\tau_k) \rangle^2, \tag{3}$$

Depending on the character of the problem it could become important to have both: delayed second-order correlations and higher-order moments. We can once more generalize Eq.(3) to obtain a cost function which makes use of the time structure *and* the information carried by the higher-order moments

$$l_3(C_{ij}) = \sum_{k=0}^{N} \sum_{i \neq j} \langle y_i(t) y_j(t+\tau_k) \rangle^2 + \sum_{k=0}^{N} \sum_{i \neq j} \langle f(y_i(t)) g(y_j(t+\tau_k)) \rangle^2, \tag{4}$$

where $\tau_0 = 0$. Using only the second term and setting $N = 0$ we can e.g. retrieve the ICA learning rule of Jutten and Herault [9] with some appropriate non-linear functions f, g. On the other side, we could also fully depend on the second -order time structure by using the first term in equation 4. So the amount of prior knowledge available for the very problem addressed would allow us to choose whether to rely more on higher -order moments or whether to use temporal information since the signals carry temporal structure. Since we analyse speech, music or biomedical signals, time structure alone is sufficient for unmixing, so – due to space limitations in this contribution – we will study only the use of Eq.(3) in this work.

Gradient descent Eq.(2)-(4) need to be minimized with respect to C, for example with simple gradient descent, $\Delta C \propto -\eta \partial l / \partial C$ (η is the learning rate), or other more sophisticated, e.g. 2nd -order, minimization methods. The solution of (2)-(4) starts to become a quite tricky optimization problem, if the parameter space increases, i.e. if the number of sources is large. Clearly in (4), so far, we have to resort to gradient decent optimization, but for Eq.(2)-(3) it is possible to find a clever method to keep the computational burden manageable.

Simultaneous Diagonalisation We define a (time-lagged) sample estimate of the correlation matrix as $\Sigma_{\tau(\mathbf{x})} = \langle \mathbf{x}(t)\mathbf{x}^T(t+\tau)\rangle$. Now we can write the optimization of the cost function as a linear algebra problem and to solve it via an eigenvalue decomposition of the correlation matrices. For *two* lagged correlation matrices $\Sigma_{(\mathbf{x})} = \langle \mathbf{x}\mathbf{x}^T\rangle$ and $\Sigma_{\tau(\mathbf{x})}$ the optimization problem can be solved via simultaneous diagonalisation [11] $(\Sigma_{\tau(\mathbf{x})}\Sigma_{(\mathbf{x})}^{-1})\mathbf{A} = \mathbf{A}\Lambda$. The quality of the solution, however, depends strongly on the very choice of τ (cf.Fig.1).

Whitening and Jacobi Rotations An alternative technique that can be applied for the joint diagonalization of two (or more) matrices proceeds in two steps: (1) whitening and (2) several Jacobi rotations to achieve an approximate simultaneous diagonalisation of the matrix set. This method has the advantage to be numerically more stable than the explicit solution of the simultaneous diagonalisation problem above. In step 1 we find a linear transform $\mathbf{W}$, such that the first term in Eq.(3) is set to zero explicitly. This means we use prior problem knowledge to restrict the solution of the optimization problem and by that to speed convergence. The whitening transform $\mathbf{W}$ can be determined by a principal component analysis or by taking the inverse square root of the covariance matrix via an eigenvalue decomposition as follows [6]:

$$\mathbf{W} = \Sigma_{(\mathbf{x})}^{-\frac{1}{2}} = (\mathbf{V}\,\Lambda\,\mathbf{V}^T)^{-\frac{1}{2}} = \mathbf{V}\,\Lambda^{-\frac{1}{2}}\,\mathbf{V}^T.$$

This transform $\mathbf{W}$ gives a representation of the sensor signals $\mathbf{x}$ in a new basis. We call these transformed signals $\mathbf{z}$. After the pre-whitening step, any time delayed correlation matrix of the transformed signals $\mathbf{z}$ should be (approximately) a diagonal matrix up to a transformation $\mathbf{Q}$ [3]. By construction, matrix $\mathbf{Q}$ is an orthogonal matrix, i.e. $\mathbf{Q}\mathbf{Q}^T = \mathbf{I}$. For the special case of two matrices, the rotation matrix $\mathbf{Q}$ can be obtained by the eigenvalue decomposition [6] of the time delayed correlation matrix

$$\Sigma_{\tau(\mathbf{z})} \quad = \quad \langle \mathbf{z}(t)\;\mathbf{z}^T(t+\tau)\rangle = \mathbf{Q}^T\,\Sigma_{\tau(\mathbf{s})}\,\mathbf{Q} = \mathbf{Q}^T\Lambda_\tau\,\mathbf{Q}. \tag{5}$$

For more than two matrices a trick proposed by Cardoso, which is based on the method, that Jacobi [8] published in 1846, can be used. The basic idea is, that one can approximate the rotation matrix $\mathbf{Q}$ by a sequence of elementary rotations $T_k(\phi_k)$ each trying to minimize the off diagonal elements of the respective $\Sigma_{\tau(\mathbf{x})}$ matrices, where the rotation angle ϕ_k can be calculated in closed form (see Cardoso [4]). The final rotation is then obtained by $\mathbf{Q} = \prod_k T_k(\phi_k)$.

Concatenation of both transforms (whitening $\mathbf{W}$ and rotation $\mathbf{Q}$) yields an estimate of the mixing matrix:

$$\hat{\mathbf{A}} = \mathbf{W}^{-1}\mathbf{Q}.$$

Summarizing: (1) one can choose one delay τ with heuristics or prior knowledge or (2) one can resort to determine $\mathbf{Q}$ such that several time delayed correlation matrices are simultaneous approximately diagonalised.

Averaging As a further simplification we suggest to average over the set of delay matrices $\tau_k (k = 1, \ldots, m)$: $\quad \Sigma_{\text{avg}} = \frac{1}{m} \sum_{k=1}^{m} (\Sigma_{\tau_k(\mathbf{z})})$ and use this averaged correlation matrix to compute one rotation $\mathbf{Q}$, instead of taking several Jacobi rotations $T_k(\phi_k)$ for every lagged correlation matrix. This idea will of course only give a crude approximation to the true minimization of Eq.(3), but it is very efficient in particular in high dimensional problems. This approximation becomes less crude if the assumption holds that the variance of the set of $\Sigma_{\tau_k(\mathbf{z})}$ matrices around the mean Σ_{avg} is small.

3 Experimental results

To illustrate the theoretical reasoning, numerous experiments were made. In order to evaluate the separation quality of the unmixed signals $\hat{\mathbf{y}}(t) = \hat{\mathbf{A}}^{-1}\mathbf{x}(t) = (\hat{\mathbf{A}}^{-1}\mathbf{A})\mathbf{s}(t)$, first a suitable measure is needed. This shows, that the closer the matrix $P_{ij} = (\hat{\mathbf{A}}^{-1}\mathbf{A})_{ij}$ is to a permutation matrix, the better the source signals are reconstructed. The interference of source i on channel j is defined as $\frac{|p_{ij}|}{\max_k(|p_{ik}|)}$. If the true mixing matrix is known, one can define the performance index as averaged interference [3, 1]

$$SIR(P) \quad = \quad \frac{1}{n} \sum_{i=1}^{n} \left(\frac{1}{n-1} \sum_{j=1}^{n-1} \frac{|p_{ij}|}{\max_k(|p_{ik}|)} - 1 \right) . \tag{6}$$

We compiled three different test datasets, each mixed on the computer by a randomly choosen, quadratic mixing matrix. (I) filtered Gaussian sources

$$\begin{aligned}
s_1(t) &= s_1(t-1) + 0.8\, s_1(t-2) + 0.6\, s_1(t-3) + \nu_1(t) \\
s_2(t) &= s_2(t-1) + 0.7\, s_2(t-2) + 0.3\, s_2(t-3) + \nu_2(t) \\
s_3(t) &= s_3(t-1) + 0.9\, s_2(t-2) + 0.15\, s_3(t-3) + \nu_3(t) \\
s_4(t) &= \sin(2\pi\, 3\, t) \\
\text{where} \quad \nu_i(t) &\ \text{is white Gaussian noise} \sim N(0,1) \ ,
\end{aligned}$$

(II) real acoustic signals (speech, music)

$$s_1(t) \ = \ \text{speech: “Bon giorno signora!”}, \quad s_2(t) = \text{music: “piccolo flute”},$$

(III) a hybrid mixture of 17 real acoustic signals and synthetic sources (super-, subgaussian and Gaussian).

Table 1 shows the performance indices (large SIR corresponds to bad performance) for various settings of τ. Clearly the performance can decrease drastically if the wrong delay is chosen, whereas a large number of delays always gives a stable (but not necessarily optimal) solution. So, in a practical experiment, where we have no knowledge, the use of many delays will always bring us to the safe side. It is interesting to note that the performance index can

SIR	τ_{best}	τ_{worst}	$\tau_{1..10}$	$\tau_{1..50}$	τ_{avg10}	JADE v1.5
Gaussians + sine	0.0032	973.6634	0.0054	0.0045	0.0221	0.2508
speech + music	0.0049	1.7815	0.0064	0.0051	0.0050	0.0049
17 sources	0.6284	11.4242	0.2799	0.2668	0.3035	1.1977
M FLOPS	τ_{best}	τ_{worst}	$\tau_{1..10}$	$\tau_{1..50}$	τ_{avg10}	JADE v1.5
Gaussians + sine	0.96	0.96	6.10	25.37	5.52	5.32
speech + music	0.32	0.32	2.00	8.39	1.80	0.69
17 sources	13.23	13.23	85.96	351.72	76.62	999.2

Table 1: Separation performance (SIR) and number of M FLOPS (in matlab implementation) using different strategies.

vary by more than two orders of magnitude and has a rather wiggly structure (cf. Fig.1)[2]. To explain this let us remember that as an assumption of the TDSEP algorithm we needed a pronounced time structure in the signals. In cases with bad performance indices the autocorrelation functions of the sources were too similar. This leads to ill-conditioned eigenvalue problems. As a baseline comparison we show the results for the JADE algorithm [3], which uses a simultaneous diagonalisation of 4th order cumulant martrices. TDSEP with several delays is working similarly good and in some cases even better than JADE (here the generic drawbacks of JADE with respect to Gaussian signals are crucial). Comparing the number of floating point operations we observe that JADE performs well for small problems, but for large number of sources clearly TDSEP is favoured. The simple averaging approximation also yields astonishingly good results, if we remember its simplistic assumptions.

4 Conclusion

We proposed the TDSEP algorithm (and some simplifications) for blind source separation based on only time-lagged second-order correlations. TDSEP takes two steps: (1) whitening and (2) a rotation (or several elementary (Jacobi) rotations) to achieve an approximate simultaneous diagonalisation of the set of time-lagged correlation matrices. For larger artificially mixed examples it was illustrated that the algorithm works well and computationally very efficient. This fact predisposes TDSEP for the analysis of biomedical signals, where we have as much as e.g. 49 channels and long recording times [14].

Future research will further study the simultaneous use of higher-order and second-order methods for source separation and the application of TDSEP to biomedical signals.[3].

Acknowledgements A.Z. was partly funded by DFG under contracts JA 379/51 and JA 379/71. We thank N.Murata for valuable discussions.

[2] Other data sets yield similar results.

[3] Further information can be obtained by `http://candy.first.gmd.de`.

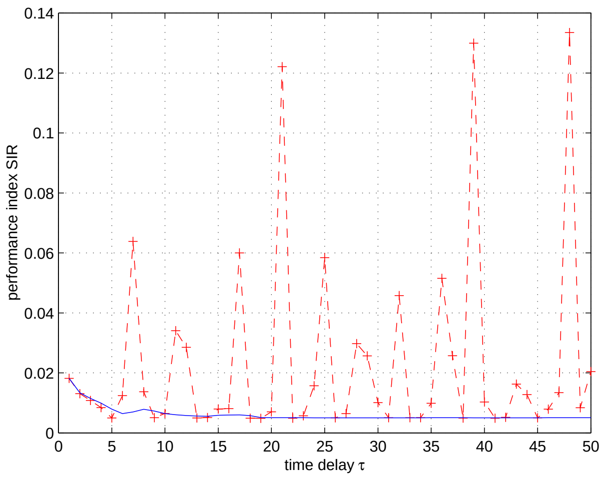

Figure 1: Performance index using one time delay ($0, \tau$, dashed) vs. several time delays ($0, \ldots, \tau$, solid) for speech and music.

References

[1] S. Amari, A. Cichocki, H. H. Yang. In *NIPS 95*, p. 882–893. 1996.

[2] A. J. Bell and T. J. Sejnowski. *Neural Computation*, 7(6):1129–1159, 1995.

[3] J.-F. Cardoso, A. Souloumiac. *IEEE Proceedings-F*, 14O(6):362–370, 1994.

[4] J.-F. Cardoso, A. Souloumiac. *SIAM J.Mat.Anal.Appl.*, 17(1):161 ff, 1996.

[5] G. Deco and D. Obradovic. *Neural Computation*, 7(2):338–348, 1995.

[6] G.H. Golub and C.F. van Loan. *Matrix Computation*. The Johns Hopkins University Press, London, 1989.

[7] A. Hyvärinen and E. Oja. *Neural Computation*, 9(7):1483–1492, 1997.

[8] C.G.J. Jacobi. *Crelle J. reine angew. Mathematik*,vol. 30, p. 51–94, 1846.

[9] Ch. Jutten and J. Herault. *Signal Processing*, p. 1–10, 1991.

[10] S. Makeig, T-P. Jung, D. Ghahremani, A.J. Bell, T.J. Sejnowski. *PNAS*, (94):10979–10984, 1997.

[11] L. Molgedey and H.G. Schuster. *Phys. Rev. Letters*, p. 3634–3637, 1994.

[12] N. Murata, K.-R. Müller, A. Ziehe, S. Amari. In *NIPS 96*, p. 599–607. The MIT Press, 1997.

[13] R.N. Vigario. *EEG and clinical Neurophysiology*, (103):395–404, 1997.

[14] A. Ziehe, K.R. Müller, G. Nolte, B.-M. Mackert, G. Curio. Artifact removal in magneto-neurography with time delayed second order correlations. submitted. 1998.