Neurocomputing 113 (2013) 153–167

Contents lists available at SciVerse ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Finding dependent and independent components from related data sets: A generalized canonical correlation analysis based method

Juha Karhunen*, Tele Hao, Jarkko Ylipaavalniemi

Department of Information and Computer Science, Aalto University School of Science, P.O. Box 15400, FI-00076 Aalto, Espoo, Finland

ARTICLE INFO

Article history:
Received 24 April 2012
Received in revised form
8 January 2013
Accepted 15 January 2013
Communicated by E.W. Lang
Available online 26 February 2013

Keywords:
Independent component analysis
Blind source separation
Canonical correlation analysis
Functional magnetic resonance imaging

ABSTRACT

In this paper, we consider an extension of independent component analysis (ICA) and blind source separation (BSS) techniques to several related data sets. The goal is to separate mutually dependent and independent components or source signals from these data sets. This problem is important in practice, because such data sets are common in real-world applications. We propose a new method which first uses a generalization of standard canonical correlation analysis (CCA) for detecting subspaces of independent and dependent components. For two data sets, this reduces to using standard CCA. Any ICA or BSS method can then be used for final separation of these components. The proposed method performs well for difficult synthetic data sets containing different types of source signals. It provides interesting and meaningful results for real-world robot grasping data and functional magnetic resonance imaging (fMRI) data. The method is straightforward to implement and computationally not too demanding. The proposed method clearly improves the separation results of several well-known ICA and BSS methods compared with the situation in which CCA or generalized CCA is not used. Not only are the signal-to-noise ratios of the separated sources often clearly higher, but our method also helps these ICA and BSS methods to separate sources that they alone cannot separate.

1. Introduction

Independent component analysis (ICA) [1–4] is a well-known technique for finding independent components or source signals in a blind (unsupervised) manner from data. While basic ICA still assumes a simple linear data model, its proper estimation requires higher-order statistics either directly or via nonlinearities. ICA provides for practical data sets often much more meaningful results (components) than standard linear techniques based on second-order statistics such as principal component analysis (PCA). In addition to ICA there exist several other techniques for blind source separation (BSS) which make somewhat different assumptions compared with ICA. These techniques will be discussed briefly in the next subsection.

ICA and BSS techniques have already many applications in different areas, and they have been extended into many directions [1–4]. In this paper, we consider an extension of ICA and BSS techniques to several related data sets. The goal is to separate mutually dependent and independent components or source signals from these data sets. This problem is important in practice, because such data sets are common in real-world applications. We propose a new method which first uses a generalization of standard canonical correlation analysis (CCA) for detecting subspaces of independent and dependent components. For two data sets, this reduces to using standard CCA. Any ICA or BSS method can then be used for final separation of these components. The proposed method performs well for difficult synthetic data sets containing different types of source signals. It provides interesting and meaningful results for real-world robot grasping data and functional magnetic resonance imaging (fMRI) data. The method is straightforward to implement and computationally not too demanding. The proposed method clearly improves the separation results of several well-known ICA and BSS methods compared with the situation in which CCA or generalized CCA is not used. Not only are the signal-to-noise ratios of the separated sources often clearly higher, but our method also helps these ICA and BSS methods to separate sources that they alone cannot separate.

1.1. Various ICA and BSS techniques

In this subsection, we briefly discuss standard ICA and some alternative blind source separation techniques. Methods belonging to these categories are later on used in our experiments for post-processing the intermediate results given CCA and its generalization.

* Corresponding author. Tel.: +358 400 817 276; fax: +358 9 4702 3277.
E-mail address: juha.karhunen@aalto.fi (J. Karhunen).
URL: http://research.ics.tkk.fi/ica/ (J. Karhunen).

0925-2312/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2013.01.018

The data model used in standard linear ICA is simply

$$\mathbf{x}(n) = \mathbf{A}\mathbf{s}(n) = \sum_{i=1}^{m} s_i(n)\mathbf{a}_i \tag{1}$$

Thus each data vector $\mathbf{x}(n)$ is expressed as a linear combination of independent components or source signals $s_i(n)$, $i = 1,2,\ldots,m$, which multiply the respective constant basis vectors $\mathbf{a}_i$. The source vector $\mathbf{s}(n) = [s_1(n),s_2(n),\ldots,s_m(n)]^T$ contains the m source signals corresponding to the n:th data vector $\mathbf{x}(n)$, and the mixing matrix $\mathbf{A} = [\mathbf{a}_1,\mathbf{a}_2,\ldots,\mathbf{a}_m]$ the basis vectors $\mathbf{a}_i$. They are in general linearly independent but non-orthogonal. They depend on the available data set $\mathbf{X} = [\mathbf{x}(1),\ldots,\mathbf{x}(N_x)]$ where N_x is the number of data vectors and $n = 1,2,\ldots,N_x$, but once they have been estimated, they are the same for all the data vectors in $\mathbf{X}$. The index n may denote discrete time, position (especially in digital images), or just the number of the sample vector. For simplicity, we assume here that both the data vector $\mathbf{x}(n) = [x_1(n),x_2(n),\ldots,x_m(n)]^T$ and the source vector $\mathbf{s}(n)$ are zero mean m-vectors, and that the mixing matrix $\mathbf{A}$ is a full-rank constant $m \times m$ matrix.

In standard linear ICA, the index n of the data vectors and source signals is not important, because the data vectors $\mathbf{x}(n)$ can be processed in arbitrary order and the results of ICA still remain the same. This is the situation if the data vectors are just samples from some multivariate statistical distribution. However, the data vectors $\mathbf{x}(n)$ are often subsequent equispaced samples from a vector-valued time series which is temporally correlated (non-white). Then the index n corresponds to discrete time instant t_n. In this case, the index n is important, because subsequent data vectors $\ldots,\mathbf{x}(n-1),\mathbf{x}(n),\mathbf{x}(n+1),\ldots$ contain additional temporal information which should be utilized for getting optimal results. Standard ICA can be applied to such time series, too, but it is suboptimal because it does not utilize this additional information. Alternative methods have been developed for extracting the source signals or independent components in such cases. They usually utilize either temporal autocorrelations directly, forming another important group of BSS methods, or in the third major group of methods smoothly changing nonstationarity of variance; see for example [1–3,5].

The application domains and assumptions made in these three major groups of BSS technique are different [1,5]. In standard ICA, it is assumed that all the independent components except for possibly one have non-Gaussian distributions and are mutually statistically independent [1,6]. Then standard ICA methods are able to separate their waveforms, leaving however the order, sign, and scaling of the separated components ambiguous. The scaling indeterminacy is usually fixed by normalizing the variances of the separated independent components to unity. The most widely used standard ICA method is currently FastICA [1,7] because of its efficient implementation and fast convergence. Therefore, it can be applied to higher dimensional ICA problems, too. We have used in our experiments the freely downloadable FastICA Matlab software package [8]. Another popular ICA method is the adaptive neural natural gradient method [1,2], which however converges slowly and requires knowledge or estimation of the type of the source signals or independent components. Super-Gaussian and sub-Gaussian sources require different nonlinearities in this method.

On the other hand, methods based on temporal autocorrelations of the source signals require that different sources have at least some different non-zero autocorrelations which they use. Contrary to standard ICA, they can then separate even Gaussian sources, but on the other hand, they fail if such temporal correlations do not exist, while standard ICA can even in this case separate non-Gaussian sources. Examples of methods based on temporal autocorrelations are the SOBI method [9] and the TDSEP method [10]. A recent review of such methods containing many more references is [11].

In the third group of BSS methods, it is assumed that the source signals have nonstationary smoothly changing variances. Such methods have been introduced in [12,13]. If the assumptions made in them are valid, they can separate even Gaussian temporally uncorrelated (white) sources that ICA and temporal autocorrelation methods are not able to handle appropriately. A fourth class of BSS methods employs time–frequency representations (see in [3, Chapter 11]), but we shall not discuss them in this paper.

Some attempts have been made to combine different types of BSS methods so that they would be able to separate wider classes of source signals. In particular, methods taking into account both non-Gaussianity used in ICA and temporal correlations have been considered by several authors. Such methods are the JADE$_{TD}$ method introduced in [14], the complexity pursuit method [15], as well as the joint cumulant and correlation-based separation method in [16], and thinICA [17]. Recently, Tichavsky et al. [18] introduced a new technique of this type which they claim to outperform other such methods at least in their simulations.

In [19], Hyvärinen developed an approximate method which tries to utilize both higher-order statistics, temporal autocorrelations, and nonstationarity of variances. This method seems to be able to separate different types of sources, but it is limited by the facts that it is approximative and uses only the autocorrelation coefficient corresponding to a single time lag equal to 1, as pointed out in [18]. We have used this method, called UniBSS in its Matlab code [20], in addition to FastICA in our experiments.

1.2. An outline of our method

In this paper, we consider a generalization in which one tries to find out mutually dependent and independent components from different but related data sets. Considering first two such data sets, data vectors $\mathbf{y}(n)$ of dimension m_y belonging to the related data set $\mathbf{Y} = [\mathbf{y}(1),\ldots,\mathbf{y}(N_y)]$ are assumed to obey a similar basic linear ICA data model

$$\mathbf{y}(n) = \mathbf{B}\mathbf{r}(n) = \sum_{i=1}^{m_y} r_i(n)\mathbf{b}_i \tag{2}$$

as the data vectors $\mathbf{x}(n)$ in Eq. (1). The assumptions that we make on the m_y-dimensional basis vectors $\mathbf{b}_i$ and source signals $r_i(n)$ are exactly the same as those made on the basis vectors $\mathbf{a}_i$ and source signals $s_i(n)$ in context with Eq. (1). More generally, we have M such data sets $\mathbf{X}_1,\mathbf{X}_2,\ldots,\mathbf{X}_M$. The dimensionalities m_i of the data vectors belonging to these data sets can be different, but the number of data vectors N in them must be the same for CCA and its generalizations. If this is not the case, obviously we select N equal to the minimum number of data vectors in these data sets. The respective data vectors in each data set should also correspond to each other, for example being taken at the same time instant.

CCA, explained mathematically in the next Section 2, is an old technique [21] which uses second-order statistics for measuring the linear relationships (correlations) between two data sets. However, it has been recently applied by several authors to different real-world data analysis and signal processing problems. This is because CCA often performs quite well in practice, and using higher-order statistics and nonlinear techniques do not necessarily improve the results markedly.

In our method, we first apply a generalization of CCA to find subspaces of dependent and independent sources in the data sets $\mathbf{X}_1,\mathbf{X}_2,\ldots,\mathbf{X}_M$. The data sets are then projected onto these subspaces. After this, any suitable ICA or BSS method can be used

for final separation. Our method is described in more detail in Section 3.

1.3. Related work

The extension of ICA and BSS for separating dependent and independent source signals from two related data sets has not been studied as much as many other extensions of ICA and BSS mentioned above, but some research on this topic has been carried out.

The first author of this paper tried to generalize cross-correlation analysis based on singular value decomposition in ICA style to take into account higher-order statistics in [22]. In this paper, we modify that method so that its performance is clearly improved, and a theoretical weakness of this earlier method vanishes.

In [23], Ylipaavalniemi et al. have carried out their analysis of biomedical fMRI sources in reverse order compared with our method. They first apply standard ICA to the two related data sets separately. Then they connect dependent sources (independent components) in these data sets using CCA. The method performs pretty well but it has a theoretical weakness: ICA assumes that the sources are non-Gaussian but CCA can be derived from a probabilistic latent variable model where all the involved random variables (vectors) are Gaussian [24]. The authors of the paper [23] have themselves noticed this theoretical weakness and improved their method in two later papers. In [25], they apply to the results first provided by ICA a nonparametric CCA type model where Gaussian distributions are not assumed, getting improved results. In another more theoretical paper [26] the authors show on a general level how to apply a probabilistic CCA type model without assuming Gaussian distributions, using instead of them any noise model belonging to the exponential family of probability distributions.

In [27], the authors use standard CCA and its extension to multiple data sets for the analysis of medical imaging data, discussing the advantages of such approaches and comparing their performances to standard ICA that has been successfully applied to this type of problems. This tutorial review is largely based on the research papers [28,29].

Koetsier et al. have presented in [30] an unsupervised neural algorithm called exploratory correlation analysis for the extraction of common features in multiple data sources. This method is closely related with canonical correlation analysis. In an earlier paper [31] Lai and Fyfe extended their neural implementation of CCA to nonlinear and kernel CCA with application to blind source separation.

Gutmann and Hyvärinen [32] have recently introduced a method based on nonstationary variances for finding dependent sources from related data sets. Their method as well as most other methods assume that in each of these data sets there is one source signal that is dependent on one source signal in the other data sets, while these sources are independent of all other sources.

Akaho et al. [33] have considered an ICA style generalization of canonical correlation analysis which they call multimodal independent component analysis. In their method, standard linear ICA is first applied to both data sets $\mathbf{x}$ and $\mathbf{y}$ separately. Then the corresponding dependent components of the two ICA expansions are identified using a natural gradient type learning rule.

2. Canonical correlation analysis (CCA)

CCA [34,35] measures the linear relationships between two multidimensional data sets $\mathbf{X}$ and $\mathbf{Y}$ using their second-order statistics, that is, autocovariances and cross-covariances. It finds two bases, one for both $\mathbf{X}$ and $\mathbf{Y}$, that are optimal with respect to correlations and it also finds the corresponding correlations. In other words, CCA finds the two bases in which the cross-correlation matrix between the data sets $\mathbf{X}$ and $\mathbf{Y}$ becomes diagonal and the correlations of the diagonal are maximized. In CCA we need not assume that the data vectors $\mathbf{x} \in \mathbf{X}$ and $\mathbf{y} \in \mathbf{Y}$ obey the linear models (1) and (2), respectively (even though they can always be represented by using such linear expansions). The dimensions m_x and m_y of the respective vectors $\mathbf{x}$ and $\mathbf{y}$ can be different, but they are assumed to have zero means. Furthermore, the data sets $\mathbf{X}$ and $\mathbf{Y}$ must have equal number of vectors. An important property of canonical correlations is that they are invariant to affine transformations of the variables, which does not hold for ordinary correlation analysis [35].

Consider first the case where only one pair of basis vectors is sought, namely the ones corresponding to the largest canonical correlation. For this, consider the linear combinations $x = \mathbf{x}^T\mathbf{w_x}$ and $y = \mathbf{y}^T\mathbf{w_y}$ of the random vectors $\mathbf{x} \in \mathbf{X}$ and $\mathbf{y} \in \mathbf{Y}$. The function to be maximized in CCA is the normalized correlation coefficient ρ between these two projections

$$\rho = \frac{\mathrm{E}\{xy\}}{\sqrt{\mathrm{E}\{x^2\}\mathrm{E}\{y^2\}}} = \frac{\mathbf{w}_\mathbf{x}^T\mathbf{C_{xy}}\mathbf{w_y}}{\sqrt{\mathbf{w}_\mathbf{x}^T\mathbf{C_{xx}}\mathbf{w_x}\mathbf{w}_\mathbf{y}^T\mathbf{C_{yy}}\mathbf{w_y}}} \tag{3}$$

where $\mathbf{C_{xy}} = E\{\mathbf{x}\mathbf{y}^T\}$ is the cross-covariance matrix of $\mathbf{x}$ and $\mathbf{y}$, and $\mathbf{C_{xx}} = E\{\mathbf{x}\mathbf{x}^T\}$ as well as $\mathbf{C_{yy}} = E\{\mathbf{y}\mathbf{y}^T\}$ are their autocovariance matrices. The maximum of ρ with respect to the weight vectors $\mathbf{w_x}$ and $\mathbf{w_y}$ defines the maximum canonical correlation.

The i:th canonical correlation is defined for $\mathbf{x}$ by the weight vector $\mathbf{w}_{xi} : x_i = \mathbf{x}^T\mathbf{w}_{xi}$, and for $\mathbf{y}$ by $\mathbf{w}_{yi} : y_i = \mathbf{y}^T\mathbf{w}_{yi}$. Different canonical correlations are uncorrelated: $E\{x_ix_j\} = E\{y_iy_j\} = E\{x_iy_j\} = 0$. It turns out that these canonical correlations can be computed by solving the eigenvector [34,35]

$$\mathbf{C_{xx}^{-1}}\mathbf{C_{xy}}\mathbf{C_{yy}^{-1}}\mathbf{C_{yx}}\mathbf{w_x} = \rho^2\mathbf{w_x}$$

$$\mathbf{C_{yy}^{-1}}\mathbf{C_{yx}}\mathbf{C_{xx}^{-1}}\mathbf{C_{xy}}\mathbf{w_y} = \rho^2\mathbf{w_y} \tag{4}$$

where $\mathbf{C_{yx}} = E\{\mathbf{y}\mathbf{x}^T\}$. The eigenvalues ρ^2 are squared canonical correlations and the eigenvectors $\mathbf{w_x}$ and $\mathbf{w_y}$ are normalized canonical correlation basis vectors. Only non-zero solutions to these equations are usually of interest, and their number is equal to at most the smaller one of the dimensions of the vectors $\mathbf{x}$ and $\mathbf{y}$.

Eq. (4) becomes simpler if the data vectors $\mathbf{x}$ and $\mathbf{y}$ are first whitened [1], which is the usual practice in many ICA algorithms, for example in FastICA. After prewhitening, both $\mathbf{C_{xx}}$ and $\mathbf{C_{yy}}$ become unit matrices, and noting that $\mathbf{C_{yx}} = \mathbf{C}_\mathbf{xy}^T$ Eq. (4) reduces to

$$\mathbf{C_{xy}}\mathbf{C}_\mathbf{xy}^T\mathbf{w_x} = \rho^2\mathbf{w_x}$$

$$\mathbf{C_{yx}}\mathbf{C}_\mathbf{yx}^T\mathbf{w_y} = \rho^2\mathbf{w_y} \tag{5}$$

But these are just the defining equations for the singular value decomposition (SVD) [36] of the cross-covariance matrix $\mathbf{C_{xy}}$

$$\mathbf{C_{xy}} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T = \sum_{i=1}^{L} \rho_i\mathbf{u}_i\mathbf{v}_i^T \tag{6}$$

There $\mathbf{U}$ and $\mathbf{V}$ are orthogonal square matrices ($\mathbf{U}^T\mathbf{U} = \mathbf{I}$, $\mathbf{V}^T\mathbf{V} = \mathbf{I}$) containing the singular vectors $\mathbf{u}_i$ and $\mathbf{v}_i$. In our case, these singular vectors are the basis vectors providing canonical correlations. In general, the dimensions of the matrices $\mathbf{U}$ and $\mathbf{V}$ are respectively $m_x \times m_x$ and $m_y \times m_y$. Consequently, the dimensions m_x and m_y of the singular vectors $\mathbf{u}_i$ and $\mathbf{v}_i$ are also generally different, and the same as the dimensions of the data vectors $\mathbf{x}$ and $\mathbf{y}$, respectively. The pseudodiagonal $m_x \times m_y$ matrix

$$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{D} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \tag{7}$$

consists of a diagonal matrix $\mathbf{D}$ containing the non-zero singular values appended with zero matrices so that the matrix $\mathbf{\Sigma}$ is compatible with the different dimensions of $\mathbf{x}$ and $\mathbf{y}$. These non-zero singular values are just the non-zero canonical correlations. If the cross-covariance matrix $\mathbf{C_{xy}}$ has full rank, their number L is the smaller one of the dimensions m_x and m_y of the data vectors $\mathbf{x}$ and $\mathbf{y}$.

3. Our method for two related data sets

We first preprocess the data vectors $\mathbf{x} \in \mathbf{X}$ and $\mathbf{y} \in \mathbf{Y}$ following the models (1) and (2) by subtracting their mean vectors from them if they are non-zero. After this, these data vectors are whitened separately

$$\mathbf{v_x} = \mathbf{V_x x}, \quad \mathbf{v_y} = \mathbf{V_y y} \tag{8}$$

The whitening matrices $\mathbf{V_x}$ and $\mathbf{V_y}$ should not be confused with the singular vector matrix $\mathbf{V}$ in (6). Whitening can be carried out in many ways [1,2], typically standard principal component analysis (PCA) is used to that end. That is, whitening is based on the eigendecompositions of the autocovariance matrices $\mathbf{C_{xx}}$ and $\mathbf{C_{yy}}$. The whitening matrix for $\mathbf{x}$ is then

$$\mathbf{V_x} = \mathbf{\Lambda}^{-1/2}\mathbf{E} \tag{9}$$

where the columns of the matrix $\mathbf{E}$ contain the eigenvectors of $\mathbf{C_{xx}}$, and the diagonal matrix $\mathbf{\Lambda}$ contains the respective eigenvalues in the same order. The whitening matrix $\mathbf{V_y}$ for $\mathbf{y}$ is computed quite similarly using the eigendecomposition of $\mathbf{C_{yy}}$. After whitening, the cross-covariances (cross-correlations) of different components of the whitened data vectors $\mathbf{v_x}$ and $\mathbf{v_y}$ are zero, while their variances equal to 1. Thus whitening normalizes the data with respect to its second-order statistics. When PCA in Eq. (9) is used for whitening, it is also possible to compress the dimensionality of the data and possibly filter out some noise by retaining in $\mathbf{\Lambda}$ only the largest PCA eigenvalues and in $\mathbf{E}$ the corresponding principal eigenvectors, but we have not used this option.

After whitening, we estimate the cross-covariance matrix $\mathbf{C_{v_x v_y}}$ of the whitened data vectors $\mathbf{v_x}$ and $\mathbf{v_y}$ in standard manner

$$\widehat{\mathbf{C}}_{\mathbf{v_x v_y}} = \frac{1}{N}\sum_{t=1}^{N} \mathbf{v_x}(t)\mathbf{v_y}^T(t) \tag{10}$$

There N is the smaller of the numbers N_x and N_y of the data vectors in the two data sets $\mathbf{X}$ and $\mathbf{Y}$, respectively.

We then perform singular value decomposition of the estimated cross-covariance matrix $\widehat{\mathbf{C}}_{\mathbf{v_x v_y}}$ quite similarly as for $\mathbf{C_{xy}}$ in Eq. (6). Inspecting the magnitudes of the singular values $\hat{\rho}_i$ in the pseudodiagonal matrix $\mathbf{\Sigma}$, we then divide the matrices $\mathbf{U}$ and $\mathbf{V}$ of singular vectors into two submatrices

$$\mathbf{U} = [\mathbf{U}_1 \ \mathbf{U}_2], \quad \mathbf{V} = [\mathbf{V}_1 \ \mathbf{V}_2] \tag{11}$$

There $\mathbf{U}_1$ and $\mathbf{V}_1$ correspond to dependent components for which the respective singular values are larger than the chosen threshold value κ, and $\mathbf{U}_2$ and $\mathbf{V}_2$ to the independent components for which the respective singular values are smaller than κ. Due to the whitening, all the singular values $\hat{\rho}_i$ of the matrix (10) lie between 0 and 1. In most of our experiments the threshold value $\kappa = 0.5$ was found to be suitable. The data are then projected using these submatrices into subspaces corresponding to the dependent and independent components by computing

$$\mathbf{U}_1^T\mathbf{X}, \quad \mathbf{U}_2^T\mathbf{X}, \quad \mathbf{V}_1^T\mathbf{Y}, \quad \mathbf{V}_2^T\mathbf{Y} \tag{12}$$

where $\mathbf{X} = [\mathbf{x}(1), \ldots, \mathbf{x}(N_x)]$ and $\mathbf{Y} = [\mathbf{y}(1), \ldots, \mathbf{y}(N_y)]$.

Finally, we apply any suitable ICA or BSS method separately to each of these 4 projected data sets for separating the source signals contained in these subspaces. It should be noted that we include in the submatrices $\mathbf{U}_2$ and $\mathbf{V}_2$ also the singular vectors corresponding to small or even zero singular values for being able to separate all the sources in $\mathbf{X}$ and $\mathbf{Y}$.

In the following, we present several somewhat intuitive and heuristic justifications to the proposed method which anyway in our opinion should largely explain its good performance. First, let us denote the separating matrices after the whitening step in (8) by $\mathbf{W}_\mathbf{x}^T$ for $\mathbf{v_x}$ and respectively by $\mathbf{W}_\mathbf{y}^T$ for $\mathbf{v_y}$. A basic result in the theory of ICA and BSS is that after whitening the separating matrices $\mathbf{W_x}$ and $\mathbf{W_y}$ become orthogonal: $\mathbf{W}_\mathbf{x}^T\mathbf{W_x} = \mathbf{I}$, $\mathbf{W}_\mathbf{y}^T\mathbf{W_y} = \mathbf{I}$ [1]. Thus

$$\widehat{\mathbf{s}} = \mathbf{W}_\mathbf{x}^T\mathbf{V_x x} = \mathbf{W}_\mathbf{x}^T\mathbf{V_x A s} = \mathbf{P_s D_s s} \tag{13}$$

The vector $\widehat{\mathbf{s}}$ on the left hand side contains the estimated sources. A basic ambiguity in the blind ICA and BSS methods is that they can appear in different order and have different scales than the original sources [1]. This has been taken into account in Eq. (13) by multiplying the source vector $\mathbf{s}$ on the right-hand side by a diagonal scaling matrix $\mathbf{D_s}$ and a permutation matrix $\mathbf{P_s}$, which changes the order of the elements in the column vector $\mathbf{D_s s}$ [37].

Assuming that there are as many linearly independent mixtures $\mathbf{x}$ and $\mathbf{W_y}$ as sources $\mathbf{s}$, so that the mixing matrix $\mathbf{A}$ is a full-rank square matrix, we get from the two last equations of (13)

$$\mathbf{A} = (\mathbf{W}_\mathbf{x}^T\mathbf{V_x})^{-1}\mathbf{P_s D_s} = \mathbf{V_x}^{-1}\mathbf{W_x P_s D_s} \tag{14}$$

due to the orthogonality of the matrix $\mathbf{W_x}$. Quite similarly, we get for the another mixing matrix $\mathbf{B}$ in (2)

$$\mathbf{B} = (\mathbf{W}_\mathbf{y}^T\mathbf{V_y})^{-1}\mathbf{P_r D_r} = \mathbf{V_y}^{-1}\mathbf{W_y P_r D_r} \tag{15}$$

where $\mathbf{D_r}$ is the diagonal scaling matrix and $\mathbf{P_r}$ the permutation matrix associated to the estimate $\widehat{\mathbf{r}}$ of the source vector $\mathbf{r}$.

Consider now the cross-covariance matrix after whitening. It is

$$\mathbf{C_{v_x v_y}} = \mathbf{V_x}\mathrm{E}\{\mathbf{x}\mathbf{y}^T\}\mathbf{V}_\mathbf{y}^T = \mathbf{V_x A Q B}^T\mathbf{V}_\mathbf{y}^T \tag{16}$$

Here the matrix $\mathbf{Q} = E\{\mathbf{s}\mathbf{r}^T\}$ is a diagonal matrix, if the sources signals in the source vectors $\mathbf{s}$ and $\mathbf{r}$ are pairwise dependent but otherwise independent of each other. Inserting $\mathbf{A}$ and $\mathbf{B}$ from Eqs. (14) and (15) into (16) yields

$$\mathbf{C_{v_x v_y}} = (\mathbf{W_x P_s})(\mathbf{D_s Q D}_\mathbf{r}^T)(\mathbf{W_y P_r})^T \tag{17}$$

But this is exactly the same type of expansion as the singular value decomposition (6) of the whitened cross-covariance matrix $\mathbf{C_{v_x v_y}}$. First, $\mathbf{W_x P_s}$ is a product of an orthogonal matrix $\mathbf{W_x}$ and permutation matrix $\mathbf{P_s}$, which here changes the order of the columns in the matrix $\mathbf{W_x}$ [37]. Thus $\mathbf{W_x P_s}$ is still an orthogonal matrix having the same column vectors as $\mathbf{W_x}$ but generally in different order. The matrix $\mathbf{W_x P_s}$ corresponds to the orthogonal matrix $\mathbf{U}$ in (6), and quite similarly the orthogonal matrix $\mathbf{W_y P_r}$ corresponds to the orthogonal matrix $\mathbf{V}$ in (6). Finally, the matrix $\mathbf{D_s Q D}_\mathbf{r}^T$ is a product of three diagonal matrices and hence a diagonal matrix which corresponds to the diagonal matrix $\mathbf{\Sigma}$ in (6).

Thus on the assumptions made above the SVD of the whitened cross-covariance matrix provides a solution that has the same structure as the separating solution. Even though we cannot from this result directly deduce that the SVD of the whitened cross-covariance matrix (that is, CCA) would provide a separating solution, this seems to hold in simple cases at least as shown by our experiments. At least CCA when applied to the data sets $\mathbf{X}$ and $\mathbf{Y}$ using (12) provides already partial separation, helping several ICA or BSS methods to achieve clearly better results in difficult cases.

Another justification is that CCA, or SVD of whitened data vectors, uses second-order statistics (cross-covariances) only for separation, while standard ICA algorithms such as FastICA use for

separation higher-order statistics only after the data has been normalized with respect to their second-order statistics by whitening them. Combining both second-order statistics and higher-order statistics by first performing CCA and then post-processing the results using a suitable ICA or BSS method can be expected to provide better results than using solely second-order or higher-statistics only for separation.

Our third justification is that dividing the separation problem into subproblems using the matrices in (12) probably helps. Solving two lower dimensional subproblems is usually easier than solving a higher dimensional separation problem. And if the mixtures are difficult to separate, consisting of several types of sources which could be super-Gaussian, sub-Gaussian, Gaussian, temporally correlated, or nonstationary sources, the complexity of the sources and mixtures in the subproblems to be solved after CCA may be reduced.

We can modify the SVD based method introduced above to include higher-order statistics via nonlinearities by using instead of the plain cross-covariance matrix $\mathbf{C}_{\mathbf{v_x v_y}} = E\{\mathbf{v_x} \mathbf{v_y}^T\}$ a generalized cross-covariance matrix of the type

$$\mathbf{G}_{\mathbf{v_x v_y}} = \mathrm{E}\{\mathbf{f}(\mathbf{v_x})\mathbf{v_y}^T + \mathbf{v_x}\mathbf{f}(\mathbf{v_y}^T)\} \tag{18}$$

where $\mathbf{f}(\mathbf{z})$ is a suitably chosen nonlinearity applied componentwise to its argument vector $\mathbf{z}$; we have tried $\mathbf{f}(\mathbf{z}) = \tanh(\mathbf{z})$ (suitably scaled). However, this modification did not have any noticeable effect on the results in our experiments. The reason is probably that ICA methods such as FastICA already use higher-order statistics and nonlinearities for final separation.

But including in a similar manner temporal correlations into the computations by using the generalized cross-covariance matrix

$$\mathbf{G}_{\mathbf{v_x v_y}} = \sum_{i=1}^{K} \mathrm{E}\{\mathbf{v_x}(n)\mathbf{v_y}^T(n-d_i) + \mathbf{v_x}(n)\mathbf{v_y}^T(n+d_i)\} \tag{19}$$

where d_i, $i = 1, \ldots, K$ are the chosen time delays, can sometimes improve the separation results quite a lot. Recall that here n corresponds to the discrete time instant t_n, and $n-d_i$ to the discrete time instant t_{n-d_i} which is d_i equispaced samples before to the sample n. In fact, in this way we can apply our method even to a single data set $\mathbf{X}$. The other related data set $\mathbf{Y}$ which we originally do not have at our disposal is created by time delaying the data set $\mathbf{X}: \mathbf{Y}(n) = \mathbf{X}(n-d_i)$ where the time delay d_i can be positive or negative. In our experiments it turned out that better results can be obtained by using two different time delays $d_1 = 1$ and $d_2 = 2$ in (19). Thus the original data set was correlated with four time delayed data sets $\mathbf{X}(n-2)$, $\mathbf{X}(n-1)$, $\mathbf{X}(n+1)$, and $\mathbf{X}(n+2)$.

This method for a single data set is related with the method proposed by Friman et al. [38]. They have successfully applied CCA to blind separation of sources in a single data set $\mathbf{X}(n)$ by using in CCA $\mathbf{X}(n)$ and $\mathbf{X}(n-1)$, even though in their method there is no division into subspaces of dependent and independent sources as in our method. Their method has been analyzed theoretically in [39] where it is proved that it separates source signals successfully.

4. Extension to several data sets

In a pioneering paper [40], Kettenring introduced and discussed five different generalizations of standard CCA to three or more data sets, albeit only two of them were completely new. These generalizations are based on somewhat different optimization criteria and orthogonality constraints, but seem in practical experiments to yield pretty similar results. The most popular of these criteria is so-called maximum variance generalization of CCA [40,41]. It can be optimized and the respective canonical vectors estimated using the procedure described in [40,41]. This optimization method is, however, computationally somewhat complicated. It first requires computation of the singular value decompositions of all the M data sets $\mathbf{X}_k$, $k = 1, \ldots, M$. From them, an $L \times L$ matrix is formed where

$$L = \sum_{k=1}^{M} m_k \tag{20}$$

is the sum of the dimensionalities m_k of the data vectors in the sets $\mathbf{X}_k$, $k = 1, \ldots, M$. The desired generalized canonical vectors are then computed from the eigenvectors of this $L \times L$ matrix.

We do not discuss this procedure in more detail because an easier solution is available. Via, Santamaria, and Perez have considered in [41] a generalization of CCA to several data sets within a least-squares regression framework, and shown that it is equivalent to the maximum variance generalization. Their computational method does not require singular value decompositions of the data sets. In the following, we present and use this method as a part of our method.

Assume that we have at our disposal M data sets $\mathbf{X}_k$, $k = 1, \ldots, M$ having the same number N of data vectors. The data vectors appear as column vectors in these data sets, and their dimensionalities m_k are in general different for each set $\mathbf{X}_k$. Denote the successive (generalized) canonical vectors of the data set $\mathbf{X}_k$ by $\mathbf{h}_k^{(i)}$ and the respective canonical variables by $\mathbf{z}_k^{(i)} = \mathbf{X}_k^T \mathbf{h}_k^{(i)}$. The estimated cross-correlation matrices[1] are denoted by $\mathbf{C}_{kl} = \mathbf{X}_k \mathbf{X}_l^T$.

The least-squares type generalization of CCA can then be formulated as the problem of sequentially maximizing the generalized canonical correlation

$$\rho^{(i)} = \frac{1}{M} \sum_{k=1}^{M} \rho_k^{(i)} \tag{21}$$

where

$$\rho_k^{(i)} = \frac{1}{M-1} \sum_{l=1, l \neq k}^{M} \rho_{kl}^{(i)} \tag{22}$$

and $\rho_{kl}^{(i)} = \mathbf{h}_k^{(i)T} \mathbf{C}_{kl} \mathbf{h}_l^{(i)}$. In this case, the energy constraint which is needed for avoiding trivial solution is [41]

$$\frac{1}{M} \sum_{k=1}^{M} \mathbf{h}_k^{(i)T} \mathbf{C}_{kk} \mathbf{h}_k^{(i)} = 1 \tag{23}$$

The orthogonality constraints are for $i \neq j$

$$\mathbf{z}^{(i)T} \mathbf{z}^{(j)} = 0 \tag{24}$$

$$\mathbf{z}^{(i)} = \frac{1}{M} \sum_{k=1}^{M} \mathbf{z}_k^{(i)} \tag{25}$$

This least-squares generalization of CCA can be rewritten as a function of distances. For extracting the i:th eigenvector, the generalized CCA problem consists of minimizing with respect to the M canonical vectors $\mathbf{h}_k^{(i)}$ the cost function

$$J^{(i)} = \frac{1}{2M(M-1)} \sum_{k,l=1}^{M} \|\mathbf{X}_k \mathbf{h}_k^{(i)} - \mathbf{X}_l \mathbf{h}_l^{(i)}\|^2$$

$$= \frac{1}{M} \sum_{k=1}^{M} \|\mathbf{z}_k^{(i)}\|^2 - \rho^{(i)} \tag{26}$$

subject to the constraints (23) and (24), which implies $J^{(i)} = 1 - \rho^{(i)}$.

The solution of this generalized CCA problem can be obtained by using the method of Lagrange multipliers [41]. This leads to

[1] The scaling factor $1/N$ can be omitted here.

the generalized eigenvector problem

$$\frac{1}{M-1}(\mathbf{C}-\mathbf{D})\mathbf{h}^{(i)} = \rho^{(i)}\mathbf{D}\mathbf{h}^{(i)} \tag{27}$$

where

$$\mathbf{h}^{(i)} = [\mathbf{h}_1^{(i)T}, \mathbf{h}_2^{(i)T}, \ldots, \mathbf{h}_M^{(i)T}]^T \tag{28}$$

is a "supervector" formed by stacking the i:th canonical vectors of the M data sets $\mathbf{X}_k$, $k=1,\ldots,M$. The respective block matrices are

$$\mathbf{C} = \begin{bmatrix} \mathbf{C}_{11} & \ldots & \mathbf{C}_{1M} \\ \vdots & \ddots & \vdots \\ \mathbf{C}_{M1} & \ldots & \mathbf{C}_{MM} \end{bmatrix} \tag{29}$$

$$\mathbf{D} = \begin{bmatrix} \mathbf{C}_{11} & \ldots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \ldots & \mathbf{C}_{MM} \end{bmatrix} \tag{30}$$

Thus $\mathbf{D}$ is an $L \times L$ block diagonal matrix whose diagonal blocks are the autocorrelation matrices $\mathbf{C}_{ii}$, $i=1,\ldots,M$, of the M data sets. The matrix $\mathbf{C}-\mathbf{D}$ is an $L \times L$ block off-diagonal matrix which contains all the cross-correlation matrices $\mathbf{C}_{kl}$, $k \neq l$, of the M data sets but not their autocorrelation matrices. The solutions for this least-squares or maximum variance generalization of CCA are obtained as the eigenvectors associated with the largest eigenvalues of (27). These eigenvectors can be computed using standard numerical methods, or alternatively using a deflation type neural recursive least-squares algorithm introduced and discussed in [41].

A couple of notes are in order here. First, Eq. (4) defining standard CCA for two data sets can be written in the form (27) after some manipulation, see [35,41]. Then in (27) $\mathbf{h}(i) = [\mathbf{w}_{xi}^T, \mathbf{w}_{yi}^T]^T$. If we denote the matrix on the left-hand side of (27) by $\mathbf{O}$ (off-diagonal), (27) is equivalent to the non-symmetric eigenproblem

$$\mathbf{D}^{-1}\mathbf{O}\mathbf{h}^{(i)} = \rho^{(i)}\mathbf{h}^{(i)} \tag{31}$$

which could in principle have complex-valued eigenvectors and -values. However, Eq. (31) can be written as

$$\mathbf{O}^{1/2}\mathbf{D}^{-1}\mathbf{O}^{1/2}(\mathbf{O}^{1/2}\mathbf{h}^{(i)}) = \rho^{(i)}(\mathbf{O}^{1/2}\mathbf{h}^{(i)}) \tag{32}$$

which is a symmetric eigenproblem for the eigenvector $\mathbf{O}^{1/2}\mathbf{h}^{(i)}$. Hence the eigenvalues and -vectors of (27) are real-valued.[2]

Our method for M related data sets $\mathbf{X}_k$, $k=1,\ldots,M$ proceeds now as follows. We first estimate all the cross-correlation matrices $\mathbf{C}_{kl}$, where $k,l=1,\ldots,M$, similarly as in (10) and form then estimates of the matrices $\mathbf{C}$ and $\mathbf{D}$. We then compute the d principal generalized eigenvectors $\mathbf{h}^{(1)},\ldots,\mathbf{h}^{(d)}$, corresponding to the d largest eigenvalues from (27) or (31). Here $d \leq \min(m_1,\ldots,m_M)$. From these stacked eigenvectors we get the vectors $\mathbf{h}_k^{(1)},\ldots,\mathbf{h}_k^{(d)}$ corresponding to each data set $\mathbf{X}_k$. We then orthonormalize these vectors, yielding vectors $\mathbf{g}_k^{(i)}$, $i=1,\ldots,d$, and orthogonal projection operator

$$\mathbf{P}_{D,k} = [\mathbf{g}_k^{(1)},\ldots,\mathbf{g}_k^{(d)}] \tag{33}$$

onto the subspace spanned by them, corresponding to the dependent components in the data set $\mathbf{X}_k$. The data sets are then mapped to these basis vectors

$$\mathbf{P}_{D,k}^T\mathbf{X}_k, \quad k=1,\ldots,M \tag{34}$$

and the dependent components (sources) of each data set are found by applying any suitable ICA or BSS method to the projected data sets (34).

[2] This presumes formally that the matrix $\mathbf{O}^{1/2}$ exists. In practice, these eigenvalues and -vectors are always real-valued.

A question now arises how to estimate the independent components (sources) in each data set. A first idea would be to use the generalized eigenvectors corresponding to the smallest eigenvalues in a similar manner as above. However, if we have for example three data sets $\mathbf{X}_1$, $\mathbf{X}_2$, and $\mathbf{X}_3$ of data vectors having respectively the dimensionalities $m_1=5$, $m_2=4$, and $m_3=6$, $L=15$ and Eq. (31) has 15 stacked eigenvectors $\mathbf{h}^{(i)}$, $i=1,\ldots,15$. From them we get 15 vectors $\mathbf{h}_k^{(i)}$ for each data set $\mathbf{X}_k$. These vectors are clearly linearly dependent.

Therefore, a better solution is to construct a subspace which is orthogonal to the subspace defined by the projection operator $\mathbf{P}_{D,k}$ in (33) for each data set $\mathbf{X}_k$. An orthonormal basis for this subspace can be computed for example by taking m_k-d random vectors of dimension m_k and orthonormalizing them against the d vectors $\mathbf{g}_k^{(i)}$ in (33) and each other. The resulting vectors are used to define a projection operator

$$\mathbf{P}_{I,k} = [\mathbf{g}_k^{(d+1)},\ldots,\mathbf{g}_k^{(m_k)}] \tag{35}$$

corresponding to the independent components in $\mathbf{X}_k$. The data are then mapped onto these subspaces

$$\mathbf{P}_{I,k}^T\mathbf{X}_k, \quad k=1,\ldots,M \tag{36}$$

and the independent components are estimated by applying any suitable ICA or BSS method to the projected data sets (36).

Mathematical analysis of our method in the case of three or more data sets seems to be quite difficult, but the method is justified by good experimental results.

5. Experimental results

5.1. Simulated data, two related data sets

We have made experiments with both synthetically generated and real-world data sets. In experiments with synthetically generated data, the true sources are known and their mixtures are in our experiments usually created by mixing the sources synthetically using random numbers as mixture coefficients. Because the true source signals are known, it is possible to assess the performance of the methods using a suitable separability criterion. For real-world data, the true sources are usually unknown, and the results can be assessed qualitatively only.

In the first series of experiments described here, we used source signals which are difficult to separate. They have been defined in the Matlab code UniBSS.m [20] and explained in the respective paper [19]. There are a total of six source signals which are all stochastic, containing at least some random component. Such sources are more appropriate than deterministic sources, for instance sinusoidal signals, which are often used to illustrate the separation results of an ICA or BSS method, but visual inspection of the quality of the separation results is more difficult for them. The four first sources are generated using a first-order autoregressive model so that the two first of them are super-Gaussian and the third and fourth source are Gaussian. Furthermore, the first and third sources had identical temporal autocovariances, and similarly the second and fourth sources. The fifth and sixth sources have smoothly changing variances.

These six sources have been purposely designed so that standard ICA methods such as FastICA or the natural gradient method [2,1] based on non-Gaussianity and higher-order statistics are able to separate the two first sources only. Methods based on temporal statistics such as [9,10] are not able to separate any of them because there is no source with a unique temporal autocovariance sequence. Methods utilizing smoothly changing variances such as [12,13] are able to separate only the fifth and sixth sources. Methods combining temporal correlations

and non-Gaussianity [15,14] would be able to separate the four first sources. Only the approximative method introduced in [19] could separate all these six sources.

We picked the first three sources and the fifth source from the UniBSS.m code [20] to the first data set **X**. One statistical realization of these sources is shown in Fig. 1. We took the second and third sources to the second data set **Y**, added with the fourth and sixth sources in [20]. These sources are shown similarly in Fig. 2. Thus in the data sets **X** and **Y** there are two completely dependent sources, while the remaining two sources in them are statistically independent of all the other sources.

In this series of experiments, we used 5000 data vectors and source signal values ($t=1,2,\ldots,5000$) for providing enough data to the UniBSS method [19]. The other tested methods: CCA, FastICA, TDSEP, and their combinations require much less samples, but also their performance improves with more samples, except for CCA which performs equally well with 500 samples only. Because the results can vary a lot for single realizations of these sources and their mixtures, we computed the averages of the signal-to-noise ratios of the separated sources over 100 random realizations of the sources and the data sets **X** and **Y**. In each realization, the elements of the 4×4 mixing matrices were Gaussian random numbers.

The signal-to-noise ratios (SNR's) (or signal-to-interference ratios) of the estimated source signals were computed for each realization of the data sets and each source from the formula

$$\mathrm{SNR}(i) = 10\log_{10}\frac{\frac{1}{N}\sum_{t=1}^{N} s_i(t)^2}{\frac{1}{N}\sum_{t=1}^{N}[s_i(t)-\hat{s}_i(t)]^2} \tag{37}$$

where the numerator is the average power of the i:th source $s_i(t)$ over the N samples, and the denominator is the respective power of the difference $s_i(t)-\hat{s}_i(t)$ between the source signal $s_i(t)$ and its estimate $\hat{s}_i(t)$. We computed the averages of these SNR's over the 100 realizations for each source and its estimate, and quite similarly for the sources $r_i(t)$ of the other data set **Y**.

We not only tried our CCA based method and its combinations applying either FastICA, TDSEP, or UniBSS for post-processing to achieve better separation, but also compared it with two methods introduced by other authors for the same problem. The first compared method introduced in [32] assumes that the dependent sources in the two data sets are active simultaneously. The second compared method [28] uses multiset canonical correlation analysis. Theoretically its results should coincide with plain CCA for two data sets but in practice this may not hold due to problems such as deflationary nature of the algorithm mentioned in a later paper [29].

The separation results for the four sources contained in the first data set **X** are shown in Table 1, and for the four sources in the other data set **Y** in Table 2. For clarity, we have numbered these sources by 1–4 in the data set **X** and by 5–8 in **Y**. We set (somewhat arbitrarily) the threshold of successful separation to 10 dB based on visual inspection. Tables 1 and 2 show that CCA alone yields fairly similar separation results for all the eight sources which already lie at our separation threshold. FastICA can separate clearly the two first sources but fails for the three last sources. The TDSEP method separates well four sources, the other sources lie at the separation threshold. This method performs better than expected in context with the definition of the type of six sources. The reason is obviously that only four of the six sources are used in the data sets **X** and **Y**, making the separation problem easier. The UniBSS method separates well all the sources. The results are qualitatively similar if the dependent and independent sources are selected otherwise among the six original sources.

Combining CCA with post-processing using the FastICA, TDSEP, or UniBSS methods improves the results for all these methods, so that also FastICA and TDSEP can now separate well all the sources in this difficult separation problem. The methods introduced in

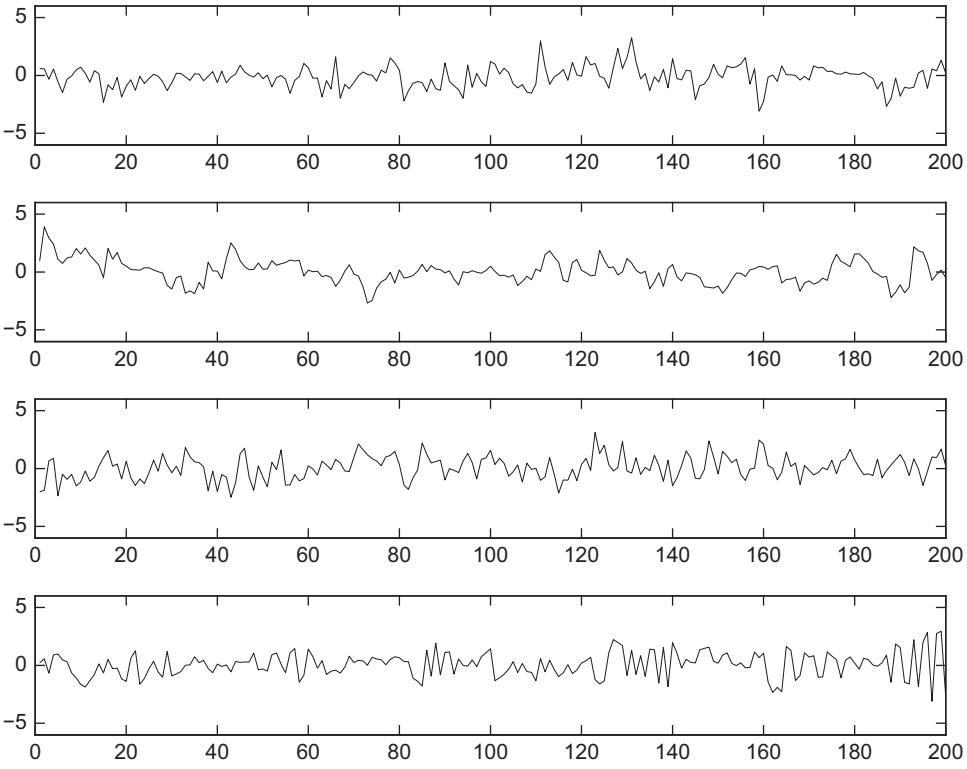

Fig. 1. Two hundred samples of the original source signals in first data set **X**. The two first sources are non-Gaussian, the third one is temporally correlated Gaussian, and the last source has smoothly changing nonstationary variance.

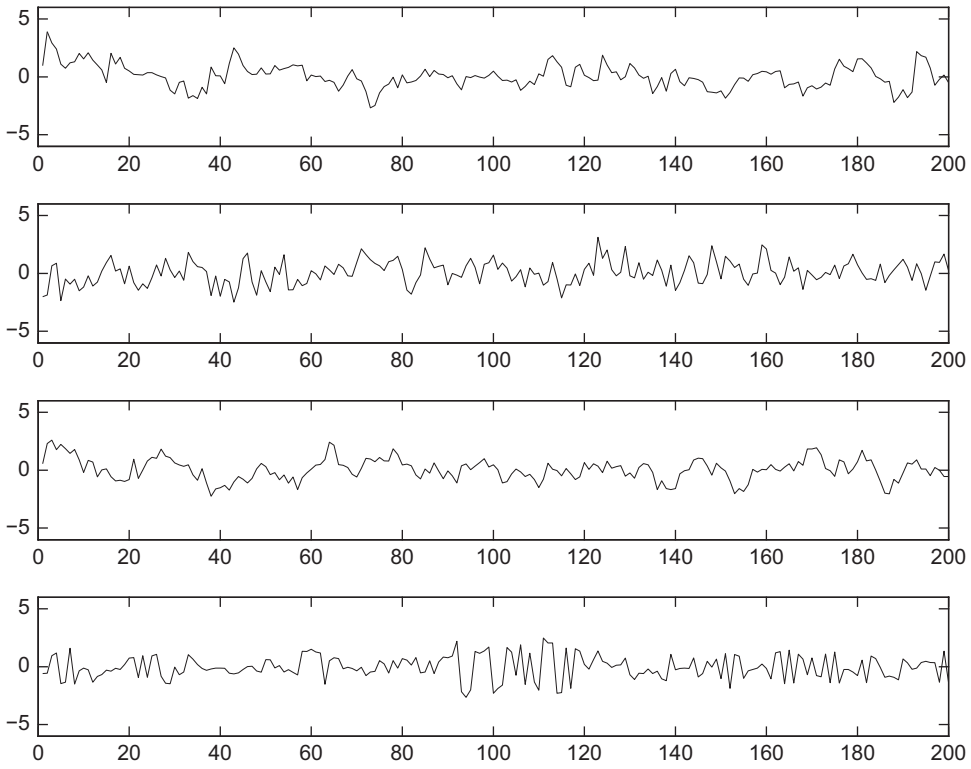

Fig. 2. Two hundred samples of the original source signals in second data set **Y**. The first one is non-Gaussian, the second and third sources are temporally correlated Gaussians, and the last source has smoothly changing nonstationary variance. The two first sources in **Y** are the same as the second and third sources in the first data set **X**.

Table 1
Signal-to-noise ratios (dB) of different methods for the source signals 1–4 in the first data set **X**.

Method	Source 1	Source 2	Source 3	Source 4
CCA	10.3	9.9	10.1	10.3
FastICA	22.5	14.1	9.4	10.6
TDSEP	10.0	30.5	10.0	27.5
UniBSS	33.9	40.7	27.6	28.5
CCA + FastICA	29.3	20.0	21.0	29.4
CCA + TDSEP	30.7	37.9	34.8	30.2
CCA + UniBSS	33.7	48.4	39.2	32.7
Method in [32]	25.7	9.8	9.4	23.1
Method in [28]	12.5	11.4	11.3	13.2

Table 2
Signal-to-noise ratios (dB) of different methods for the source signals 5–8 in the second data set **Y**.

Method	Source 5	Source 6	Source 7	Source 8
CCA	9.9	10.1	10.5	10.5
FastICA	9.5	4.6	4.2	5.2
TDSEP	9.7	26.4	9.8	28.8
UniBSS	37.1	27.0	28.6	29.0
CCA + FastICA	21.1	21.9	13.1	13.2
CCA + TDSEP	37.9	34.8	31.6	33.1
CCA + UniBSS	49.4	39.2	31.0	33.0
Method in [32]	9.8	9.4	9.5	9.5
Method in [28]	11.4	11.3	3.6	3.9

[32,28] provide clearly lower signal-to-noise ratios, failing for some sources. Using CCA combined with the FastICA or TDSEP methods is in practice often preferable over using the UniBSS method. The UniBSS method requires much more samples for reliable results, and different types of nonlinearities for sub-Gaussian and super-Gaussian sources. The FastICA and TDSEP methods do not suffer from this limitation. In the data sets **X** and **Y** there are only super-Gaussian and Gaussian sources, and therefore the same nonlinearity can be used in the UniBSS method for all of them.

To give an idea about the practical quality of separation corresponding to different SNR's, Fig. 3 shows one original source signal and its estimates provided by different methods for a single realization of the data sets. This source signal is a super-Gaussian source which is the third source in the first set $\mathbf{s}(t)$ of sources and the second dependent source in the second set $\mathbf{r}(t)$ of sources. Only 200 first samples are shown to make the details of the estimates better discernible. The signal-to-noise ratios of the estimates provided by different methods are 1.92 dB for plain FastICA, 15.5 dB for the combination of CCA and FastICA, 37.9 dB for the combination of CCA and UniBSS, 33.5 dB for plain UniBSS, and 3.5 dB for plain CCA.

Inspecting Fig. 3 visually shows that even though the SNR of plain CCA is poor, 3.5 dB only, it is anyway able to approximate some parts of the original source signal, for example the last samples, but for the other parts it fails. The combination of CCA and FastICA is clearly able to separate the source adequately with the SNR of 15.5 dB. The much better SNR's of the UniBSS method and the method combining CCA with UniBSS do not show up in the visual quality of separation results notably. Obviously finding differences in the quality of these estimates would require looking at finer details of the separation results.

We made experiments also with artificially generated mixtures of real-world speech sources taken from [42], and with other methods such as JADE [43] and SOBI [9]. The results were qualitatively similar though somewhat less convincing. In this case, especially the UniBSS method benefits from CCA preprocessing, because it would require different nonlinearities for super-Gaussian and sub-Gaussian, and we have used only one type of nonlinearity. CCA preprocessing improves the separation quality

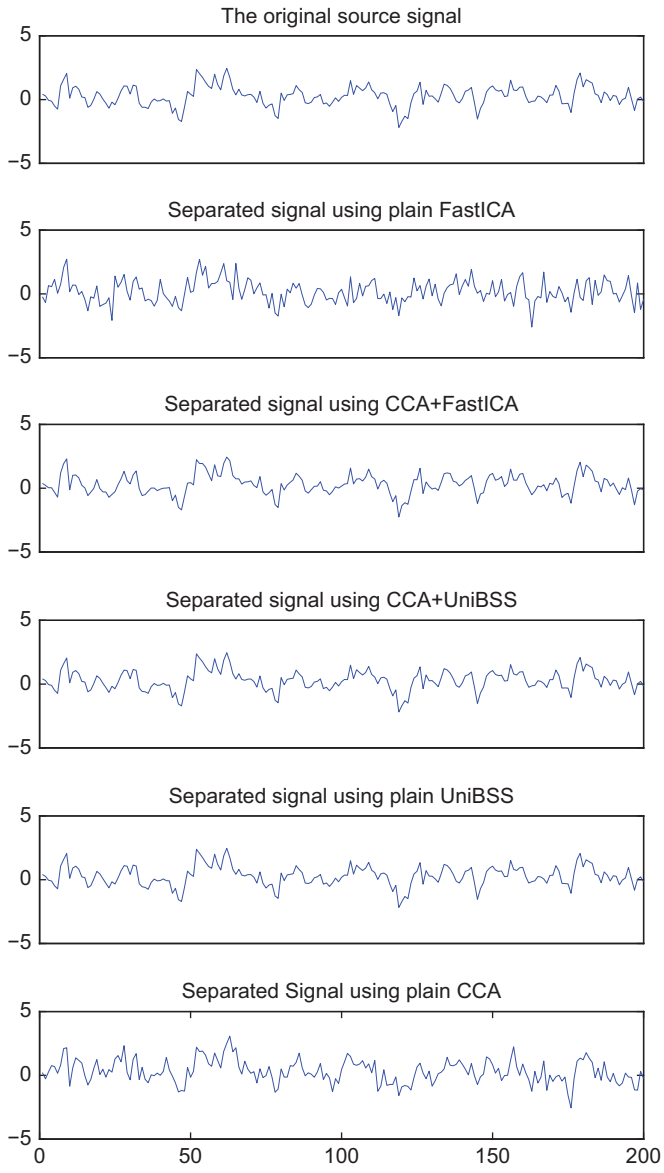

Fig. 3. First 200 samples of the original super-Gaussian third source signal in the data set $\mathbf{X}$ and its estimates given by different methods.

of the JADE method, but the overall results are not so good as for FastICA and TDSEP and are therefore not shown in Tables 1 and 2. The SOBI method fails completely, and CCA preprocessing does not improve its performance.

5.2. Simulated data, three related data sets

We used the same six source signals defined in the Matlab code UniBSS.m [20] as in the experiments with two related data sets. Furthermore, we generated three more sources in a similar manner, so that one of them was super-Gaussian, one temporally correlated Gaussian, and one had a smoothly changing variance. Due to the construction of these difficult source signals, almost all ICA and BSS methods fail to separate all of them from their mixtures. From these nine source signals we constructed three sets $\mathbf{X}_1$, $\mathbf{X}_2$, and $\mathbf{X}_3$ of 5-dimensional data vectors using randomly chosen mixing matrices. In each of these data sets there were three same sources, namely sources 1 and 2 which were super-Gaussian and source 5 which has a smoothly changing variance. Sources 3 and 4 in each data set were different and independent of all the other sources. We used 2000 data vectors and source signal values ($t=1,2,\ldots,2000$) for providing enough data especially to the UniBSS and TDSEP methods.

The results for the data sets $\mathbf{X}_1$, $\mathbf{X}_2$, and $\mathbf{X}_3$ are presented in Tables 3–5, respectively. On the first row of the tables are the results of the generalized CCA (GCCA) without any postprocessing. It shows some progress towards separation, and the results for the independent third and fourth sources are around the separation border already. FastICA [1,7,8], based on non-Gaussianity, is able to separate the non-Gaussian first and second sources in all the data sets, but fails for other types of sources as expected. The TDSEP method [10] based on temporal autocorrelations is able to separate at least marginally all the five sources in the first data set $\mathbf{X}_1$, but fails though not badly for most sources in the other two data sets $\mathbf{X}_2$ and $\mathbf{X}_3$. The UniBSS method [19,20] is able to separate all the sources, though some of them rather marginally. It may benefit from the construction of the sources using a first-order autoregressive model as its uses just this first autocorrelation.

Preprocessing using generalized canonical correlation analysis (GCCA) improves the separation results for most sources and all

Table 3
Signal-to-noise ratios (dB) of different methods for the source signals S1–S5 in the first data set $\mathbf{X}_1$.

Method	S1	S2	S3	S4	S5
GCCA	4.6	4.7	10.2	10.2	4.5
FastICA	18.3	16.8	9.9	6.1	6.9
TDSEP	15.5	18.8	10.2	10.2	16.8
UniBSS	27.5	26.4	31.7	24.8	23.9
GCCA+FastICA	26.1	25.7	15.5	15.2	23.8
GCCA+TDSEP	16.4	22.1	10.3	10.5	17.6
GCCA+UniBSS	32.5	33.5	25.9	24.2	28.3
Method in [32]	25.0	27.1	6.9	6.7	24.7
Method in [28]	6.2	5.8	6.2	6.1	4.9

Table 4
Signal-to-noise ratios (dB) of different methods for the source signals S1–S5 in the second data set $\mathbf{X}_2$.

Method	S1	S2	S3	S4	S5
GCCA	4.6	4.7	9.9	9.8	4.5
FastICA	17.3	16.1	5.4	6.9	5.3
TDSEP	7.7	17.9	7.9	8.7	8.1
UniBSS	26.0	28.3	11.1	18.5	10.7
GCCA+FastICA	26.1	25.8	12.4	12.3	23.8
GCCA+TDSEP	16.4	22.1	19.1	19.3	17.6
GCCA+UniBSS	31.8	33.3	21.7	21.9	27.7
Method in [32]	25.1	28.6	17.5	21.2	24.9
Method in [28]	6.2	5.8	2.5	2.3	4.9

Table 5
Signal-to-noise ratios (dB) of different methods for the source signals S1–S5 in the third data set $\mathbf{X}_3$.

Method	S1	S2	S3	S4	S5
GCCA	4.6	4.7	10.2	10.1	4.5
FastICA	14.7	13.8	4.1	3.8	3.9
TDSEP	11.9	8.8	9.1	9.0	8.8
UniBSS	25.9	27.6	13.8	12.9	10.6
GCCA+FastICA	26.1	25.8	10.1	10.2	23.8
GCCA+TDSEP	16.4	22.1	25.1	24.5	17.6
GCCA+UniBSS	32.6	33.7	19.2	19.4	28.7
Method in [32]	24.6	28.6	9.9	10.2	24.5
Method in [28]	6.2	5.8	8.8	9.2	4.9

the tested methods, FastICA, TDSEP, and UniBSS. Not only are the SNR's of separated sources often much higher but GCCA preprocessing helps FastICA and TDSEP to separate sources that they alone are not able to separate. These results are qualitatively similar as for two data sets in the previous subsection.

We also again compared our method with two methods introduced by other authors for the same problem. The first compared method [32] assumes that the dependent sources in the data sets are active simultaneously. From Tables 3–5 one can see that it performs quite well for the dependent first, second, and fifth sources in all the three data set, but fails for the independent third and fourth source in the first data set $\mathbf{X}_1$, and lies at separation border for these sources in the third data set $\mathbf{X}_3$. The second compared method [28] uses multiset canonical correlation analysis. It makes some progress towards separation for most sources, but fails at least marginally for all of them in this difficult separation task.

We tested also the dependence of the methods on the number of samples N in the data sets. Generalized CCA (GCCA) performs in practice equally well using 500 samples (data vectors) only, but the other methods FastICA, TDSEP, and UniBSS provide much better results when the number of samples increases. Even the UniBSS method fails to separate some of the sources when the number of samples is 500 or 1000.

5.3. Single data set

Fig. 4 shows another data set of seven biomedical sources, containing both sub-Gaussian and super-Gaussian sources which have quite different characteristics. This is the data set ABio7.mat of the signal processing benchmarks in the ICALAB Toolbox [42]. In Table 6, we present separation results provided by different methods in the case of a single data set $\mathbf{X}$ which is a random mixture of these sources with a square 7×7 mixing matrix $\mathbf{A}$. For the CCA method and preprocessing using it, the other related data set $\mathbf{Y}$ was created using time delays $d_1 = 1$ and $d_2 = 2$ in (19). Thus the original data set $\mathbf{X}(n)$ was correlated with four time delayed data sets $\mathbf{X}(n-2)$, $\mathbf{X}(n-1)$, $\mathbf{X}(n+1)$, and $\mathbf{X}(n+2)$. This gave the best results of the several related data sets that we tried. In general, applying our method to a single data set requires creating artificially a related reference data set which is usually highly correlated with the original data set. It has turned out that separating all the sources successfully is then often more difficult than for two naturally existing data sets which are related.

The performance of various methods tested is this time different from previous simulations. Now CCA containing time delays performs the best of all the methods tested in the sense

Table 6
The average signal-to-ratios over 100 realizations for the seven sources S1–S7 for a single biomedical data set.

Method	S1	S2	S3	S4	S5	S6	S7
CCA	22.3	24.4	29.6	23.2	14.5	21.1	14.4
FastICA	11.4	14.6	13.4	29.9	11.8	18.4	33.0
UniBSS	2.8	4.7	15.0	18.0	3.0	5.3	4.0
TDSEP	4.4	24.1	27.4	21.7	14.4	26.8	3.1
CCA + FastICA	11.3	11.9	12.2	25.0	13.4	18.8	33.7
CCA + UniBSS	9.8	18.2	30.5	18.1	14.1	24.3	15.0
CCA + TDSEP	4.4	24.1	27.4	21.7	14.4	26.8	3.1

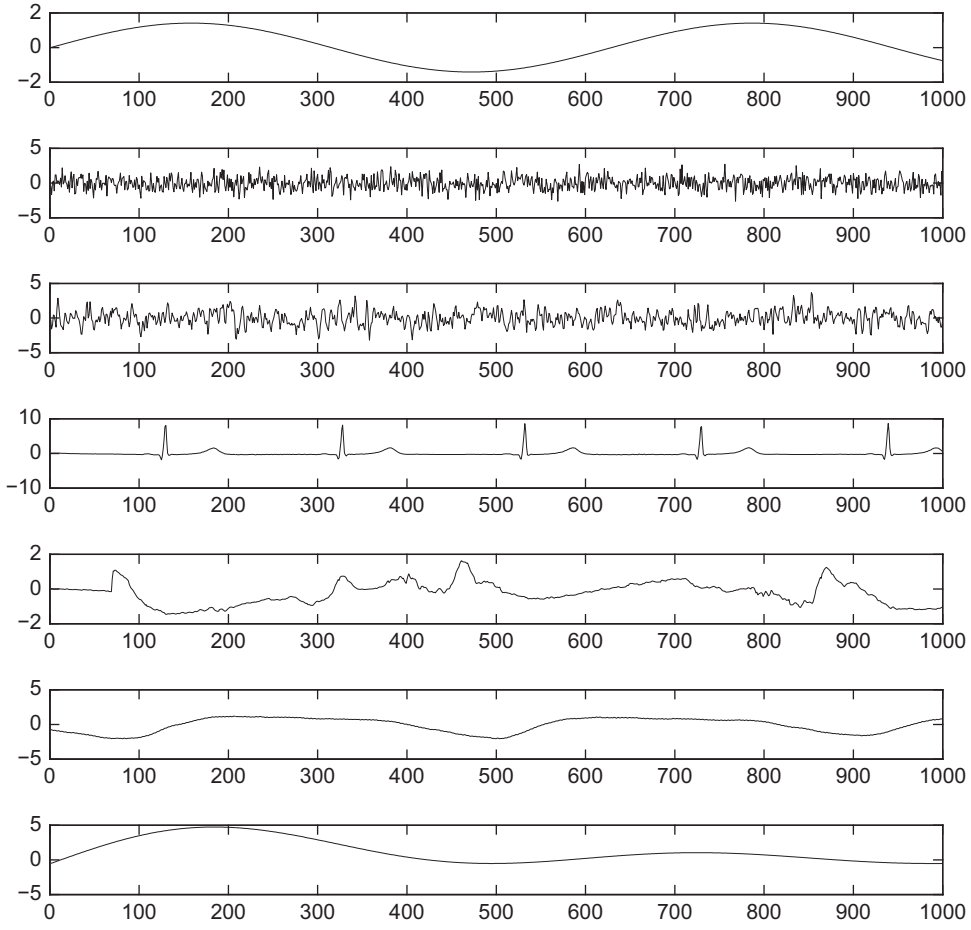

Fig. 4. First 1000 samples of the biomedical data set from ICALAB.

that it is able to separate all the sources with a clear margin. FastICA separates all the sources, too, but its signal-to-noise ratios are poorer for five of the seven sources, being much worse for the three first sources. FastICA with CCA preprocessing provides about the same quality of separation than FastICA alone. The TDSEP method is not able to separate the first and last source. Preprocessing with CCA does not help TDSEP, because the separation results are the same as for TDSEP alone. This is probably due to the fact that the TDSEP method already uses time delays (temporal correlations), and therefore adding time-delayed data sets do not help it. The most general UniBSS method is able to separate only two of the seven sources. A major reason for this is that we used the same nonlinearity for all the sources, and in this data set there are both sub-Gaussian and super-Gaussian sources which require different types of nonlinearities. Changing this nonlinearity from $g(u) = -\tanh(u)$ which is suitable for super-Gaussian sources to $g(u) = -u + \tanh(u)$ which is recommended in [19] for sub-Gaussian sources did not improve the results of the UniBSS method. However, CCA preprocessing helps it greatly, so that the combined CCA followed by UniBSS method can clearly separate six of the seven sources, and fails with quite narrow margin for the first source only.

5.4. Robot grasping data

This real-world robot data set consists of samples from a robot arm that is used for picking off and sorting different types of garbage from a conveyor belt. In this experimental setting there are several sensors in different parts of the robot arm. The sensor data used in our experiments consist of two data sets. First, there is the wrist which guides the arm of the robot to turn so that its grasping hand containing three fingers moves to a correct position. This force sensitive wrist data set **X** consists of four attributes: three of them are used to represent the status of movement in Euclidean three dimensional space. The fourth attribute is used to represent the status of the rotation in one direction. The other related data set **Y** consists of 7-dimensional position information about the wrist using Euclidean distance measure and standard quaternion representation in computer graphics and robotics. A mathematical model describing the relationship between these two data sets is not known. Quite probably the data model of this paper and its assumptions hold as an approximation only.

We can argue that there should be some dependent changes in the force sensitive wrist data set **X** as well as independent changes with respect to the position information data set **Y**. For instance, when the wrist is sent to grab a rather heavy object, the wrist sensor data set not only expresses the position information, but also provides some feedback on holding a heavy thing in the robot arm. Furthermore, when the arm is moving into some direction, the wrist sensor data should indicate the status of the wrist along with the change in the position. Therefore, these robot data sets should suit well for testing our methods. The goal is to separate the wrist signals to the dependent parts, which have strong relationships between the relative moments, and to the independent parts, showing the impacts from the external world, such as grabbing a heavy object.

We first preprocessed both the data sets by making their means zero and by whitening them. Furthermore, the second originally 7-dimensional position data set **Y** was transformed to a 4-dimensional data set, too, by converting the 4-dimensional quaternion representation to Euler angles in space. Furthermore, we found in our experiments that better results are obtained by using first-order and second-order differences of subsequent values of each component for the latter data set **Y**. Because the original components represent position information, these first-order differences approximate their first derivative with respect to time, which is local velocity. Second-order differences approximate the second derivative of position with respect to time, which is local acceleration in the direction of the respective coordinate.

Using second-order differences can be justified by the classical law of physics: The force $F = ma$, where m is the mass of the object and a is its acceleration. Here F is the external force applied to the objects handled by the robot, and it is thus linearly proportional to the acceleration.

Figs. 5 and 6 show the results for the experiment where we used second-order differences for the position data set **Y** using then CCA followed by the UniBSS method. The four singular values in the diagonal matrix **D** in (6) were 0.580, 0.340, 0.132, and 0.035. In this case, the first two singular values correspond to mutually dependent components, and the last quite small one to independent components in the two data sets **X** and **Y**. The third singular value 0.132 is relatively small, and it was deemed to correspond another pair of independent components. Inspecting the first and second dependent components of the data sets

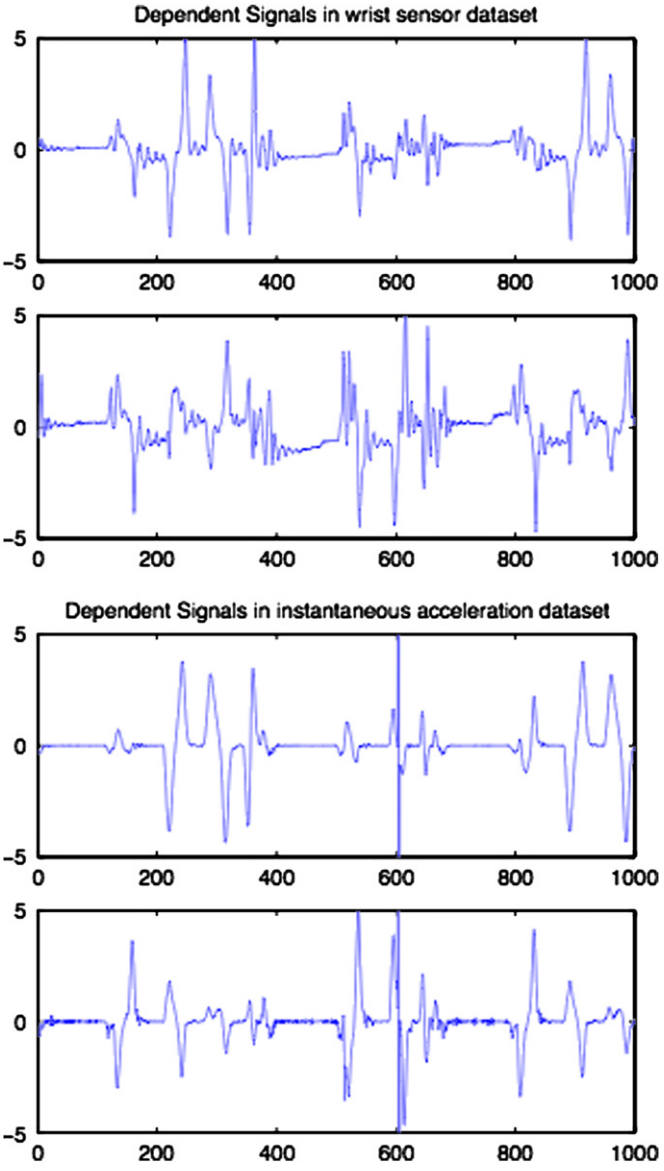

Fig. 5. Dependent signals in the robot data sets for the wrist data (two top subfigures) and for the instantaneous acceleration data (two bottom subfigures).

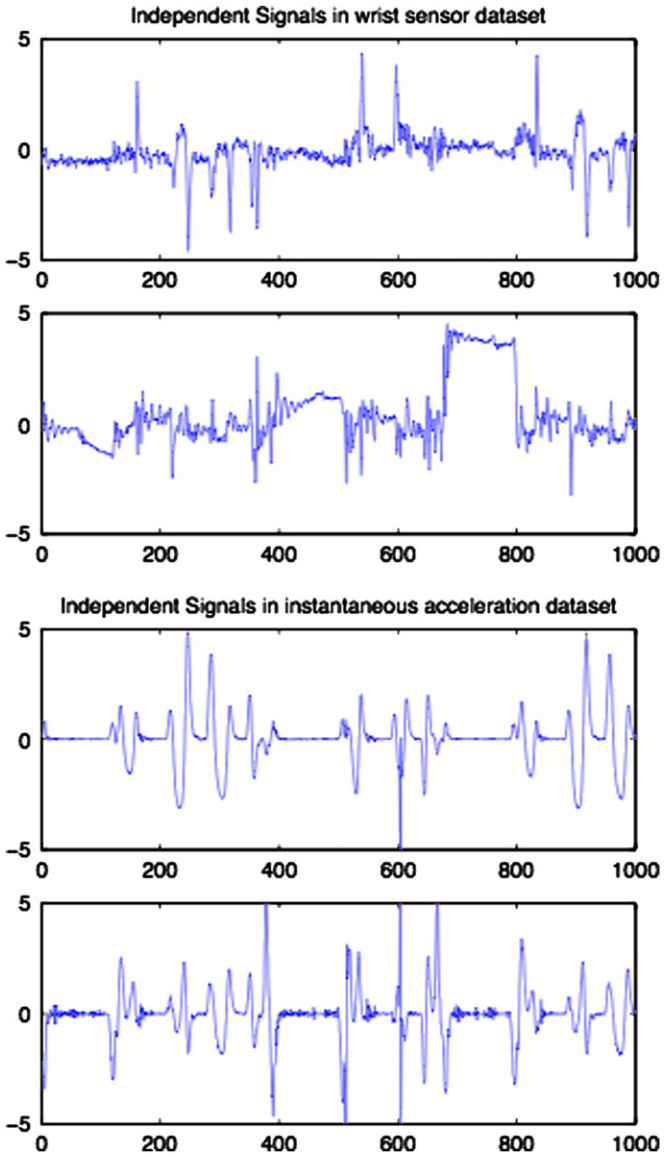

Fig. 6. Independent signals in the robot data sets for the wrist data (two top subfigures) and for the instantaneous acceleration data (two bottom subfigures).

depicted in Fig. 5 shows clearly their dependence. Here one must recall the sign ambiguity in ICA and BSS methods: if the separated source has different sign than the original one, peaks correspond to bumps and vice versa. On the other hand, especially the second components in Fig. 6 are quite clearly independent.

5.5. Real-world fMRI data

We tested the usefulness of our method with data from a functional magnetic resonance imaging (fMRI) study [44], where it is described in more detail. We used the measurements of two healthy adults while they were listening to spoken safety instructions in 30 s intervals, interleaved with 30 s resting periods. In these experiments we used slow feature analysis (SFA) described in detail in [45] for post-processing the results given by CCA, because it gave better results than FastICA. FastICA tended to overfit the separated components in this application. Slow feature analysis, which tries to learn invariant or slowly varying features from its input vectors, did not suffer from this deficiency. All the data were acquired at the Advanced Magnetic Imaging Centre of Aalto University, using a 3.0 T MRI scanner (Signa EXCITE 3.0 T; GE Healthcare, Chalfont St. Giles, UK).

Figs. 7 and 8 show the results of applying our method to the two datasets and separating 11 components from the dependent subspaces **U1** and **V1**. The left hand sides of these figures show the time courses of the separated sources, where the shaded areas correspond to periods of spoken instructions and the white areas resting periods. The right hand sides of these figures show the corresponding spatial patterns. Our method provided most of these components in correct order, but a few components were ordered manually to match best each other. The method was applied to the fMRI in a spatial manner, so that the spatial patterns are the identified components and the corresponding time courses are the projection vectors. The method was applied to the whole head measurements without preselected regions of interest, and the data matrices contained 80 time points with 263,361 samples, which were not compressed. The spatial patterns are not thresholded, but the color intensities are based on comparing the weight values to a standardized Gaussian distribution, where the red colors correspond to the positive tail of the pdf and the blue colors to the negative tail, respectively. Positive spatial weight values mean that the time course in that region corresponds to the time course shown on the left, whereas negative values mean that the time course is the inverse.

The consistency of the components across the subjects is quite good. The first component shows a global hemodynamic contrast, where large areas inside the brain have negative values and the surface of the brain is positive. The clear contrast could also be a scanning related artifact or an effect produced by the standard fMRI preprocessing of the datasets.

The activity in the second component is focused on the primary auditory cortices. The time course of the activity also closely follows the stimulation blocks. The third component shows a weakly task-related activity, with positive regions around the anterior and posterior cingulate gyrus. These areas have been identified in many studies to be part of a bigger network related with novelty of the stimulus, introspection and default-state-network. The areas of activation in the fourth component partly overlap with those in the third one. However, in this case the activation is positive in the anterior part and negative in the posterior. This clearly shows that the activity of these areas is too complex to be described by a single component.

The rest of the components are not directly stimulus related, but the activated areas have been consistently identified in the earlier studies. Some of them appear to be well-known supplementary audio and language processing areas in the brain.

These results are promising and in good agreement with the ones reported in [44]. Generally, the activated areas identified by our method are the same as or very close to the ones previously reported. There are some differences when compared to the earlier FastICA results, as the method seems to enhance contrasts within the components. There are both strongly positive and negative values in each component. Furthermore, the first component has not been identified by using FastICA. Future experiments are needed with multiple datasets for interpreting the found components more thoroughly, and a more extensive comparison with existing ICA and BSS methods using real-world data should be carried out.

6. Discussion

Even though the UniBSS method [19] performed on an average best in these experiments, it has some drawbacks. First, it requires at least of the order of 1000 samples to work appropriately, while for example FastICA needs less samples for

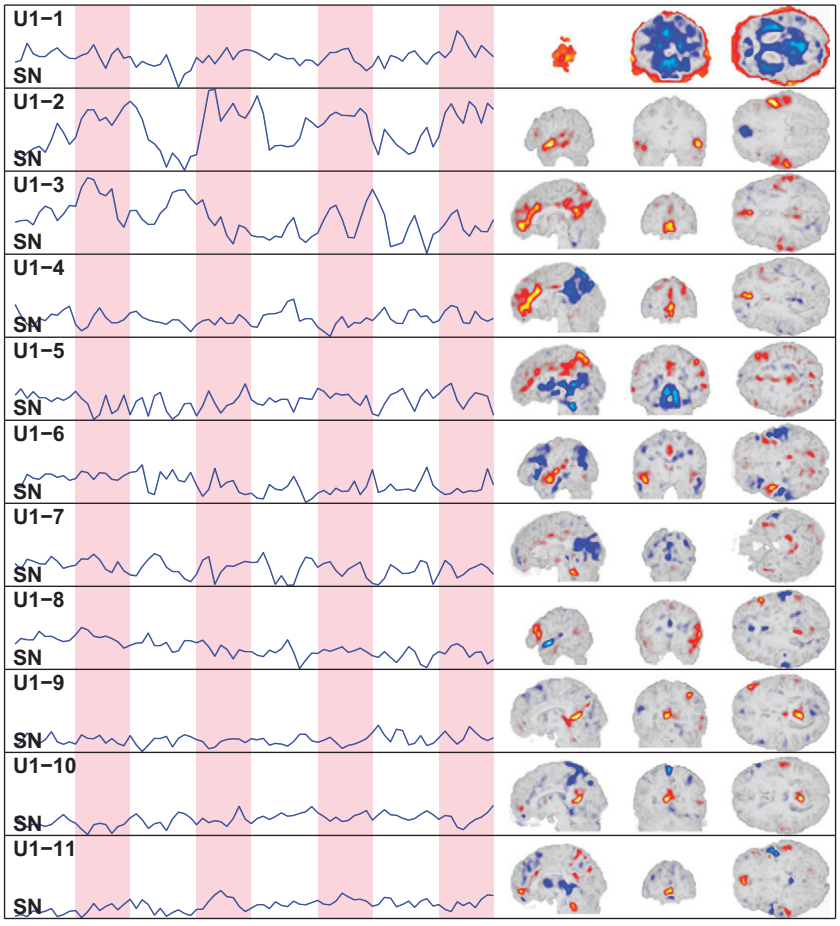

Fig. 7. Experimental results with fMRI data. Each row shows one of the 11 separated components. The activation time course with the stimulation blocks for reference is shown on the left, and the corresponding spatial pattern on three coincident slices on the right. Components from the first dataset.

providing pretty good estimates of the sources if there are just a few of them. Second, the UniBSS method requires many iterations and it does not converge uniformly. It may already provide good estimates but then still with more iterations deviate far away from a good solution, giving then rather poor estimates of the source signals. This can happen several times until the method eventually permanently converges to a good solution. A third drawback of the UniBSS method is that just like the well-known natural gradient algorithm [1,2], it requires different types of nonlinearities for super-Gaussian and sub-Gaussian source signals. Thus one should know or somehow be able to estimate how many super-Gaussian and sub-Gaussian sources the data set contains, otherwise the UniBSS methods fails to separate some sources. In our experiments with synthetically generated data this was not a problem because all the sources were either super-Gaussian or Gaussian. However, FastICA and TDSEP methods do not suffer from this limitation. In practice, using them together with CCA or generalized CCA is often a preferable choice over using the UniBSS method.

Canonical correlation analysis is based on second-order statistics, that is, autocovariances and cross-covariances of the two related data sets. Furthermore, like PCA it can be derived from a probabilistic model in which all the involved random vectors are Gaussian [24]. We are not aware of a probabilistic model for the least-squares generalization of CCA that we have used, but it also uses second-order statistics only, collected into the matrices (29) and (30). In our method, this is not so great limitation as one might expect, because all the information including higher-order statistics and non-Gaussianity contained in the related data sets are retained in mapping them to the subspaces corresponding to their dependent and independent components in (34) and (36), or (12) for two data sets.

The division into these subspaces is now based on inspection of the magnitudes of singular values of the cross-covariance matrix of whitened data sets. One could argue that also higher-order statistics should be taken into account in determining these subspaces. However, even this is often not critical because the final goal is to separate all the sources in the related two data sets irrespective of how dependent or independent they are from each other and in which way they are divided into these subspaces.

We tested our method also in experiments in which the sources were only partly dependent or the number of dependent and independent sources were purposely incorrect. Our method performed still pretty well in these cases. However, it does not work well for underdetermined mixtures where there are more source signals than mixtures. On the other hand, overdetermined mixtures having more mixtures than sources pose no problems, our method clarified this situation excellently. Another remark from our experiments is that if an ICA or a BSS method fails completely in a separation task, then preprocessing with CCA does not help.

A final remark concerns our data model. Contrary to the compared methods [32,28], it does not assume that in the two data sets there are always pairs of sources that are mutually dependent but independent of all the other sources. Thus our method could be applied to the case of independent subspaces in which there are subpaces of dependent sources in each data set. However, the theoretical argument (17) does not hold then any more because $\mathbf{Q} = E\{\mathbf{s}\mathbf{r}^T\}$ is no longer a diagonal matrix.

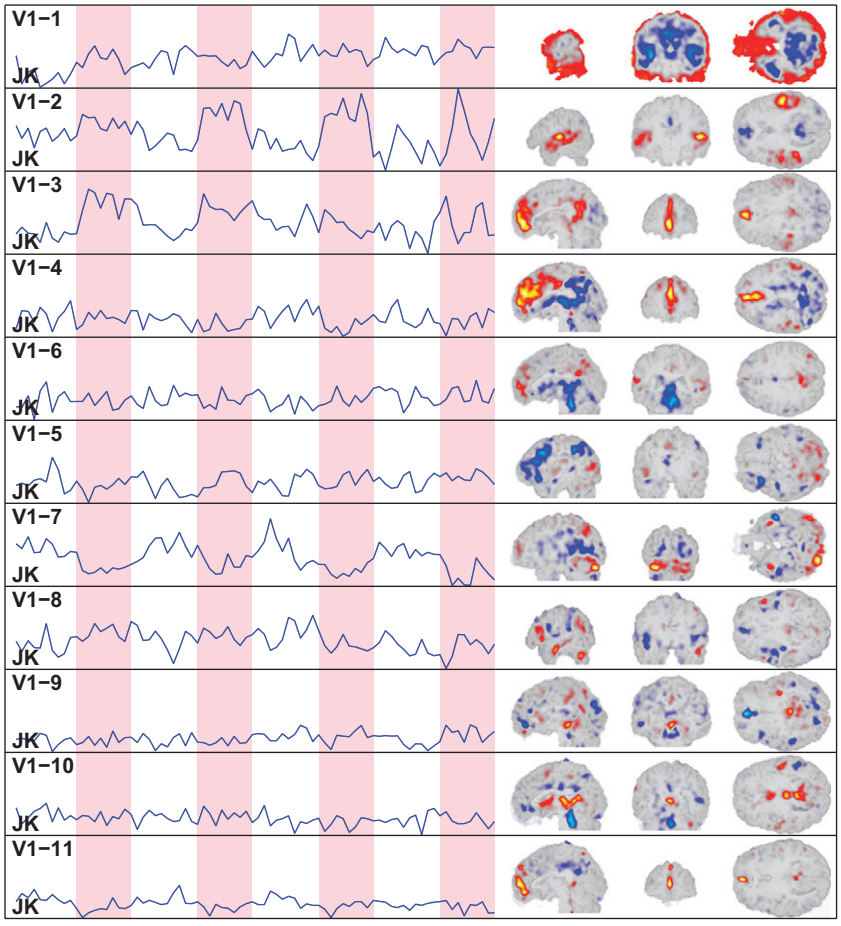

Fig. 8. Experimental results with fMRI data. Each row shows one of the 11 separated components. The activation time course with the stimulation blocks for reference is shown on the left, and the corresponding spatial pattern on three coincident slices on the right. Components from the second dataset.

7. Conclusions

In this paper, we have introduced a method based on canonical correlation analysis (CCA) and its maximum variance generalization using least-squares formulation for blind source separation from related data sets. The goal is to separate mutually dependent and independent components or source signals from these data sets. We use CCA and this its generalization for first detecting subspaces of independent and dependent components. Any ICA or BSS method can after this be used for final separation of these components. The proposed method performs quite well for synthetic data sets for which the assumed data model holds exactly. It provides interesting and meaningful results for real-world robot grasping and functional magnetic resonance imaging (fMRI) data. The method is straightforward to implement and computationally not too demanding. The proposed method improves clearly the separation results of several well-known ICA and BSS methods compared with the situation in which generalized CCA is not used.

Acknowledgments

The authors are grateful to Dr. Harri Valpola for providing us the robot grasping data set and helping to interpret the results on it. Dr. Valpola is with ZenRobotics Ltd, and also with the Department of Information and Computer Science in Aalto University School of Science, Espoo, Finland. This work was supported by the Adaptive Informatics Research Centre of Excellence of the Academy of Finland.

References

[1] A. Hyvärinen, J. Karhunen, E. Oja, Independent Component Analysis, Wiley, 2001.
[2] A. Cichocki, S.-I. Amari, Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications, Wiley, 2002.
[3] P. Comon, C. Jutten (Eds.), Handbook of Blind Source Separation: Independent Component Analysis and Applications, Academic Press, 2010.
[4] S. Roberts, R. Everson (Eds.), Independent Component Analysis: Principles and Practice, Cambridge University Press, 2001.
[5] J.-F. Cardoso, The three easy routes to independent component analysis; contrasts and geometry, in: Proceedings of the Third International Conference on Independent Component Analysis and Blind Signal Separation (ICA2001), San Diego, CA, USA, 2001, pp. 1–6.
[6] P. Comon, Independent component analysis—a new concept, Signal Process. 36 (1994) 287–314.
[7] A. Hyvärinen, Fast and robust fixed-point algorithms for independent component analysis, IEEE Trans. Neural Networks 10 (3) (1999) 626–634.
[8] A. Hyvärinen, et al., The FastICA Package for Matlab, 2005, Available at 〈http://research.ics.tkk.fi/ica/fastica/〉.
[9] A. Belouchrani, K.A. Meraim, J.-F. Cardoso, E. Moulines, A blind source separation technique based on second order statistics, IEEE Trans. Signal Process. 45 (2) (1997) 434–444.
[10] A. Ziehe, K.-R. Müller, TDSEP—an efficient algorithm for blind source separation using time structure, in: Proceedings of the International Conference on Artificial Neural Networks (ICANN'98), Skövde, Sweden, 1998, pp. 675–680.
[11] A. Yeredor, Second-order methods based on color, in: P. Comon, C. Jutten (Eds.), Handbook of Blind Source Separation: Independent Component Analysis and Applications, Academic Press, 2010, pp. 227–279.
[12] D.-T. Pham, J.-F. Cardoso, Blind separation of instantaneous mixtures of non stationary sources, IEEE Trans. Signal Process. 49 (9) (2001) 1837–1848.

[13] A. Hyvärinen, Blind source separation by nonstationarity of variance: a cumulant-based approach, IEEE Trans. Neural Networks 12 (6) (2001) 1471–1474.
[14] K.-R. Müller, P. Philips, A. Ziehe, Jade$_{TD}$: combining higher-order statistics and temporal information for blind source separation (with noise), in: Proceedings of International Workshop on Independent Component Analysis and Signal Separation (ICA'99), Aussois, France, 1999, pp. 87–92.
[15] A. Hyvärinen, Complexity pursuit: separating interesting components from time-series, Neural Comput. 13 (4) (2001) 883–898.
[16] J. Gorodnitzky, A. Belouchrani, Joint cumulant and correlation based signal separation with application to EEG data, in: Proceedings of the Third International Conference on Independent Component Analysis and Blind Signal Separation (ICA2001), San Diego, CA, USA, 2001, pp. 475–480.
[17] S. Cruces-Alvarez, A. Cichocki, Combining blind source extraction with joint approximate diagonalization: thin algorithm for ICA, in: Proceedings of the Third International Conference on Independent Component Analysis and Blind Signal Separation (ICA2001), San Diego, CA, USA, 2001, pp. 740–745.
[18] P. Tichavsky, Z. Koldowsky, A. Yeredor, G. Gomez-Herrero, E. Doron, A hybrid technique for blind separation of non-Gaussian and time-correlated sources using a multicomponent approach, IEEE Trans. Neural Networks 19 (3) (2008) 421–430.
[19] A. Hyvärinen, A unifying model for blind separation of independent sources, Signal Process. 85 (7) (2005) 1419–1427.
[20] A. Hyvärinen, Basic Matlab Code for the Unifying Model for BSS, 2003–2006, Available at 〈www.cs.helsinki.fi/u/ahyvarin/code/UniBSS.m〉.
[21] H. Hotelling, Relations between two sets of variates, Biometrika 28 (1936) 321–377.
[22] J. Karhunen, T. Ukkonen, Extending ICA for finding jointly dependent components from two related data sets, Neurocomputing 70 (2007) 2969–2979.
[23] J. Ylipaavalniemi, E. Savia, S. Malinen, R. Hari, R. Vigario, S. Kaski, Dependencies between stimuli and spatially independent fMRI sources: towards brain correlates of natural stimuli, NeuroImage 48 (1) (2009) 176–185.
[24] F. Bach, M. Jordan, A probabilistic interpretation of canonical correlation analysis, Technical Report 688, Department of Statistics, University of California, Berkeley, CA, USA, 2005, Available at 〈http://www.di.ens.fr/~fbach/〉.
[25] E. Savia, A. Klami, S. Kaski, Fast dependent components for fMRI analysis, in: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'09), Taipei, Taiwan, 2009, pp. 1737–1740.
[26] A. Klami, S. Virtanen, S. Kaski, Bayesian exponential family projections for coupled data sources, in: P. Grunwald, P. Spirtes (Eds.), Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence (UAI 2010), AUAI Press, Corvallis, OR, USA, 2010, pp. 286–293.
[27] N. Correa, T. Adali, Y.-Q. Li, V. Calhoun, Canonical correlation analysis for data fusion and group inferences, IEEE Signal Process. Mag. 27 (4) (2010) 39–50.
[28] Y.-Q. Li, T. Adali, W. Wang, V. Calhoun, Joint blind source separation by multiset canonical correlation analysis, IEEE Trans. Signal Process. 57 (10) (2009) 3918–3928.
[29] M. Anderson, X.-L. Li, T. Adali, Nonorthogonal independent vector analysis using multivariate Gaussian model, in: V. Vigneron, et al., (Eds.), Proceedings of the Ninth International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA 2010), St. Malo, France, Lecture Notes in Computer Science, vol. 6365, Springer Verlag, Berlin, 2010, pp. 354–361.
[30] J. Koetsier, D. MacDonald, D. Charles, C. Fyfe, Exploratory correlation analysis, in: Proceedings of the 10th European Symposium on Artificial Neural Networks (ESANN2002), Bruges, Belgium, 2002, pp. 483–488.
[31] P. Lai, C. Fyfe, Kernel and nonlinear canonical correlation analysis, Int. J. Neural Syst. 10 (5) (2000) 365–377.
[32] M. Gutmann, A. Hyvärinen, Extracting coactivated features from multiple data sets, in: T. Honkela, W. Duch, M. Girolami, S. Kaski (Eds.), Proceedings of the International Conference on Artificial Neural Networks (ICANN 2011), Espoo, Finland, Lecture Notes in Computer Science, vol. 6791, Springer Verlag, Berlin, 2011, pp. 323–330.
[33] S. Akaho, Y. Kiuchi, S. Umeyama, MICA: multidimensional independent component analysis, in: Proceedings of the 1999 International Joint Conference on Neural Networks (IJCNN'99), IEEE Press, Washington, DC, USA, 1999, pp. 927–932.
[34] A. Rencher, Methods of Multivariate Analysis, second ed., Wiley, 2002.
[35] M. Borga, Canonical Correlation: A Tutorial, 2001, Available at 〈www.imt.liu.se/~magnus/cca/tutorial/〉.
[36] S. Haykin, Modern Filters, MacMillan, 1989.
[37] Wikipedia, The Free Encyclopedia, Articles "Orthogonal matrix" and "Permutation matrix" 〈http://en.wikipedia.org/wiki/〉.
[38] O. Friman, M. Borga, P. Lundberg, H. Knutsson, Exploratory fMRI analysis by autocorrelation maximization, NeuroImage 16 (2) (2002) 454–464.
[39] W. Liu, D. Mandic, A. Cichocki, Analysis and online realization of the CCA approach for blind source separation, IEEE Trans. Neural Networks 18 (5) (2007) 1505–1510.
[40] J. Kettenring, Canonical analysis of several sets of variables, Biometrika 58 (3) (1971) 433–451. Available in Electronic form at 〈http://www.jstor.org/stable/2334380〉..
[41] J. Via, I. Santamaria, J. Perez, A learning algorithm for adaptive canonical correlation analysis of several data sets, Neural Networks 20 (2007) 139–152.
[42] A. Cichocki et al., ICALAB for Signal Processing MATLAB Toolbox, Available at 〈http://www.bsp.brain.riken.jp/ICALAB/〉.
[43] J. Cardoso, A. Souloumiac, Blind beam forming for non Gaussian signals, IEE Proc.: F 140 (6) (1993) 362–370.
[44] J. Ylipaavalniemi, R. Vigario, Analyzing consistency of independent components: an fMRI illustration, NeuroImage 39 (1) (2008) 169–180.
[45] L. Wiskott, T. Sejnowski, Slow feature analysis: unsupervised learning of invariances, Neural Comput. 14 (2002) 715–770.

Juha Karhunen received the DSc (Tech.) degree from the Helsinki University of Technology in 1984. Since 1976, he has been in the Laboratory of Computer and Information Science at Helsinki University of Technology, Espoo, Finland, where he became a Professor in Computer Science in 1999 (specialization area: neural networks and signal processing). After changes in organization, this laboratory has been merged to form the Department of Information and Computer Science at Aalto University School of Science.

Juha Karhunen's current research interests include unsupervised variational Bayesian learning, deep learning, independent component analysis, blind source separation, and their applications. Prof. Karhunen has published more than 100 conference and journal papers. He is a co-author of the book A. Hyvärinen, J. Karhunen, and E. Oja, "Independent Component Analysis", Wiley 2001, which has become a standard reference in the fields of independent component analysis and blind source separation.

Tele Hao was born in China. He has studied in the International Macadamia (Machine Learning and Data Mining) Study Programme at Aalto University Department of Information and Computer Science, where he is currently working. He got the Diploma Engineer (Master of Science) degree in December 2012. Tele Hao's research interests include blind signal separation and deep learning. He has already several conference papers on this topics.

Jarkko Ylipaavalniemi received his Master's degree in Computer and Information Science from the Helsinki University of Technology, Finland, in 2005. He is a PhD candidate in the School of Science in the Aalto University, working under the supervision of Prof. Erkki Oja and Dr. Ricardo Vigarion. His research interests include machine learning, data-driven approaches, and functional brain imaging.