SPATIOTEMPORAL BLIND SOURCE SEPARATION USING DOUBLE-SIDED APPROXIMATE JOINT DIAGONALIZATION

Fabian J. Theis, Peter Gruber, Ingo R. Keck, Anke Meyer-Bäse[*] *and Elmar W. Lang*

Institute of Biophysics, University of Regensburg, 93040 Regensburg, Germany
phone: +49 941 943 2924, fax: +49 941 943 2479, email: fabian@theis.name
[*]Department of Electrical and Computer Engineering, Florida State University
Tallahassee, FL 32310-6046, USA

ABSTRACT

In independent component analysis (ICA) the common task is to achieve either spatial or temporal independence by linearly mapping into a feature space. If the data possesses both spatial and temporal structures such as a sequence of images or 3d-scans taken at fixed time intervals, we can require the transformed data to be as independent as possible in both domains. First introduced by Stone using a joint entropy energy function, spatiotemporal ICA is a promising method for real-world data analysis. We propose a novel algorithm for performing spatiotemporal ICA by jointly diagonalizing various source conditions such as higher-order cumulants of the mixtures, both in time *and* in space. Similar to algebraic ICA algorithms, this provides a robust method for data analysis, which is confirmed by simulations.

1. INTRODUCTION

Blind source separation (BSS) describes the task of recovering the unknown mixing process and the underlying sources of an observed data set. Currently, many BSS algorithms assume either independence (ICA) or auto-decorrelation of the sources, see for instance [3] and references therein. Spatiotemporal ICA in comparison to the more common methods of either spatial or temporal analysis tries to achieve both spatial and temporal separation by optimizing a joint energy function. First proposed by Stone et al [6], it is a promising method, which has potential applications in biomedical data analysis. We extend his approach by generalizing algebraic ICA algorithms to the spatiotemporal case.

2. BLIND SOURCE SEPARATION

We consider the following *blind source separation* (BSS) problem: Let $\mathbf{x}(t)$ be an (observed) stationary m-dimensional stochastical process (with not necessarily discrete time t) and $\mathbf{A}$ a full rank matrix such that

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) + \mathbf{n}(t) \qquad (1)$$

where the n-dimensional source signals $\mathbf{s}(t)$ fulfill additional properties such as:

- they are stochastically independent: $p_{\mathbf{s}}(s_1, \ldots, s_n) = p_{s_1}(s_1) \ldots p_{s_n}(s_n)$,
- each source is sparse i.e. it contains a certain number of zeros or has a low p-norm for small p,
- for all τ, they have diagonal *autocovariances* $E(\mathbf{s}(t+\tau)\mathbf{s}(t)^\top)$ (zero-mean $\mathbf{s}(t)$ are assumed).

In the following, we will derive a BSS algorithm framework for spatiotemporal data sets. Thereby, one of the above conditions is denoted by the term *source condition*, if we do not want to specialize on a single case. The additive noise $\mathbf{n}(t)$ is modelled by a stationary, temporally and spatially white zero-mean process with variance σ^2. As usual, we further assume that at most as many sources as sensors are to be extracted, i.e. $m \geq n$.

$\mathbf{x}(t)$ is observed, and the goal is to recover $\mathbf{A}$ and $\mathbf{s}(t)$. Having found $\mathbf{A}$, $\mathbf{s}(t)$ can be estimated by $\mathbf{A}^\dagger\mathbf{x}(t)$, which is optimal in the maximum-likelihood sense. Here † denotes the pseudo-inverse of $\mathbf{A}$, which equals the inverse in the case of $m = n$. So the BSS task reduces to the estimation of the mixing matrix $\mathbf{A}$, hence the additive noise $\mathbf{n}$ is often neglected (after whitening). Note that in the following we will assume that all signals are real-valued. Extensions to the complex case are straightforward.

3. SPATIOTEMPORAL BSS

In contrast to the theory, real-world data sets often possess structure in addition to the necessary instantaneous independence required by ICA. For example fMRI measurements contain both temporal and spatial indices so a data entry $x = x(a,b,c,t)$ can depend on position (a,b,c) as well as time t. More generally, we want to consider data sets $x(\mathbf{r},t)$ depending on two indices $\mathbf{r}$ and t, where $\mathbf{r} \in \mathbb{R}^n$ can be a multidimensional index and t indexes the time axis. In reality this generalized random process is realized by a finite number of samples. For example in the case of fMRI scans we could assume $t \in [1:T] := \{1,2,\ldots,T\}$ and $\mathbf{r} \in [1:h] \times [1:w] \times [1:d]$, where T is the number of scans, which were of size $h \times w \times d$. So the number of spatial observations is ${}^{\mathbf{s}}m := hwd$ and the number of temporal observations ${}^{\mathbf{t}}m = T$.

3.1 Spatial and temporal BSS

For such multi-structured data, two methods of BSS analysis exist. In *temporal BSS*, the data is interpreted to contain a measured time series $x_{\mathbf{r}}(t) := x(\mathbf{r},t)$ for each spatial location $\mathbf{r}$. The goal is then to apply BSS to the *temporal observation vector* ${}^{\mathbf{t}}\mathbf{x}(t) := (x_{\mathbf{r}_{111}}(t), \ldots, x_{\mathbf{r}_{hwd}}(t))^\top$ containing ${}^{\mathbf{s}}m$ entries i.e. consisting of ${}^{\mathbf{s}}m$ spatial observations. In other words we are looking for a decomposition ${}^{\mathbf{t}}\mathbf{x}(t) = {}^{\mathbf{t}}\mathbf{A}\,{}^{\mathbf{t}}\mathbf{s}(t)$ with the *temporal mixing matrix* ${}^{\mathbf{t}}\mathbf{A}$ and *temporal sources* ${}^{\mathbf{t}}\mathbf{s}(t)$, possibly of lower dimension.

This contrasts to so-called *spatial BSS*, where the data is considered to be composed of T spatial patterns $\mathbf{x}_t(\mathbf{r}) := x(\mathbf{r},t)$. *Spatial BSS* tries to decompose the *spatial observation vector* ${}^{\mathbf{s}}\mathbf{x}(\mathbf{r}) := (x_{t_1}(\mathbf{r}), \ldots, x_{t_T}(\mathbf{r}))^\top \in \mathbb{R}^{{}^{\mathbf{t}}m}$ into ${}^{\mathbf{s}}\mathbf{x}(\mathbf{r}) = {}^{\mathbf{s}}\mathbf{A}\,{}^{\mathbf{s}}\mathbf{s}(\mathbf{r})$ with a *spatial mixing matrix* ${}^{\mathbf{s}}\mathbf{A}$ and *spatial sources* ${}^{\mathbf{s}}\mathbf{s}(\mathbf{r})$, possibly of lower dimension.

Often, the spatial multi-dimensional index $\mathbf{r}$ is contracted into a one-dimensional index r, for instance by row, column or slice concatenation. Then the data set $x(r,t) =: x_{rt}$ can be represented by a data matrix $\mathbf{X}$ of dimension ${}^{\mathbf{s}}m \times {}^{\mathbf{t}}m$, and the goal is to determine a source matrix $\mathbf{S}$, either spatially or temporally.

3.2 Preprocessing – mean removal

By subtracting first the temporal (sample) mean ${}^{\mathbf{t}}\mu_{\mathbf{X}} := (1/{}^{\mathbf{t}}m \sum_t x_{rt})_r$ of $\mathbf{X}$ to get $\tilde{\mathbf{X}}$ and then the spatial mean ${}^{\mathbf{s}}\mu_{\tilde{\mathbf{X}}} = (1/{}^{\mathbf{s}}m \sum_r \tilde{x}_{rt})_t$, we can assume that the mixtures are spatiotemporally centered. This corresponds to allowing for *affine* linear transformations both temporally and spatially. The coefficients of the centered data set $\bar{\mathbf{X}}$ can simply be calculated by

$$\bar{x}_{r_0t_0} = x_{r_0t_0} - \frac{1}{{}^{\mathbf{s}}m}\sum_r x_{rt_0} - \frac{1}{{}^{\mathbf{t}}m}\sum_t x_{r_0t} + \frac{1}{{}^{\mathbf{s}}m\,{}^{\mathbf{t}}m}\sum_{r,t} x_{rt}.$$

3.3 Why factorization into three terms fails

The data set $\mathbf{X}$ consists of temporal observations in the rows and spatial observations in the columns. One possible extension of the common source separation would be to require the source conditions (for instance perfect independence) both temporally and spatially. In order to achieve such a separation it could be allowed to transform the data both spatially and temporally, so the goal is to determine mixing matrices ${}^{\mathbf{s}}\mathbf{A}$ and ${}^{\mathbf{t}}\mathbf{A}$ with

$$\mathbf{X} = {}^{\mathbf{t}}\mathbf{A}\mathbf{S}\,{}^{\mathbf{s}}\mathbf{A}^\top, \tag{2}$$

where $\mathbf{S}$ fulfills the spatiotemporal conditions fully. In the following, we will show why such a 'three-term' factorization approach fails in most cases.

Almost all source conditions include decorrelation i.e. principal component analysis, typically as preprocessing step or incorporated into the algorithm itself. If we require $\mathbf{S}$ to be spatiotemporally decorrelated, we would be searching for matrices ${}^{\mathbf{s}}\mathbf{W}$ and ${}^{\mathbf{t}}\mathbf{W}$ such that $\mathbf{Y} := {}^{\mathbf{t}}\mathbf{W}\mathbf{X}\,{}^{\mathbf{s}}\mathbf{W}^\top$ has vanishing spatiotemporal covariances. Since $\mathbf{X}$ and hence $\mathbf{Y}$ are spatiotemporally centered, this means $\mathbf{Y}\mathbf{Y}^\top \propto \mathbf{I}$ and $\mathbf{Y}^\top\mathbf{Y} \propto \mathbf{I}$. One such set of whitening matrices $\mathbf{W}$ (and hence all since they all are constructed from each other by left-multiplication by orthogonal matrices) can be constructed as follows:

Consider the singular value decomposition $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$ of $\mathbf{X}$. Here $\mathbf{D}$ is a diagonal nonnegative square matrix of size $\min\{{}^{\mathbf{s}}m, {}^{\mathbf{t}}m\}$, and $\mathbf{U}$ and $\mathbf{V}$ are pseudo-orthogonal meaning that they have orthogonal columns ($\mathbf{U}^\top\mathbf{U} = \mathbf{V}^\top\mathbf{V} = \mathbf{I}$). Defining ${}^{\mathbf{t}}\mathbf{W} := \mathbf{D}^{-1/2}\mathbf{U}^\top$ and ${}^{\mathbf{s}}\mathbf{W} := \mathbf{D}^{-1/2}\mathbf{V}^\top$ yields the desired result as can be easily checked. But $\mathbf{Y} = {}^{\mathbf{t}}\mathbf{W}\mathbf{X}\,{}^{\mathbf{s}}\mathbf{W}^\top = \mathbf{D}^{-1/2}\mathbf{U}^\top\mathbf{X}\mathbf{V}\mathbf{D}^{-1/2} = \mathbf{D}^{-1/2}\mathbf{U}^\top\mathbf{U}\mathbf{D}\mathbf{V}^\top\mathbf{V}\mathbf{D}^{-1/2} = \mathbf{I}$ so simple spatiotemporal whitening already renders the source data set trivial. Any whitening matrix factorizes over the above matrices $\mathbf{W}$, hence this represents an inherent problem of double-sided whitening or, for that matter, of *any* factorization given by equation (2).

3.4 The solution: spatiotemporal matrix factorization

Temporal BSS is equivalent to the matrix factorization $\mathbf{X} = {}^{\mathbf{t}}\mathbf{A}\,{}^{\mathbf{t}}\mathbf{S}$, whereas spatial BSS implies the factorization $\mathbf{X}^\top = {}^{\mathbf{s}}\mathbf{A}\,{}^{\mathbf{s}}\mathbf{S}$ or equivalently $\mathbf{X} = {}^{\mathbf{s}}\mathbf{S}^\top\,{}^{\mathbf{s}}\mathbf{A}^\top$. Hence

$$\mathbf{X} = {}^{\mathbf{t}}\mathbf{A}\,{}^{\mathbf{t}}\mathbf{S} = {}^{\mathbf{s}}\mathbf{S}^\top\,{}^{\mathbf{s}}\mathbf{A}^\top \tag{3}$$

So both source separation models can be interpreted as matrix factorization problems; in the temporal case restrictions such as independence are put onto the second factor, in the spatial case onto the first one. In order to achieve a spatiotemporal model, which includes both these conditions, a three term approach has turned out to be too general to yield useful results. But equation (3) gives an idea how to proceed. Instead of recovering a single source data set which fulfills the source conditions spatiotemporally we try to find *two* source matrices, a spatial and a temporal source matrix, and the conditions are put onto the matrices separately. So the *spatiotemporal BSS* model can be formulated by the factorization problem

$$\mathbf{X} = {}^{\mathbf{s}}\mathbf{S}^\top\,{}^{\mathbf{t}}\mathbf{S} \tag{4}$$

with spatial source matrix ${}^{\mathbf{s}}\mathbf{S}$ and temporal source matrix ${}^{\mathbf{t}}\mathbf{S}$, which both have to fulfill the source conditions as much as possible. Later we will specify in more detail what we mean by 'as much as possible' using a weighted cost function. Any spatiotemporal model should have extremal solutions of spatial respectively temporal BSS depending on the weight — we will confirm this property later for our proposed model.

The source conditions are typically invariant under scaling and transformation, so the above model contains the same indeterminacy — indeed the spatial and temporal sources can interchange scaling ($\mathbf{L}$) and permutation ($\mathbf{P}$) matrices, ${}^{\mathbf{s}}\mathbf{S}^\top\,{}^{\mathbf{t}}\mathbf{S} = (\mathbf{L}^{-1}\mathbf{P}^{-1}\,{}^{\mathbf{s}}\mathbf{S})^\top(\mathbf{L}\mathbf{P}\,{}^{\mathbf{t}}\mathbf{S})$. Apart from that, in the case in which the conditions are fulfilled perfectly, the proofs of temporal uniqueness [4, 7] can easily be transferred to the above problem. However, if the source conditions hold jointly but only approximately for ${}^{\mathbf{s}}\mathbf{S}$ and ${}^{\mathbf{t}}\mathbf{S}$, uniqueness results are unknown so far.

After having successfully separated the data, the previously subtracted spatiotemporal mean can be incorporated into the sources (to get first-order equality in the model (4) in the case of non-centered mixtures) by adding the transformed spatiotemporal means: The new non-centered spatial sources are estimated by ${}^{\mathbf{s}}\mathbf{S} + {}^{\mathbf{t}}\mathbf{S}^{\dagger\top}\,{}^{\mathbf{s}}\mu_{\mathbf{X}}$ and the non-centered temporal sources by ${}^{\mathbf{t}}\mathbf{S} + {}^{\mathbf{s}}\mathbf{S}^{\dagger\top}\,{}^{\mathbf{t}}\mu_{\mathbf{X}}$.

4. AN ALGORITHM FOR SPATIOTEMPORAL BSS

Stone [6] first proposed the model from equation (4), where he employs a joint energy function based on mutual entropy and infomax. Apart from the many parameters used in the algorithm, the involved gradient descent optimization is susceptible to noise, local minima and inappropriate initializations, so we propose a novel, more robust algebraic approach based on joint diagonalization in the following.

4.1 Source conditions

In order to work within a general BSS framework, we allowed different source conditions, see section 2. We will now make the further restriction that such a source condition can be formulated by a criterion specifying the diagonality of a set of matrices, which can be estimated from the data.

We will formulate the conditions for an m-dimensional centered random vector $\mathbf{x}$. The expectation operator is denoted by $E(\mathbf{x}) \in \mathbb{R}^m$. If N realizations i.e. samples $\mathbf{x}(1), \ldots, \mathbf{x}(N)$ of $\mathbf{x}$ are given, E is estimated by the sample mean $\frac{1}{N}\sum_i \mathbf{x}(i)$ as usual.

Let $\mathbf{C}_1(\mathbf{x})$ be a square matrix that is to be diagonalized, depending on the source condition — often multiple such $\mathbf{C}_1(\mathbf{x}), \ldots, \mathbf{C}_K(\mathbf{x})$ are constructed for a single source condition, for example:

- If the sources are to be decorrelated, the matrix $\mathbf{C}_1(\mathbf{x})$ is simply the estimated *covariance* $\mathbf{C}_1(\mathbf{x}) := \mathbf{R}_{\mathbf{x}} := E(\mathbf{x}\mathbf{x}^\top)$.
- If the sources are assumed to be independent (ICA), then the fourth-order cross cumulants of the sources have to be trivial. In order to find transformations of the mixtures fulfilling this property, the well-known JADE algorithm [2] jointly diagonalizes the *contracted quadricovariance matrices* defined by $\mathbf{C}_{ij}(\mathbf{x}) := E\left(\mathbf{x}^\top\mathbf{E}_{ij}\mathbf{x}\mathbf{x}\mathbf{x}^\top\right) - \mathbf{R}_{\mathbf{x}}\mathbf{E}_{ij}\mathbf{R}_{\mathbf{x}} - \operatorname{tr}(\mathbf{E}_{ij}\mathbf{R}_{\mathbf{x}})\mathbf{R}_{\mathbf{x}} - \mathbf{R}_{\mathbf{x}}\mathbf{E}_{ij}\mathbf{R}_{\mathbf{x}}$. Here $\mathbf{E}_{ij}$ is a set of eigen-matrices of $\mathbf{C}_{ij}$, $1 \le i, j \le m$. One simple choice is to use m^2 matrices $\mathbf{E}_{ij}$ with zeros everywhere except 1 at index (i, j). More elaborate choices of eigen-matrices (with only $m(m+1)/2$ or even m entries) are discussed in [3], section 4.C.
- Instead of diagonalizing fourth-order (contracted) cumulants, other-order moments can be used such as *third-order cumulants* in order to account for non-symmetric, skew data: $\mathbf{C}_i(\mathbf{x}) := E\left(\mathbf{x}_i\mathbf{x}\mathbf{x}^\top\right)$ Here $1 \le i \le m$. This can be further extended by jointly diagonalizing different-order cumulants as proposed in the eJADE algorithm [5].
- Another source assumption can be made in the case of non i.i.d. signals (and different source power spectra). Then source identification can be performed by diagonalization of the *autocovariances* $\mathbf{C}_\tau(\mathbf{x}) := E\left(\mathbf{x}(t+\tau)\mathbf{x}(t)^\top\right)$ for a given set of delays τ. The so-called AMUSE algorithm uses a single τ, whereas SOBI [1] jointly diagonalizes a whole set of such delays.

- Finally, for data sets that possess multidimensional parametrizations as for example sets of images or 3d-scans, the above approach can be generalized to the diagonalization of *multidimensional autocovariances* $\mathbf{C}_{\tau_1,\ldots,\tau_M}(\mathbf{x}) := E\left(\mathbf{x}(t_1+\tau_1,\ldots,t_M+\tau_M)\mathbf{x}(t_1,\ldots,t_M)^\top\right)$ for a single or multiple given delay vectors $(\tau_1,\ldots,\tau_M)$. This is the basic principle of the multidimensional SOBI (mdSOBI) algorithm [8].

Other choices of condition matrices $\mathbf{C}_i(\mathbf{x})$ are possible. We only require two properties (which are fulfilled by the above examples): the matrices $\mathbf{C}_i(\mathbf{s})$ must be diagonal for all i when evaluated for the source random vector $\mathbf{s}$; furthermore they must transform as $\mathbf{C}_i(\mathbf{W}\mathbf{x}) = \mathbf{W}\mathbf{C}_i(\mathbf{x})\mathbf{W}^\top$ for all matrices $\mathbf{W}$. Finally note that using the substitution $\bar{\mathbf{C}}_i(\mathbf{x}) := \mathbf{C}_i(\mathbf{x}) + \mathbf{C}_i(\mathbf{x})^\top$, we can assume $\mathbf{C}_i(\mathbf{x})$ to be symmetric.

4.2 Approximate joint diagonalization

Many BSS algorithms employ diagonalization techniques on some of the above source conditions to identify a mixing matrix. Given a set of symmetric matrices $\mathscr{C} := \{\mathbf{C}_1,\ldots,\mathbf{C}_K\}$, such a matrix can be found by minimizing

$$\sum_{k=1}^{K} \operatorname{off}\left(\hat{\mathbf{A}}^\top \mathbf{C}_i \hat{\mathbf{A}}\right) \tag{5}$$

with respect to the orthogonal matrix $\hat{\mathbf{A}}$, where off denotes the sum of the off-diagonal terms. A global minimum of this function is called *joint diagonalizer* of $\mathscr{C}$. A sufficient criterion for existence of such a joint diagonalizer is that all elements of $\mathscr{C}$ commute. Algorithms for performing joint diagonalization include gradient descent on the function from equation (5), iterative construction of $\mathbf{A}$ by Givens rotation in two coordinates [2] or an iterative two-step recovery of $\mathbf{A}$ [9], where the latter algorithm can also search for non-orthogonal matrices $\mathbf{A}$. Joint diagonalization has been used in BSS using cumulant matrices [2] or temporal autocovariances [1].

Note that in practice minimization of the off-sums only gives an *approximate joint diagonalizer* — in the case of finite samples, the source condition matrices are only estimates and hence they only approximately share the same eigenstructure, so the value of equation (5) cannot be rendered precisely zero but only approximately.

4.3 Double-sided joint diagonalization

Now we can finally derive an algorithm for the spatiotemporal BSS problem (4); it is based on the joint diagonalization of source conditions posed not only temporally but also spatially.

Shifting to matrix notation, we interpret $\mathbf{C}_i(\mathbf{X}) := \mathbf{C}_i({}^{\mathbf{t}}\mathbf{x}(t))$ as a temporal condition matrix, whereas $\mathbf{C}_i(\mathbf{X}^\top) := \mathbf{C}_i({}^{\mathbf{s}}\mathbf{x}(r))$ is to denote the corresponding spatial condition matrix. Application of the spatiotemporal mixing model from equation (4) together with the transformation properties of $\mathbf{C}_i$ yields

$$\begin{aligned} \mathbf{C}_i(\mathbf{X}) &= \mathbf{C}_i({}^{\mathbf{s}}\mathbf{S}^\top {}^{\mathbf{t}}\mathbf{S}) = {}^{\mathbf{s}}\mathbf{S}^\top \mathbf{C}_i({}^{\mathbf{t}}\mathbf{S}) {}^{\mathbf{s}}\mathbf{S} \\ \mathbf{C}_i(\mathbf{X}^\top) &= \mathbf{C}_i({}^{\mathbf{t}}\mathbf{S}^\top {}^{\mathbf{s}}\mathbf{S}) = {}^{\mathbf{t}}\mathbf{S}^\top \mathbf{C}_i({}^{\mathbf{s}}\mathbf{S}) {}^{\mathbf{t}}\mathbf{S}, \end{aligned}$$

so

$$\begin{aligned} \mathbf{C}_i({}^{\mathbf{t}}\mathbf{S}) &= {}^{\mathbf{s}}\mathbf{S}^{\dagger\top} \mathbf{C}_i(\mathbf{X}) {}^{\mathbf{s}}\mathbf{S}^\dagger \\ \mathbf{C}_i({}^{\mathbf{s}}\mathbf{S}) &= {}^{\mathbf{t}}\mathbf{S}^{\dagger\top} \mathbf{C}_i(\mathbf{X}^\top) {}^{\mathbf{t}}\mathbf{S}^\dagger \end{aligned} \tag{6}$$

because ${}^*m \geq n$ and hence ${}^*\mathbf{S}{}^*\mathbf{S}^\dagger = \mathbf{I}$. By assumption the matrices $\mathbf{C}_i({}^*\mathbf{S})$ are as diagonal as possible. Hence we can find one of the source vectors by jointly diagonalizing either $\mathbf{C}_i(\mathbf{X})$ or $\mathbf{C}_i(\mathbf{X}^\top)$ for all i. The other source vector can then be calculated by equation (4). Of course we would only be using either temporal or spatial properties, so this corresponds to only temporal or spatial BSS, see section 3.1.

In order to include the full spatiotemporal data, we have to find diagonalizers for both $\mathbf{C}_i(\mathbf{X})$ and $\mathbf{C}_i(\mathbf{X}^\top)$ such that they satisfy the spatiotemporal model (4). As $\mathbf{X}$ (or matrices derived from it) have to be diagonalized in terms of both columns and rows, we want to call this task *double-sided approximate joint diagonalization*. This process will be reduced to the common approximate joint diagonalization in the following.

For the remainder of this section, let us assume the (unrealistic) case of ${}^{\mathbf{s}}m = {}^{\mathbf{t}}m = n$ — we will deal with the general problem in the next section. Then all matrices, which in general can be assumed to be of full rank, are now even invertible, and by model (4) we get ${}^{\mathbf{s}}\mathbf{S}^\top = \mathbf{X}{}^{\mathbf{t}}\mathbf{S}^{-1}$. Applying this to equations (6) together with an inversion of the second equation yields

$$\begin{aligned} \mathbf{C}_i({}^{\mathbf{t}}\mathbf{S}) &= {}^{\mathbf{t}}\mathbf{S}\ \mathbf{X}^\dagger \mathbf{C}_i(\mathbf{X}) \mathbf{X}^{\dagger\top}\ {}^{\mathbf{t}}\mathbf{S}^\top \\ \mathbf{C}_i({}^{\mathbf{s}}\mathbf{S})^{-1} &= {}^{\mathbf{t}}\mathbf{S}\quad\ \mathbf{C}_i(\mathbf{X}^\top)^{-1}\ {}^{\mathbf{t}}\mathbf{S}^\top. \end{aligned} \tag{7}$$

Note that we also assume that the condition matrices are invertible. So the double-sided joint diagonalization can be simply performed by jointly diagonalizing the twice as large set of matrices $\{\mathbf{X}^\dagger\mathbf{C}_i(\mathbf{X})\mathbf{X}^{\dagger\top}, \mathbf{C}_i(\mathbf{X}^\top)^{-1} \mid i = 1,\ldots\}$.

Furthermore we can now finally specify what we mean by achieving spatiotemporal BSS 'as much as possible' — we simply measure the error term of the above joint diagonalization criterion. Moreover, either spatial or temporal separation can be favored by introducing a weighting factor $\alpha \in [0,1]$. The set for approximate joint diagonalization is then defined by

$$\{\alpha\mathbf{X}^\dagger\mathbf{C}_i(\mathbf{X})\mathbf{X}^{\dagger\top}, (1-\alpha)\mathbf{C}_i(\mathbf{X}^\top)^{-1} \mid i = 1,\ldots\}. \tag{8}$$

If $\mathbf{A}$ is a diagonalizer of (8) in the sense of section 4.2, then the sources can be estimated by ${}^{\mathbf{t}}\hat{\mathbf{S}} = \mathbf{A}^{-1}$ and ${}^{\mathbf{s}}\hat{\mathbf{S}} = \mathbf{A}^\top\mathbf{X}^\top$. Joint diagonalization is usually performed by optimizing an off-diagonal criterion such as (5), so different scale factors in the matrices indeed yield different optima if the diagonalization cannot be achieved fully. According to equations (7), the higher α the more temporal separation is stressed. In the limit case $\alpha = 1$ only the temporal criterion is optimized, so temporal BSS is performed, whereas for $\alpha = 0$ a spatial BSS is calculated.

In practice, in order to be able to weight the matrix sets using α appropriately, a normalization by multiplication by a constant separately within the two sets seems to be appropriate. Only then can we guarantee equal scales of the two matrix sets. Furthermore note that we cannot assume that the diagonalizer is orthogonal, so a more general non-orthogonal joint diagonalization algorithm such as ACDC [9] has to be used.

4.4 Dimension reduction

In principle, diagonalization of the matrix set from (8) can now be used to perform spatiotemporal BSS — but only in the case of equal dimensions. Furthermore, apart from computational issues involving the high dimensionality, the BSS estimate would be very poor, simply due to the fact that in the estimates of the source condition matrices, either in $\mathbf{C}_i(\mathbf{X})$ or in $\mathbf{C}_i(\mathbf{X}^\top)$ equal or less samples than signals are available! Hence dimension reduction is essential.

Our goal is to extract only $n \ll \min\{{}^{\mathbf{s}}m, {}^{\mathbf{t}}m\}$ sources. Similar to section 3.3, we consider the singular value decomposition $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$ of $\mathbf{X}$. Permute the diagonal matrix $\mathbf{D}$ (and corresponding columns of $\mathbf{U}$ and $\mathbf{V}$) such that $\mathbf{D}$ contains the eigenvalues in decreasing order in its main diagonal. By only choosing the first n columns of $\mathbf{U}$ and $\mathbf{V}$ and the upper-left $n \times n$ submatrix of $\mathbf{D}$, we get a decomposition again denoted by $\hat{\mathbf{X}} := \mathbf{U}\mathbf{D}\mathbf{V}^\top$, which is an estimate of $\mathbf{X}$ using only the n largest eigenvalues. The matrices $\mathbf{U} \in \mathbb{R}^{{}^{\mathbf{s}}m \times n}$ and $\mathbf{V} \in \mathbb{R}^{{}^{\mathbf{t}}m \times n}$ are again pseudo-orthogonal, and $\mathbf{D}$ is diagonal. So

$$\mathbf{X} \approx \mathbf{U}\mathbf{D}\mathbf{V}^\top = \left(\mathbf{U}\mathbf{D}^{1/2}\right)\left(\mathbf{V}\mathbf{D}^{1/2}\right)^\top.$$

This is a matrix factorization of $\mathbf{X}$ into two decorrelated signals $\mathbf{U}\mathbf{D}^{1/2}$ and $\mathbf{V}\mathbf{D}^{1/2}$. After dimension reduction, the spatiotemporal

BSS model (4) can only hold approximately: $\mathbf{X} \approx \hat{\mathbf{X}} = {}^{\mathbf{s}}\mathbf{S}^\top {}^{\mathbf{t}}\mathbf{S}$ — now ${}^{\mathbf{s}}\mathbf{S}$ and ${}^{\mathbf{t}}\mathbf{S}$ are of reduced (row) size n. Plugging this model into the above equation together with the pseudo-orthogonality of $\mathbf{U}$ and $\mathbf{V}$ yields $\left(\mathbf{U}\mathbf{D}^{-1/2}\right)^\top {}^{\mathbf{s}}\mathbf{S}^\top {}^{\mathbf{t}}\mathbf{S} \left(\mathbf{V}\mathbf{D}^{-1/2}\right) = \mathbf{I}$. Hence $\mathbf{W} := {}^{\mathbf{t}}\mathbf{S}\mathbf{V}\mathbf{D}^{-1/2}$ is an invertible $n \times n$ matrix.

The first equation of (7) still holds in the more general case and we get (using $\mathbf{W}$ from above and $\mathbf{V}^\dagger = \mathbf{V}^\top$):

$$\begin{aligned}
\mathbf{C}_i({}^{\mathbf{t}}\mathbf{S}) &= {}^{\mathbf{t}}\mathbf{S}\hat{\mathbf{X}}^\dagger \mathbf{C}_i(\hat{\mathbf{X}})\hat{\mathbf{X}}^{\dagger\top} {}^{\mathbf{t}}\mathbf{S}^\top \\
&= {}^{\mathbf{t}}\mathbf{S}\mathbf{V}^{\dagger\top}\mathbf{D}^{-1}\mathbf{U}^\dagger \mathbf{C}_i(\hat{\mathbf{X}})\mathbf{U}^{\dagger\top}\mathbf{D}^{-1}\mathbf{V}^\dagger {}^{\mathbf{t}}\mathbf{S}^\top \\
&= \mathbf{W}\mathbf{C}_i\left(\mathbf{D}^{-1/2}\mathbf{U}^\dagger\hat{\mathbf{X}}\right)\mathbf{W}^\top \\
&= \mathbf{W}\mathbf{C}_i(\mathbf{D}^{1/2}\mathbf{V}^\top)\mathbf{W}^\top.
\end{aligned}$$

The second equation of (7) cannot hold for $n < {}^*m$, but we can derive a similar result from (6), where we use $\mathbf{W}^{-1} = \mathbf{D}^{-1/2}\mathbf{V}^\dagger {}^{\mathbf{t}}\mathbf{S}^\dagger = \mathbf{D}^{-1/2}\mathbf{V}^\top {}^{\mathbf{t}}\mathbf{S}^\dagger$:

$$\begin{aligned}
\mathbf{C}_i({}^{\mathbf{s}}\mathbf{S}) &= {}^{\mathbf{t}}\mathbf{S}^{\dagger\top}\mathbf{C}_i(\mathbf{X}^\top){}^{\mathbf{t}}\mathbf{S}^\dagger \\
&= {}^{\mathbf{t}}\mathbf{S}^{\dagger\top}\mathbf{V}\mathbf{D}^{1/2}\mathbf{C}_i(\mathbf{D}^{1/2}\mathbf{U}^\top)\mathbf{D}^{1/2}\mathbf{V}^\top {}^{\mathbf{t}}\mathbf{S}^\dagger \\
&= \mathbf{W}^{-\top}\mathbf{C}_i(\mathbf{D}^{1/2}\mathbf{U}^\top)\mathbf{W}^{-1}
\end{aligned}$$

which we can now invert to get $\mathbf{C}_i({}^{\mathbf{s}}\mathbf{S})^{-1} = \mathbf{W}\mathbf{C}_i(\mathbf{D}^{1/2}\mathbf{U}^\top)^{-1}\mathbf{W}^\top$.

Hence diagonality of the spatial and temporal source conditions can be easily calculated in terms of this new reduced coordinate system. The set of diagonalization matrices from equation (8) can now be rewritten as

$$\{\alpha\mathbf{C}_i(\mathbf{D}^{1/2}\mathbf{V}^\top),\ (1-\alpha)\mathbf{C}_i(\mathbf{D}^{1/2}\mathbf{U}^\top)^{-1} \mid i = 1, \ldots\} \qquad (9)$$

which can be easily calculated once the SVD of $\mathbf{X}$ is known. If $\mathbf{A}$ is a joint diagonalizer of (9), the sources are estimated by ${}^{\mathbf{t}}\hat{\mathbf{S}} = \mathbf{A}^\top\mathbf{D}^{1/2}\mathbf{V}^\top$ and ${}^{\mathbf{s}}\hat{\mathbf{S}} = \mathbf{A}^{-1}\mathbf{D}^{1/2}\mathbf{U}^\top$.

4.5 Matlab implementation

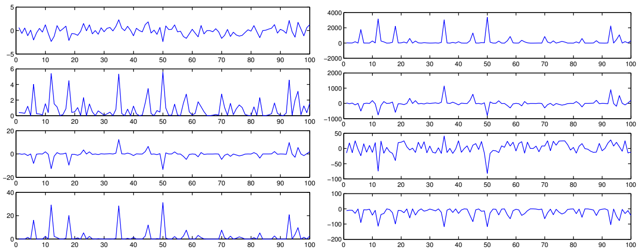
In the experiments we use the JADE-like fourth-order cumulants criterion to perform spatiotemporal ICA; we call the resulting algorithm *spatiotemporal JADE (stJADE)* for short. Our software package, available at `http://fabian.theis.name/` implements all the details of stJADE. The package contains all the files needed to reproduce the results described in this paper.

5. SIMULATIONS

We present the performance of stJADE on a toy example. Consider $n = 4$ temporal sources ${}^{\mathbf{t}}\mathbf{S}$ with ${}^{\mathbf{t}}m = 100$ samples, each drawn uniformly from $[-1, 1]$. Furthermore, let n spatial sources ${}^{\mathbf{s}}\mathbf{S}$, again with ${}^{\mathbf{s}}m = 100$ be constructed as follows: let $v(r)$ be ${}^{\mathbf{s}}m$ samples of a normal distribution. Then set ${}^{\mathbf{s}}S_{ir} := v(r)^i$. Finally set $\mathbf{X} := {}^{\mathbf{s}}\mathbf{S}^\top {}^{\mathbf{t}}\mathbf{S}$ according to the spatiotemporal BSS model. So the temporal sources are fully independent, whereas the spatial sources are strongly dependent.

The stJADE algorithm is applied with $\alpha = 0.5$ and orthogonal matrix recovery. Figure 1 shows the spatial sources together with the recoveries using stJADE. The algorithm is able to recover the (independent) temporal sources well with a mean signal-to-noise ratio (SNR) of 13.3 dB. Due to the strong spatial dependencies, it finds only 3 of the 4 spatial sources. If we vary the weighting, we get similar results (mean SNR of 13.8 dB for temporal recovery) when using temporal structure only ($\alpha = 1$), and worse results (mean SNR of 8.9 dB for temporal recovery) when performing spatial separation ($\alpha = 0$). This is to be expected due to the broken spatial diagonality of the cumulants.

For comparison, we also apply Stone's spatiotemporal infomax algorithm [6]. It is unable to detect the temporal sources (mean SNR of −2.5 dB). However, it partially recovers two of the spatial sources, but these have high SNR at two of the original sources, not only one. We note that these results are somewhat difficult to judge due to the many parameters involved in Stone's algorithm.



(a) spatial sources ${}^{\mathbf{s}}\mathbf{S}$ (b) recovered ${}^{\mathbf{s}}\hat{\mathbf{S}}$ using stJADE

Figure 1: stJADE toy example. (a) shows the original dependent spatial sources, (b) the recoveries using stJADE. It is able to estimate 3 of the 4 spatial sources well with SNRs of 36, 14 and 21 dB respectively.

6. CONCLUSION

We have proposed a novel spatiotemporal BSS algorithm. It is based on the double-sided joint diagonalization as generalization of the often applied 'single-sided' joint diagonalization in temporal-only BSS. The algorithm can be applied to a whole set of source conditions; in the simulations, we use fourth-order cumulants and hence a spatiotemporal version of JADE to separate signals, thereby outperforming Stone's spatiotemporal infomax considerably. Preliminary results for fMRI data sets are promising, and in future works, we will present more extensive studies of such data along with comparisons of various source conditions.

Acknowledgements

The authors gratefully acknowledge partial financial support by the DFG (GRK 638) and the BMBF (project 'ModKog').

REFERENCES

[1] A. Belouchrani, K. Abed Meraim, J.-F. Cardoso, and E. Moulines. A blind source separation technique based on second order statistics. *IEEE Transactions on Signal Processing*, 45(2):434–444, 1997.

[2] J.-F. Cardoso and A. Souloumiac. Blind beamforming for non gaussian signals. *IEE Proceedings - F*, 140(6):362–370, 1993.

[3] A. Cichocki and S. Amari. *Adaptive blind signal and image processing*. John Wiley & Sons, 2002.

[4] P. Comon. Independent component analysis - a new concept? *Signal Processing*, 36:287–314, 1994.

[5] E. Moreau. A generalization of joint-diagonalization criteria for source separation. *IEEE Transactions on Signal Processin*, 49(3):530–541, 2001.

[6] J.V. Stone, J. Porrill, N.R. Porter, and I.W. Wilkinson. Spatiotemporal independent component analysis of event-related fmri data using skewed probability density functions. *NeuroImage*, 15(2):407–421, 2002.

[7] F.J. Theis. A new concept for separability problems in blind source separation. *Neural Computation*, 16:1827–1850, 2004.

[8] F.J. Theis, A. Meyer-Bäse, and E.W. Lang. Second-order blind source separation based on multi-dimensional autocovariances. In *Proc. ICA 2004*, volume 3195 of *Lecture Notes in Computer Science*, pages 726–733, Granada, Spain, 2004.

[9] A. Yeredor. Non-orthogonal joint diagonalization in the least-squares sense with application in blind source separation. *IEEE Trans. Signal Processing*, 50(7):15451553, 2002.