Weighted ICA-CPA

Lieven De Lathauwer*

In [12] a technique was presented for combining Independent Component Analysis (ICA) and Canonical / Parallel Factor Analysis, or Canonical Polyadic Analysis (CPA). The technique was called ICA-CPA. In this report we sketch how the different entries in [12, Eq. (25)] can be optimally weighted in the high SNR case. The derivation builds upon the results obtained in [4, 7, 3, 8, 9, 10, 13]. We use the same notation as in [12].

Let us first consider the real variant of [12, Eq. (4)] and assume that the mixing matrix is estimated on the basis of symmetrized covariance matrix estimates, as in [10, 13]. Given a tensor $\mathcal{U} \in \mathbb{R}^{L \times L \times K}$ in which the estimated symmetrized covariance matrices are stacked, uniform weighting corresponds to minimizing

$$f_1(\mathbf{M}, \mathbf{D}) = \sum_{l_1=1}^{L} \sum_{l_2=l_1+1}^{L} \sum_{k=1}^{K} (u_{l_1 l_2 k} - \sum_{r=1}^{R} m_{l_1 r} m_{l_2 r} d_{kr})^2. \tag{1}$$

(The diagonal entries are not taken into account for technical reasons.) Define a selection matrix $\mathbf{S}_{L^2} \in \mathbb{R}^{\frac{1}{2}L(L-1) \times L^2}$ that is such that, for any matrix $\mathbf{X} = \mathbf{X}_{[1;2]} \in \mathbb{R}^{L \times L}$, $\mathbf{S}_{L^2}\mathbf{x}_{[1,2]}$ contains the entries of the strictly upper triangular part of $\mathbf{X}$. Cost function f_1 can also be expressed as

$$f_1(\mathbf{M}, \mathbf{D}) = (\sum_{r=1}^{R} \mathbf{m}_r \otimes \mathbf{m}_r \otimes \mathbf{d}_r - \mathbf{u}_{[1,2,3]})^T \cdot (\mathbf{S}_{L^2} \otimes \mathbf{I}_K)^T \cdot (\mathbf{S}_{L^2} \otimes \mathbf{I}_K) \cdot (\sum_{r=1}^{R} \mathbf{m}_r \otimes \mathbf{m}_r \otimes \mathbf{d}_r - \mathbf{u}_{[1,2,3]}). \tag{2}$$

*L. De Lathauwer is with the Group Science, Engineering and Technology of the Katholieke Universiteit Leuven Campus Kortrijk, E. Sabbelaan 53, 8500 Kortrijk, Belgium, tel.: +32-(0)56-32.60.62, fax: +32-(0)56-24.69.99, e-mail: Lieven.DeLathauwer@kuleuven-kortrijk.be. He is also with the Department of Electrical Engineering (ESAT), Research Division SCD, of the Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium, tel.: +32-(0)16-32.86.51, fax: +32-(0)16-32.19.70, e-mail: Lieven.DeLathauwer@esat.kuleuven.be, URL: http://homes.esat.kuleuven.be/∼delathau/home.html.

Instead of minimizing f_1, we could also minimize

$$f_2(\mathbf{M},\mathbf{D}) = (\sum_{r=1}^{R} \mathbf{m}_r \otimes \mathbf{m}_r \otimes \mathbf{d}_r - \mathbf{u}_{[1,2,3]})^T \cdot (\mathbf{S}_{L^2} \otimes \mathbf{I}_K)^T \cdot \mathbf{W} \cdot (\mathbf{S}_{L^2} \otimes \mathbf{I}_K) \cdot (\sum_{r=1}^{R} \mathbf{m}_r \otimes \mathbf{m}_r \otimes \mathbf{d}_r - \mathbf{u}_{[1,2,3]}), \tag{3}$$

in which $\mathbf{W} \in \mathbb{R}^{\frac{1}{2}L(L-1)K \times \frac{1}{2}L(L-1)K}$ is a properly chosen weight matrix. It was shown in [7] that the optimal weight matrix (in the sense of mean square error) is given by the inverse of the covariance of the sample statistic $(\mathbf{S}_{L^2} \otimes \mathbf{I}_K) \cdot \mathbf{u}_{[1,2,3]}$. In practice, the optimal weight matrix can be taken equal to a consistent estimate, while preserving asymptotic optimality [4]. This is the approach taken in [13]. The technique is expensive because $\mathbf{W}$ is a large matrix. The computational complexity can be significantly reduced when the mixing matrix is close to the identity matrix [3, 10]. At high SNR, $\mathbf{W}$ can then be well approximated by a block-diagonal matrix, consisting of $\frac{1}{2}L(L-1)$ blocks of size $(K \times K)$. In [10] explicit expressions are given for the diagonal blocks for different types of Gaussian processes. The overall procedure is then as follows. In a first step, the sources are nearly separated by means of a fast algorithm based on suboptimal weighting (e.g. SOBI [1]). This yields a first estimate $\hat{\mathbf{M}}_1 \in \mathbb{R}^{L \times R}$ of the mixing matrix. The source estimates are used to estimate the block-diagonal matrix $\mathbf{W}$. Minimization of f_2 yields a matrix $\hat{\mathbf{M}}_2 \in \mathbb{R}^{R \times R}$. The overall mixing matrix estimate is given by $\hat{\mathbf{M}} = \hat{\mathbf{M}}_1 \cdot \hat{\mathbf{M}}_2$. To obtain an even better estimate, one may iterate a few times between alternating updates of $\mathbf{W}$ and $\hat{\mathbf{M}}_2$.

In ICA-CPA, the mixing matrix has the Khatri-Rao structure $\mathbf{M} = \mathbf{A} \odot \mathbf{B}$, where $\mathbf{A} \in \mathbb{R}^{I \times R}$ and $\mathbf{B} \in \mathbb{R}^{J \times R}$. Computation of the product $\hat{\mathbf{M}} = \hat{\mathbf{M}}_1 \cdot \hat{\mathbf{M}}_2$ would destroy this structure. Therefore we minimize

$$\begin{aligned} f_3(\mathbf{A},\mathbf{B},\mathbf{D}) = (\textstyle\sum_{r=1}^{R} \mathbf{a}_r \otimes \mathbf{b}_r \otimes \mathbf{a}_r \otimes \mathbf{b}_r \otimes \mathbf{d}_r - \mathbf{u}_{[1,2,3,4,5]})^T \cdot (\mathbf{S}_{I^2J^2} \otimes \mathbf{I}_K)^T \cdot \mathbf{W} \\ \cdot (\mathbf{S}_{I^2J^2} \otimes \mathbf{I}_K) \cdot (\textstyle\sum_{r=1}^{R} \mathbf{a}_r \otimes \mathbf{b}_r \otimes \mathbf{a}_r \otimes \mathbf{b}_r \otimes \mathbf{d}_r - \mathbf{u}_{[1,2,3,4,5]}), \end{aligned} \tag{4}$$

in which $\mathbf{u}_{[1,2,3,4,5]}$ represents the original set of covariance matrices, i.e., without prior approximate unmixing. The computational complexity can now be significantly reduced as follows. First we compute estimates $\hat{\mathbf{A}}_1$ and $\hat{\mathbf{B}}_1$ by means of a suboptimally weighted scheme. For the nearly separated sources $\hat{\mathbf{M}}_1^{\dagger} \cdot \mathbf{x} = (\hat{\mathbf{A}}_1 \odot \hat{\mathbf{B}}_1)^{\dagger} \cdot \mathbf{x}$, a block-diagonal weighting matrix $\tilde{\mathbf{W}} \in \mathbb{R}^{\frac{1}{2}R(R-1)K \times \frac{1}{2}R(R-1)K}$ can be computed as above. Optimal weighting can be implemented by replacing the matrix

$$(\mathbf{S}_{I^2J^2} \otimes \mathbf{I}_K)^T \cdot \mathbf{W} \cdot (\mathbf{S}_{I^2J^2} \otimes \mathbf{I}_K)$$

in (4) by the matrix

$$(\hat{\mathbf{M}}_1^{\dagger,T} \otimes \hat{\mathbf{M}}_1^{\dagger,T} \otimes \mathbf{I}_K) \cdot (\mathbf{S}_{I^2J^2} \otimes \mathbf{I}_K)^T \cdot \tilde{\mathbf{W}} \cdot (\mathbf{S}_{I^2J^2} \otimes \mathbf{I}_K) \cdot (\hat{\mathbf{M}}_1^{\dagger} \otimes \hat{\mathbf{M}}_1^{\dagger} \otimes \mathbf{I}_K).$$

However, this large matrix never needs to be formed explicitly. Namely, for any tensor $\mathcal{V} \in \mathbb{R}^{I \times J \times I \times J \times K}$ the vector

$$\tilde{\mathbf{v}} = (\hat{\mathbf{M}}_1^{\dagger} \otimes \hat{\mathbf{M}}_1^{\dagger} \otimes \mathbf{I}_K) \cdot \mathbf{v}_{[1,2,3,4,5]} \tag{5}$$

is a representation of $\mathcal{V}_{[1,2;3,4;5]} \cdot_1 \hat{\mathbf{M}}_1^{\dagger} \cdot_2 \hat{\mathbf{M}}_1^{\dagger} \in \mathbb{R}^{R \times R \times K}$ and can be computed as such, without forming $\hat{\mathbf{M}}_1^{\dagger} \otimes \hat{\mathbf{M}}_1^{\dagger} \otimes \mathbf{I}_K$. In addition, the pseudo-inverse $\hat{\mathbf{M}}_1^{\dagger}$ is given by [11]

$$(\hat{\mathbf{A}}_1 \odot \hat{\mathbf{B}}_1)^{\dagger} = ((\hat{\mathbf{A}}_1^H \cdot \hat{\mathbf{A}}_1) * (\hat{\mathbf{B}}_1^H \cdot \hat{\mathbf{B}}_1))^{-1} \cdot (\hat{\mathbf{A}}_1 \odot \hat{\mathbf{B}}_1)^H$$

if $(\hat{\mathbf{A}}_1^H \cdot \hat{\mathbf{A}}_1) * (\hat{\mathbf{B}}_1^H \cdot \hat{\mathbf{B}}_1)$ is nonsingular, which is generically the case if $R \leqslant IJ$. The latter is a consequence of [5, Lemma 2.2]. Combining our two observations, vector $\tilde{\mathbf{v}}$ can be computed efficiently as follows. First we compute the entries $f_{r_1 r_1 r_2 r_2 k}$, $1 \leqslant r_1, r_2 \leqslant R$, $1 \leqslant k \leqslant K$, of the tensor

$$\mathcal{F} = \mathcal{V} \cdot_1 \hat{\mathbf{A}}_1^T \cdot_2 \hat{\mathbf{B}}_1^T \cdot_3 \hat{\mathbf{A}}_1^T \cdot_4 \hat{\mathbf{B}}_1^T \tag{6}$$

and stack them in a tensor $\tilde{\mathcal{F}} = [\tilde{f}_{r_1 r_2 k} = f_{r_1 r_1 r_2 r_2 k}] \in \mathbb{R}^{R \times R \times K}$. Next, we compute

$$\mathcal{G} = \mathcal{F} \cdot_1ig((\hat{\mathbf{A}}_1^T \cdot \hat{\mathbf{A}}_1) * (\hat{\mathbf{B}}_1^T \cdot \hat{\mathbf{B}}_1)\big)^{-1} \cdot_2 \big((\hat{\mathbf{A}}_1^T \cdot \hat{\mathbf{A}}_1) * (\hat{\mathbf{B}}_1^T \cdot \hat{\mathbf{B}}_1)\big)^{-1}. \tag{7}$$

Vector $\tilde{\mathbf{v}}$ in (5) is then equal to $\mathbf{g}_{[1,2,3]}$.

When the tensor $\mathcal{V}$ has rank 1,

$$\mathcal{V} = \mathbf{x}_1 \circ \mathbf{y}_1 \circ \mathbf{x}_2 \circ \mathbf{y}_2 \circ \mathbf{z},$$

with $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^I$, $\mathbf{y}_1, \mathbf{y}_2 \in \mathbb{R}^J$ and $\mathbf{z} \in \mathbb{R}^K$, then the computation can be further simplified. Expressions (5–7) imply that

$$\begin{aligned}\mathcal{G} = &\Big[((\hat{\mathbf{A}}_1^T \cdot \hat{\mathbf{A}}_1) * (\hat{\mathbf{B}}_1^T \cdot \hat{\mathbf{B}}_1))^{-1} \cdot \mathrm{diag}((\hat{\mathbf{A}}_1^T \cdot \mathbf{x}_1) \cdot (\hat{\mathbf{B}}_1^T \cdot \mathbf{y}_1)^T)\Big] \\ &\circ \Big[((\hat{\mathbf{A}}_1^T \cdot \hat{\mathbf{A}}_1) * (\hat{\mathbf{B}}_1^T \cdot \hat{\mathbf{B}}_1))^{-1} \cdot \mathrm{diag}((\hat{\mathbf{A}}_1^T \cdot \mathbf{x}_2) \cdot (\hat{\mathbf{B}}_1^T \cdot \mathbf{y}_2)^T)\Big] \circ \mathbf{z}.\end{aligned} \tag{8}$$

This shows that it is not needed to explicitly construct the rank-1 tensor $\mathcal{V}$ either.

The weighted cost function can be minimized by means of exact line search. Substitution of [12, Eq. (26)] in the expression of the function yields a polynomial of degree 10 in λ, just like in the minimization of [12, Eq. (25)]. The coefficients of the polynomial can be computed efficiently by making use of (7–8) and the block-diagonality of $\tilde{\mathbf{W}}$.

In the complex case, we need to distinguish between two types of problems. If one works with Hermitean matrices, like in the complex variant of [6], then the derivation above can be repeated with just the transpose replaced by the complex conjugated transpose. However, in [1] for instance, all but one covariance matrices are unsymmetric. One then additionally needs to replace the matrix $\mathbf{S}_{I^2J^2} \otimes \mathbf{I}_K$ in (4) by a selection matrix $\tilde{\mathbf{S}}_{L^2K} \in \mathbb{R}^{(L^2K - \frac{1}{2}L(L+1) - (K-1)L) \times L^2K}$ that drops the diagonals of all matrices and the lower triangular part of only the Hermitean matrix. The actual line search works in the same way in both cases. Exact line search with a real parameter leads to the minimization of a polynomial of degree 10. Line search with a complex parameter leads to a cost function that depends on the step size in a similar way as the cost function in [12, Eq. (25)]. Hence, for the determination of the step size one can work in analogy with [12, Appendix].

If the mixing matrix is estimated on the basis of a covariance matrix and cumulant slices, then one can develop a similar reasoning starting from [8].

The fact that, in the minimization of a cost function that is expressed in terms of the statistics of the mixed data, the computational complexity can nevertheless be reduced by only estimating statistics of nearly unmixed data, can be used in other applications where a post-processing as in [3, 8, 10] would destroy the structure of the mixing matrix. One can for instance think of applications in which the mixing matrix has an affine structure. An example is the convolutive variant of SOBI, in which the mixing matrix has a Toeplitz-block structure [2].

Acknowledgments

Research supported by: (1) Research Council K.U.Leuven: GOA-MaNet, CoE EF/05/006 Optimization in Engineering (OPTEC), CIF1 and STRT1/08/023 (2) F.W.O.: (a) project G.0427.10N, (b) Research Communities ICCoS, ANMMM and MLDM, (3) the Belgian Federal Science Policy Office: IUAP P6/04 (DYSCO, "Dynamical systems, control and optimization", 2007–2011), (4) EU: ERNSI.

References

[1] A. Belouchrani, K. Abed-Meraim, J.-F. Cardoso, and E. Moulines. A blind source separation technique using second order statistics. *IEEE Trans. Signal. Proc.*, 45:434–444, 1997.

[2] H. Bousbia-Salah, A. Belouchrani, and K. Abed-Meraim. Jacobi-like algorithm for blind signal separation of convolutive mixtures. *Electronics Letters*, 37:1049–1050, 2001.

[3] J.-F. Cardoso, S. Bose, and B. Friedlander. Output cumulant matching for source separation. In *Proc. IEEE SP Workshop on Higher-Order Stat.*, pages 44–48, Aiguablava, Spain, 1995.

[4] B. Friedlander and B. Porat. Asymptotically optimal estimation of MA and ARMA parameters of non-Gaussian processes from high-order moments. *IEEE Trans. on Automatic Control*, 35:27–35, 1990.

[5] L. De Lathauwer. A link between the canonical decomposition in multilinear algebra and simultaneous matrix diagonalization. *SIAM J. Matrix Anal. Appl.*, 28:642–666, 2006.

[6] D.-T. Pham and J.-F. Cardoso. Blind separation of instantaneous mixtures of non stationary sources. *IEEE Trans. Signal Processing*, 49:1837–1848, 2001.

[7] B. Porat and B. Friedlander. Performance analysis of parameter estimation based on high order moments. *Journal of Adaptive Control and Signal Processing*, 3:191–229, 1998.

[8] A. Smekhov and A. Yeredor. Optimization of JADE using a novel optimally weighted joint diagonalization approach. In *Proc. of the 12th European Signal Processing Conference EUSIPCO*, Vienna, Austria, September 2004.

[9] P. Tichavský, E. Doron, A. Yeredor, and J. Nielsen. A computationally affordable implementation of an asymptotically optimal BSS algorithm for AR sources. In *Proc. of EUSIPCO*, Florence, Italy, 2006.

[10] P. Tichavský and A. Yeredor. Fast approximate joint diagonalization incorporating weight matrices. *IEEE Trans. Signal. Proc.*, 57:878–891, 2009.

[11] G. Tomasi. *Practical and Computational Aspects in Chemometric Data Analysis*. PhD thesis, University of Copenhagen, 2006.

[12] M. De Vos, D. Nion, S. Van Huffel, and L. De Lathauwer. A combination of parallel factor and independent component analysis. *Tech. Report 09-04, ESAT-SISTA, K.U.Leuven (Leuven, Belgium)*, 2009.

[13] A. Yeredor. Blind separation of Gaussian sources via second-order statistics with asymptotically optimal weighting. *IEEE Signal Processing Letters*, 7:197–200, 2000.