# A Generalized Canonical Correlation Analysis Based Method for Blind Source Separation from Related Data Sets

Juha Karhunen, Tele Hao, and Jarkko Ylipaavalniemi

***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. Any ICA or BSS method can after this be used for final separation of these components. The proposed method performs well for synthetic data sets for which the assumed data model holds, and provides interesting and meaningful results for real-world 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.**

## I. Introduction

### *A. Independent component analysis and blind source separation*

Independent component analysis (ICA) and related blind source separation (BSS) methods [1], [8], [9] are nowadays well understood techniques for blind extraction of useful information from vector-valued data $\mathbf{x}$ with many applications.

The data model used in standard linear ICA is simply

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

Thus each data vector $\mathbf{x}(t)$ is expressed as a linear combination of independent components or source signals $s_i(t)$, $i = 1, 2, \ldots, m$, which multiply the respective constant basis vectors $\mathbf{a}_i$. The source vector $\mathbf{s}(t) = [s_1(t), s_2(t), \ldots, s_m(t)]^T$ contains the source signals, 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 $\mathbf{X} = [\mathbf{x}(1), \ldots, \mathbf{x}(N_x)]$ but once they have been estimated, they are the same for all the data vectors in $\mathbf{X}$. The index $t$ in the source signals $s_i(t)$ may denote time, position (especially in digital images), or just the number of the sample vector. For simplicity, we assume here that both the data vector $\mathbf{x}(t) = [x_1(t), x_2(t), \ldots, x_m(t)]^T$ and the source vector $\mathbf{s}(t)$ 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 $t$ can be dropped out, because the order of the data vectors $\mathbf{x}(t)$ is not important and can even be random. This assumption is valid if the data vectors are samples from some multivariate statistical distribution. However, the data vectors $\mathbf{x}(t)$ have often important underlying temporal structure, if they are subsequent samples from a vector-valued time series which is temporally correlated (non-white). Standard ICA can be applied to such time series, too, but it is suboptimal because it does not utilize this temporal information. Alternative methods have been developed for extracting the source signals or independent components in such cases. They usually utilize either temporal autocorrelations directly or assume that the variances of the source signals are nonstationary but smoothly changing; see for example [1], [8], [9], [11].

The application domains and assumptions made in these three major groups of BSS techniques vary to some extent [1], [11]. In standard ICA, it is assumed that all the independent components have non-Gaussian distributions except for possibly one, and they are mutually statistically independent [1]. Then standard ICA methods are able to separate their waveforms, leaving however the order, sign, and scaling of the separated components ambiquous. 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], [13] due to its efficient implementation and fast convergence which makes it applicable to higher dimensional problems, too. We have used in our experiments the freely downloadable FastICA Matlab software package [21].

Methods based on temporal autocorrelations of the source signals require that different sources have at least some different non-zero autocorrelations. Contrary to standard ICA, they can then separate even Gaussian sources, but on the other hand they fail if such temporal autocorrelations do not exist, while standard ICA can even in this case separate non-Gaussian sources. In our experiments the TDSEP method [12] performed best of this type of methods that we have tried. Temporal autocorrelation methods have been reviewed in [15].

BSS methods based on nonstationary smoothly changing variances have been introduced for example in [18], [19]. 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 Chapter 11 in [9]), but we shall not discuss them in this paper.

The authors are with the Department of Information and Computer Science, Aalto Univ. School of Science, P.O. Box 15400, FI-00076 Aalto, Espoo, Finland. Email: {firstname.lastname}@aalto.fi. URL: http://research.ics.tkk.fi/ica/.

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 [14], Hyvärinen developed an approximate method which tries to utilize both higher-order statistics, temporal autocorrelations, and nonstationarity of variances. Only the autocorrelation coefficient corresponding to a single time lag equal to 1 is used there, but the method seems anyway to be able to separate different types of sources. We have used also this method called UniBSS in its Matlab code [22] in our experiments.

ICA and BSS have been generalized into many directions from the simple linear noiseless model (1) [1], [8], [9]. 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 data such data sets, data vectors $\mathbf{y}(t)$ 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

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

as the data vectors $\mathbf{x}(t)$ in (1). The assumptions that we make on the $m_y$-dimensional basis vectors $\mathbf{b}_i$ and source signals $r_i(t)$ are exactly the same as those made on the basis vectors $\mathbf{a}_i$ and source signals $s_i(t)$ 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 canonical correlation analysis 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 correspond to each other, for example being taken at the same time instant.

In our method, we first apply a generalization of canonical correlation analysis (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 the next section.

### B. Related work

The first author considered the problem of finding dependent components from two related data sets already in [2], but the method introduced there suffers from a theoretical weakness. We modified this method and got rid of its weakness in [3]. The method presented in that paper performs much better than plain BSS and ICA methods applied directly to the data sets without using canonical correlation analysis. Not only are the signal-to-noise ratios of the separated sources often clearly higher, but the method is able to separate difficult sources for which plain ICA and BSS methods fail. In the current paper, we generalize this method for more than two data sets, and present a successful real-world application to fMRI data.

In general, the extension of ICA and BSS for separating dependent and independent source signals from 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.

In [17], 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 [20]. The authors of the paper [17] have improved their method in two later papers. In [23], they apply to the results first provided by ICA a nonparametric CCA type model where Gaussian distributions are not assumed. In another more theoretical paper [24] the authors show on a general level how to apply a probabilistic CCA type model without assuming Gaussian distributions.

In [16], 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 paper is largely based on the research papers [29], [28].

Koetsier et al. have presented in [25] 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.

Gutmann and Hyvärinen [27] 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. Our method is more general and does not suffer from such a restrictive model assumption.

Akaho and his co-authors [10] 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.

Furhermore, several authors have developed constrained ICA methods for extracting source signals which are constrained to be similar to some reference signals. This requires, however, some prior knowledge on the reference signals. In [26], Van Hulle introduces three ways to perform constrained ICA. In one of them he tries to find dependent components between two data sets by generalizing CCA, with a small-scale biomedical application.

## II. Canonical correlation analysis

Canonical correlation analysis (CCA) [4], [5] is an old statistical technique which has during the last decade become popular in various signal processing and data analysis applications, because it often provides in practice quite good and meaningful results. Standard CCA measures the linear relationships between two multidimensional datasets $\mathbf{X}$ and $\mathbf{Y}$ using their second-order statistics, autocovariances and cross-covariances. It finds two bases, one for both $\mathbf{X}$ and $\mathbf{Y}$, 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, the dimensions of the data vectors $\mathbf{x} \in \mathbf{X}$ and $\mathbf{y} \in \mathbf{Y}$ can be different, but they are assumed to have zero means. The number of the data vectors in $\mathbf{X}$ and $\mathbf{Y}$ must be the same. The exact conditions required for the canonical correlations and the problem solution are discussed in [4], [5], see also our earlier paper [3]. It turns out these canonical correlations can be computed by solving the eigenvector equations

$$\begin{aligned} \mathbf{C}_{\mathbf{xx}}^{-1}\mathbf{C}_{\mathbf{xy}}\mathbf{C}_{\mathbf{yy}}^{-1}\mathbf{C}_{\mathbf{yx}}\mathbf{w}_{\mathbf{x}} &= \rho^2\mathbf{w}_{\mathbf{x}} \\ \mathbf{C}_{\mathbf{yy}}^{-1}\mathbf{C}_{\mathbf{yx}}\mathbf{C}_{\mathbf{xx}}^{-1}\mathbf{C}_{\mathbf{xy}}\mathbf{w}_{\mathbf{y}} &= \rho^2\mathbf{w}_{\mathbf{y}} \end{aligned} \tag{3}$$

where $\mathbf{C}_{\mathbf{yx}} = \mathrm{E}\{\mathbf{y}\mathbf{x}^T\}$. The eigenvalues $\rho^2$ are squared canonical correlations and the eigenvectors $\mathbf{w}_{\mathbf{x}}$ and $\mathbf{w}_{\mathbf{x}}$ are normalized CCA basis vectors. Only non-zero solutions to these equations are usually of interest, and their number is equal to the smaller of the dimensions of the vectors $\mathbf{x}$ and $\mathbf{y}$.

The solution (3) can be simplified if the data vectors $\mathbf{x}$ and $\mathbf{y}$ are prewhitened [1], which is the usual practice in many ICA and BSS methods. After prewhitening, both $\mathbf{C}_{\mathbf{xx}}$ and $\mathbf{C}_{\mathbf{yy}}$ become unit matrices, and noting that $\mathbf{C}_{\mathbf{yx}} = \mathbf{C}_{\mathbf{xy}}^T$ Eqs. (3) become

$$\begin{aligned} \mathbf{C}_{\mathbf{xy}}\mathbf{C}_{\mathbf{xy}}^T\mathbf{w}_{\mathbf{x}} &= \rho^2\mathbf{w}_{\mathbf{x}} \\ \mathbf{C}_{\mathbf{yx}}\mathbf{C}_{\mathbf{yx}}^T\mathbf{w}_{\mathbf{y}} &= \rho^2\mathbf{w}_{\mathbf{y}} \end{aligned} \tag{4}$$

But these are just the defining equations for the singular value decomposition (SVD) [30] of the cross-covariance matrix $\mathbf{C}_{\mathbf{xy}}$:

$$\mathbf{C}_{\mathbf{xy}} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T = \sum_{i=1}^{L} \rho_i \mathbf{u}_i \mathbf{v}_i^T \tag{5}$$

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 $\mathbf{w}_{xi}$ and $\mathbf{w}_{yi}$ providing canonical correlations. In general, the dimensionalities of the matrices $\mathbf{U}$ and $\mathbf{V}$ and consequently the singular vectors $\mathbf{u}_i$ and $\mathbf{v}_i$ are different corresponding to different dimensions of the data vectors $\mathbf{x}$ and $\mathbf{y}$. The pseudodiagonal matrix

$$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{D} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \tag{6}$$

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 $\mathbf{C}_{\mathbf{xy}}$ has full rank, their number is the smaller one of the dimensions of the data vectors $\mathbf{x}$ and $\mathbf{y}$.

## III. Our method for two related data sets

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

$$\mathbf{v}_{\mathbf{x}} = \mathbf{V}_{\mathbf{x}}\mathbf{x}, \quad \mathbf{v}_{\mathbf{y}} = \mathbf{V}_{\mathbf{y}}\mathbf{y} \tag{7}$$

We use standard principal component analysis (PCA) for computing the whitening matrices $\mathbf{V}_{\mathbf{x}}$ and $\mathbf{V}_{\mathbf{y}}$ as discussed in Section 6.4 in [1]. We then estimate the cross-covariance matrix $\mathbf{C}_{\mathbf{v}_{\mathbf{x}}\mathbf{v}_{\mathbf{y}}}$ of the whitened data vectors $\mathbf{v}_{\mathbf{x}}$ and $\mathbf{v}_{\mathbf{y}}$ in standard manner:

$$\widehat{\mathbf{C}}_{\mathbf{v}_{\mathbf{x}}\mathbf{v}_{\mathbf{y}}} = \frac{1}{N}\sum_{t=1}^{N} \mathbf{v}_{\mathbf{x}}(t)\mathbf{v}_{\mathbf{y}}^T(t) \tag{8}$$

After this, we perform singular value decomposition (SVD) of the estimated cross-covariance matrix $\widehat{\mathbf{C}}_{\mathbf{v}_{\mathbf{x}}\mathbf{v}_{\mathbf{y}}}$ quite similarly as for $\mathbf{C}_{\mathbf{xy}}$ in (5). Inspecting the magnitude of the singular values in the pseudodiagonal matrix $\mathbf{\Sigma}$, we then divide the matrices $\mathbf{U}$ and $\mathbf{V}$ of singular vectors into two submatrices:

$$\mathbf{U} = [\mathbf{U}_1\ \mathbf{U}_2], \quad \mathbf{V} = [\mathbf{V}_1\ \mathbf{V}_2] \tag{9}$$

There $\mathbf{U}_1$ and $\mathbf{V}_1$ correspond to dependent components for which the respective singular values are larger than $0.5$, and $\mathbf{U}_2$ and $\mathbf{V}_2$ to the independent components for which the respective singular values are smaller. We have found experimentally that the threshold value $0.5$ is suitable. The data are then projected using these submatrices into subspaces corresponding to the dependent and independent components by computing

$$\mathbf{U}_1^T\mathbf{X}, \quad \mathbf{U}_2^T\mathbf{X}, \quad \mathbf{V}_1^T\mathbf{Y}, \quad \mathbf{V}_2^T\mathbf{Y} \tag{10}$$

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 contrary to the customary use of SVD 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 largely explain its good performance.

First, let us denote the separating matrices after the whitening step in (7) by $\mathbf{W}_{\mathbf{x}}^T$ for $\mathbf{v}_{\mathbf{x}}$ and respectively by $\mathbf{W}_{\mathbf{y}}^T$ for $\mathbf{v}_{\mathbf{y}}$. A basic result in the theory of ICA and BSS [1] is that after whitening the separating matrices $\mathbf{W}_{\mathbf{x}}$ and $\mathbf{W}_{\mathbf{y}}$ become orthogonal: $\mathbf{W}_{\mathbf{x}}^T\mathbf{W}_{\mathbf{x}} = \mathbf{I}$, $\mathbf{W}_{\mathbf{y}}^T\mathbf{W}_{\mathbf{y}} = \mathbf{I}$. Thus

$$\widehat{\mathbf{s}} = \mathbf{W}_{\mathbf{x}}^T\mathbf{V}_{\mathbf{x}}\mathbf{x} = \mathbf{W}_{\mathbf{x}}^T\mathbf{V}_{\mathbf{x}}\mathbf{A}\mathbf{s} = \mathbf{P}_{\mathbf{s}}\mathbf{D}_{\mathbf{s}}\mathbf{s} \tag{11}$$

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. (11) 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 $\mathbf{D_s s}$ [34].

Assuming that there are as many linearly independent mixtures $\mathbf{x}$ 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 (11)

$$\mathbf{A} = (\mathbf{W}_\mathbf{x}^T \mathbf{V}_\mathbf{x})^{-1} \mathbf{P_s D_s} = \mathbf{V}_\mathbf{x}^{-1} \mathbf{W_x P_s D_s} \tag{12}$$

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

$$\mathbf{B} = (\mathbf{W}_\mathbf{y}^T \mathbf{V}_\mathbf{y})^{-1} \mathbf{P_r D_r} = \mathbf{V}_\mathbf{y}^{-1} \mathbf{W_y P_r D_r} \tag{13}$$

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

$$\mathbf{C}_{\mathbf{v_x v_y}} = \mathbf{V_x} \mathrm{E}\{\mathbf{x}\mathbf{y}^T\} \mathbf{V}_\mathbf{y}^T = \mathbf{V_x A Q B}^T \mathbf{V}_\mathbf{y}^T \tag{14}$$

Here the matrix $\mathbf{Q} = \mathrm{E}\{\mathbf{s}\mathbf{r}^T\}$ 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. (12) and (13) into (14) yields

$$\mathbf{C}_{\mathbf{v_x v_y}} = (\mathbf{W_x P_s})(\mathbf{D_s Q D}_\mathbf{r}^T)(\mathbf{W_y P_r})^T \tag{15}$$

But this is exactly the same type of expansion as the singular value decomposition (5) of the whitened cross-covariance matrix $\mathbf{C}_{\mathbf{v_x v_y}}$. 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 $\mathbf{W_x}$ [34]. 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 (5), and quite similarly the orthogonal matrix $\mathbf{W_y P_r}$ corresponds to the orthogonal matrix $\mathbf{V}$ in (5). 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 (5).

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 in [3]. At least CCA when applied to the data sets $\mathbf{X}$ and $\mathbf{Y}$ using (10) 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 (10) probably helps. Solving two lower dimensional subproblems is easier than solving a higher dimensional separation problem. And if the mixtures consist of several types of sources, which could be super-Gaussian, sub-Gaussian, Gaussian, temporally correlated, or nonstationary sources, the complexity of the sources in the subproblems to be solved can be reduced.

We can somewhat heuristically modify the SVD based method introduced above to include temporal correlations into the computations by using instead of the plain cross-covariance matrix $\mathbf{C}_{\mathbf{v_x v_y}} = \mathrm{E}\{\mathbf{v_x v}_\mathbf{y}^T\}$ the generalized cross-covariance matrices

$$\mathbf{G}_{\mathbf{v_x v_y}} = \mathrm{E}\{\mathbf{v_x}(t)\mathbf{v}_\mathbf{y}^T(t) + \mathbf{v_x}(t-d)\mathbf{v}_\mathbf{y}^T(t) + \mathbf{v_x}(t)\mathbf{v}_\mathbf{y}^T(t-d)\} \tag{16}$$

where $d$ is the chosen time delay. In our experiments, we have found that a suitably chosen time delay $d$ in (16) can improve the separation results for temporally correlated sources.

## IV. Extension to several data sets

In a pioneering paper [31], 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 [31], [32]. It can be optimized and the respective canonical vectors estimated using the procedure described in [31], [32]. This optimization method is, however, computationally somewhat complicated. It requires first 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

$$L = \sum_{k=1}^{M} m_k \tag{17}$$

is the sum of the dimensionalies 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 [32] 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 by $\mathbf{h}_k^{(i)}$ and canonical variables by $\mathbf{z}_k^{(i)} = \mathbf{X}_k^T \mathbf{h}_k^{(i)}$, and the estimated cross-correlation matrices[1] as $\mathbf{C}_{kl} = \mathbf{X}_k \mathbf{X}_l^T$.

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

$$\rho^{(i)} = \frac{1}{M} \sum_{k=1}^{M} \rho_k^{(i)} \tag{18}$$

where

$$\rho_k^{(i)} = \frac{1}{M-1} \sum_{l=1, l \neq k}^{M} \rho_{kl}^{(i)} \tag{19}$$

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 [32]

$$\frac{1}{M} \sum_{k=1}^{M} \mathbf{h}_k^{(i)T} \mathbf{C}_{kk} \mathbf{h}_k^{(i)} = 1 \tag{20}$$

The orthogonality constraints are for $i \neq j$

$$\mathbf{z}^{(i)T} \mathbf{z}^{(j)} = 0 \tag{21}$$

$$\mathbf{z}^{(i)} = \frac{1}{M} \sum_{k=1}^{M} \mathbf{z}_k^{(i)}. \tag{22}$$

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

$$\begin{aligned} J^{(i)} &= \frac{1}{2M(M-1)} \sum_{k,l=1}^{M} \|\mathbf{X}_k \mathbf{h}_k^{(i)} - \mathbf{X}_l \mathbf{h}_l^{(i)}\|^2 \\ &= \frac{1}{M} \sum_{k=1}^{M} \| \mathbf{z}_k^{(i)} \|^2 - \rho^{(i)} \end{aligned} \tag{23}$$

subject to the constraints (20) and (21), which implies $J^{(i)}$ = $1 - \rho^{(i)}$.

The solutions of this generalized CCA problem can be obtained using the method of Lagrange multipliers [32]. This leads to the generalized eigenvector problem

$$\frac{1}{M-1} (\mathbf{C} - \mathbf{D}) \mathbf{h}^{(i)} = \rho^{(i)} \mathbf{D} \mathbf{h}^{(i)} \tag{24}$$

where

$$\mathbf{h}^{(i)} = [\mathbf{h}_1^{(i)T}, \mathbf{h}_2^{(i)T}, \ldots, \mathbf{h}_M^{(i)T}]^T \tag{25}$$

is a "supervector" formed by stacking the $i$:th canonical vectors of the M data sets, and the respective block matrices are

$$\mathbf{C} = \begin{bmatrix} \mathbf{C}_{11} & \ldots & \mathbf{C}_{1M} \\ \vdots & \ddots & \vdots \\ \mathbf{C}_{M1} & \ldots & \mathbf{C}_{MM} \end{bmatrix} \tag{26}$$

$$\mathbf{D} = \begin{bmatrix} \mathbf{C}_{11} & \ldots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \ldots & \mathbf{C}_{MM} \end{bmatrix} \tag{27}$$

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 (24). These eigenvectors can be computed also using a deflation type neural recursive least-squares algorithm introduced and discussed in [32].

A couple of notes are in order here. First, the equations (3) defining standard CCA for two data sets can be written in the form (24) after some manipulation, see [32], [5]. Then in (24) $\mathbf{h}(i)$ = $[\mathbf{w}_{xi}^T, \mathbf{w}_{yi}^T]^T$. If we denote the matrix on the left-hand side of (24) by $\mathbf{O}$ (off-diagonal), (24) is equivalent to the non-symmetric eigenproblem

$$\mathbf{D}^{-1} \mathbf{O} \mathbf{h}^{(i)} = \rho^{(i)} \mathbf{h}^{(i)} \tag{28}$$

which could in principle have complex-valued eigenvectors and -values. However, the equation (28) can be written as

$$\mathbf{O}^{1/2} \mathbf{D}^{-1} \mathbf{O}^{1/2} (\mathbf{O}^{1/2} \mathbf{h}^{(i)}) = \rho^{(i)} (\mathbf{O}^{1/2} \mathbf{h}^{(i)}) \tag{29}$$

which is a symmetric eigenproblem for the eigenvector $\mathbf{O}^{1/2} \mathbf{h}^{(i)}$. Hence the eigenvalues and -vectors of (24) are real-valued.

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 $\mathbf{C}_{kl}$, $k, l = 1, \ldots, M$ similarly as in (8) and form from them 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 (24) or (28). 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

$$\mathbf{P}_{D,k} = [\mathbf{g}_k^{(1)}, \ldots, \mathbf{g}_k^{(d)}] \tag{30}$$

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,

$$\mathbf{P}_{D,k}^T \mathbf{X}_k, \qquad k = 1, \ldots, M \tag{31}$$

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

and the dependent components (sources) of each data set are found by applying any suitable ICA or BSS method to the projected data sets (31).

A question now arises how to estimate the independent components (sources) in each data set. A first idea is to use the generalized eigenvectors corresponding to the smallest eigenvalues in a similar manner as above. However, if we have for example 3 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 the equation (28) 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 (30) 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 (30) and each other. The resulting vectors are used to define a projection operator

$$\mathbf{P}_{I,k} = [\mathbf{g}_k^{(d+1)}, \ldots, \mathbf{g}_k^{(m_k)}] \qquad (32)$$

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

$$\mathbf{P}_{I,k}^T \mathbf{X}_k, \qquad k = 1, \ldots, M \qquad (33)$$

and the independent components are estimated by applying any suitable ICA or BSS method to the projected data sets (33).

## V. Experimental results

### A. Simulated data

Experiments with synthetically generated data are useful and necessary, because the true source signals are known. It is then possible to assess the performance of the methods using a suitable criterion. For real-world data, the true sources are usually unknown, and the results can be assessed qualitatively only.

TABLE I
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 [27] | 25.0 | 27.1 | 6.9 | 6.7 | 24.7 |
| Method in [29] | 6.2 | 5.8 | 6.2 | 6.1 | 4.9 |

We used the 6 source signals defined in the Matlab code UniBSS.m [22] and explained in [14]. 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. The first and third source had identical temporal autocovariances, and similarly the second and fourth source. The fifth and sixth source have smoothly changing variances. Furthermore, we generated 3 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. Only the approximative UniBSS method should be able to separate all of them [14].

From these 9 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 3 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.

Because the results can vary a lot for different statistical 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. The signal-to-noise ratios (SNR's) of the estimated source signals were computed for each realization of the data sets and each source from the formula

$$\text{SNR}(i) = 10 \log_{10} \frac{\frac{1}{N}\sum_{t=1}^{N} s_i(t)^2}{\frac{1}{N}\sum_{t=1}^{N} [s_i(t) - \hat{s}_i(t)]^2} \qquad (34)$$

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 $\hat{s}_i(t)$.

TABLE II
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 [27] | 25.1 | 28.6 | 17.5 | 21.2 | 24.9 |
| Method in [29] | 6.2 | 5.8 | 2.5 | 2.3 | 4.9 |

The results for the data sets $\mathbf{X}_1$, $\mathbf{X}_2$, and $\mathbf{X}_3$ are presented in Tables I, II, and III, respectively. Based on visual inspection, we set the border value of SNR for a successful separation to 10 dB. Even an SNR of a few decibels means in practice progress towards separation, often considerable. In this case, some parts of the respective source signals are often well separated while others not. Poor results with no visible separation have typically an SNR value around 0 dB.

TABLE III
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 [27] | 24.6 | 28.6 | 9.9 | 10.2 | 24.5 |
| Method in [29] | 6.2 | 5.8 | 8.8 | 9.2 | 4.9 |

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 3rd and 4th source are around the separation border already. FastICA [1], [13], [21], 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 [12] based on temporal autocorrelations is able to marginally separate all the 5 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 [14], [22] 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 the first autocorrelation.

Preprocessing using generalized canonical correlation analysis (GCCA) improves the separation results for most sources and all 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 in our earlier paper [3] using plain CCA preprocessing for two data sets and the FastICA and UniBSS methods.

In this paper, we also compare our method with two methods introduced by other authors for the same problem. The first compared method [27] assumes that the dependent sources in the data sets are active simultaneously. From Tables I-III one can see that it performs quite well for the dependent first, second, and fifth source 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 [29] 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.

### B. Real-world fMRI data

We tested the usefulness of our method with data from a functional magnetic resonance imaging (fMRI) study [7], 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 [6] for post-processing the results given by CCA, because it gave better results than FastICA. All the data were acquired at the Advanced Magnetic Imaging Centre of Aalto University, using a 3.0 Tesla MRI scanner (Signa EXCITE 3.0 T; GE Healthcare, Chalfont St. Giles, UK).

Figures 1 and 2 show the results of applying our method to the two datasets and separating 11 components from the dependent subspaces $\mathbf{U1}$ and $\mathbf{V1}$. 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 [7]. 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.

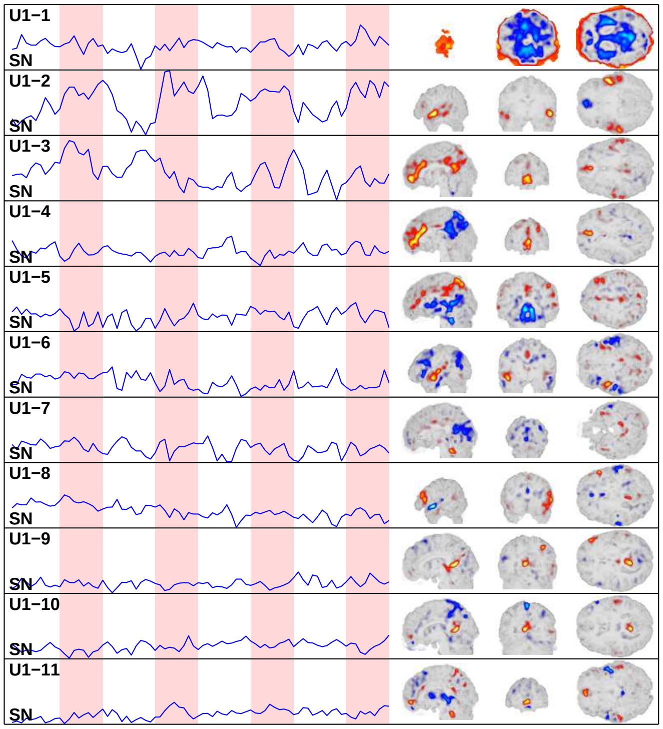

Fig. 1. Experimental results with fMRI data. Each row shows one of the 11 separated components. The activation time-course with the stimulation blocks for reference, shown on the left, and the corresponding spatial pattern on three coincident slices, on the right. Components from the first dataset.

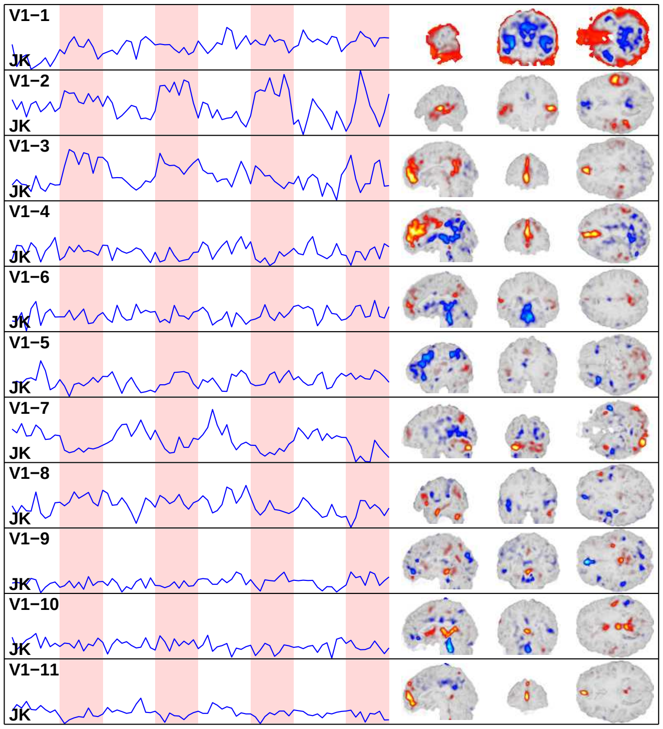

Fig. 2. Experimental results with fMRI data. Components from the second dataset.

## VI. Discussion

After writing the paper [3], we tested our method for two data sets with several other methods than FastICA and UniBSS. The results were good especially for the TDSEP method, and CCA prepocessing improved them also for the well-known algebraic ICA method JADE [33], which is based on non-Gaussianity included into computations explicitly by higher-order statistics. However, the results of the CCA followed by JADE method were not as good as for FastICA, TDSEP, and UniBSS. We tested several other ICA and BSS methods, too, and found that if a method fails completely in a separation task providing results around 0 dB, CCA preprocessing does not any more help it to achieve better results.

Even though the UniBSS method performed well in these experiments, it has some drawbacks. First, it requires at least of the order of 1000 samples to perform appropriately, while for example FastICA needs less samples for 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 move 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 well-known the natural gradient algorithm [1], [8], 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 [20]. 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 (26) and (27). 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 two related data sets are retained in mapping them to the subspaces corresponding to their dependent and independent components in (31) and (33).

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.

## VII. Conclusions

In this paper, we have introduced a method based on least-squares generalization of standard canonical correlation analysis (CCA) 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 this generalization of CCA 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 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.

**Acknowledgements** This work was supported by the Adaptive Informatics Research Centre of Excellence of the Academy of Finland.

## References

[1] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis.* Wiley, 2001.

[2] J. Karhunen and T. Ukkonen, “Extending ICA for finding jointly dependent components from two related data sets,” *Neurocomputing*, vol. 70, pp. 2969–2979, 2007.

[3] J. Karhunen and T. Hao, “Finding dependent and independent components from two related data sets”, in *Proc. of Int. J. Conf. on Neural Networks (IJCNN 2011)*, San Jose, California, USA, August 2011, pp. 457–466.

[4] A. Rencher, *Methods of Multivariate Analysis, 2nd ed.*, Wiley, 2002.

[5] M. Borga, “Canonical correlation: a tutorial”, Linköping University, Linköping, Sweden, 2001, 12 pages. Available at http://www.imt.liu.se/∼magnus/cca/tutorial/.

[6] L. Wiskott and T. Sejnowski, “Slow feature analysis: unsupervised learning of invariances”, *Neural Computation*, Vol. 14, pp. 715–770, 2002.

[7] J. Ylipaavalniemi and R. Vigario, “Analyzing consistency of independent components: An fMRI illustration”, *NeuroImage*, vol. 39, 2008, pp. 169–180.

[8] A. Cichocki and S.-I.Amari, *Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications*, Wiley, 2002.

[9] P. Comon and C. Jutten (Eds.), *Handbook of Blind Source Separation: Independent Component Analysis and Applications*, Academic Press, 2010.

[10] S. Akaho, Y. Kiuchi, and S. Umeyama, ”MICA: Multidimensional Independent Component Analysis,” in *Proc. of the 1999 Int. Joint Conf. on Neural Networks (IJCNN’99)*, Washington, DC, USA, July 1999. IEEE Press, 1999, pp. 927–932.

[11] J.-F. Cardoso, ”The three easy routes to independent component analysis; contrasts and geometry”, in *Proc. of the 3rd Int. Conf. on Independent Component Analysis and Blind Signal Separation (ICA2001)*, San Diego, California, USA, December 2001, pp. 1–6.

[12] A. Ziehe and K.-R. Müller, “TDSEP - an efficient algorithm for blind source separation using time structure,” in *Proc. of the Int. Conf. on Artificial Neural Networks (ICANN’98)*, Skövde, Sweden, 1998.

[13] A. Hyvärinen, “Fast and robust fixed-point algorithms for independent component analysis,” *IEEE Trans. on Neural Networks*, vol. 10, no. 3, pp. 626–634, 1999.

[14] A. Hyvärinen, “A unifying model for blind separation of independent sources,“ *Signal Processing*, vol. 85, no. 7, pp. 1419–1427, 2005.

[15] A. Yeredor, “Second-order methods based on color”. Chapter 7 in P. Comon and C. Jutten (Eds.), *Handbook of Blind Source Separation: Independent Component Analysis and Applications*, Academic Press, 2010, pp. 227–279.

[16] N. Correa, T. Adali, Y.-Q. Li, and V. Calhoun, “Canonical correlation analysis for data fusion and group inferences”, *IEEE Signal Processing Magazine*, vol. 27, no. 4, July 2010, pp. 39–50.

[17] J. Ylipaavalniemi et al., “Dependencies between stimuli and spatially independent fMRI sources: towards brain correlates of natural stimuli”, *NeuroImage*, vol. 48, 2009, pp. 176–185.

[18] D.-T. Pham and J.-F, Cardoso, “Blind separation of instantaneous mixtures of non stationary sources”, *IEEE Trans. on Signal Processing*, vol. 49, no. 9, 1837–1848, 2001.

[19] A. Hyvärinen, “Blind source separation by nonstationarity of variance: a cumulant-based approach,” *IEEE Trans. on Neural Networks*, vol. 12, no. 6, pp. 1471-1474, 2001.

[20] F. Bach and M. Jordan, “A probabilistic interpretation of canonical correlation analysis”. Technical Report 688, Dept. of Statistics, Univ. of California, Berkeley, CA, USA, 2005. Available at http://www.di.ens.fr/∼fbach/ .

[21] A. Hyvärinen et al., “The FastICA package for Matlab”, Helsinki Univ. of Technology, Espoo, Finland, 2005. Available at http://research.ics.tkk.fi/ica/fastica/ .

[22] A. Hyvärinen, “Basic Matlab code for the unifying model for BSS”, Univ. of Helsinki, Dept. of Mathematics and Statistics and Dept. of Computer Science, Helsinki, Finland, 2003–2006. Available at http://www.cs.helsinki.fi/u/ahyvarin/code/UniBSS.m .

[23] E. Savia, A. Klami, and S. Kaski, “Fast dependent components for fMRI analysis”, in *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP’09)*, Taipei, Taiwan, April 2009, pp. 1737–1740.

[24] A. Klami, S. Virtanen, and S. Kaski, “Bayesian exponential family projections for coupled data sources”, in P. Grunwald and P. Spirtes (Eds.), *Proc. of the 26th Conf. on Uncertainty in Artificial Intelligence (UAI 2010)*, Catalina Island, California, USA, July 2010. AUAI Press, Corvallis, Oregon, USA, 2010, pp. 286–293.

[25] J. Koetsier, D. MacDonald, D. Charles, and C. Fyfe, “Exploratory correlation analysis”, in *Proc. of the 10th European Symposium on Artificial Neural Networks (ESANN2002)*, Bruges, Belgium, April 2002, pp. 483–488.

[26] M. Van Hulle, ”Constrained subspace ICA based on mutual information optimization directly”, *Neural Computation*, vol. 20, no. 4, 2008, pp. 964–973.

[27] M. Gutmann and A. Hyvärinen, “Extracting coactivated features from multiple data sets”, in T. Honkela et al., *Lecture Notes in Computer Science*, Vol. 6791 (Proc. of ICANN2011, Espoo, Finland), pp. 323–330, Springer, 2011.

[28] M. Anderson, X.-L. Li, and T. Adali, “Nonorthogonal independent vector analysis using multivariate Gaussian model”, in V. Vigneron et al. (Eds.), *Lecture Notes in Computer Science*, Vol. 6365 (Proc. of LVA/IVA 2010, St. Malo, France), pp. 354–361, Springer, 2010.

[29] Y.-Q. Li, T. Adali, W. Wang, and V. Calhoun, “Joint blind source separation by multiset canonical correlation analysis”, *IEEE Trans. on Signal Processing*, Vol. 57, No. 10, October 2009, pp. 3918–3928.

[30] S. Haykin, *Modern Filters.* MacMillan, 1989.

[31] J. Kettenring, “Canonical analysis of several sets of variables”, *Biometrika*, Vol. 58, No. 3, December 1971, pp. 433–451. Available in electronic form at http://www.jstor.org/stable/2334380 .

[32] J. Via, I. Santamaria, and J. Perez, “A learning algorithm for adaptive canonical correlation analysis of several data sets”, *Neural Networks*, Vol. 20, 2007, pp. 139–152.

[33] J. Cardoso and A. Souloumiac, “Blind beamforming for non Gaussian signals”, *IEE Proceedins-F*, Vol. 140, No. 6, pp. 362–370, 1993.

[34] Wikipedia, the free encyclopedia. Articles “Orthogonal matrix” and “Permutation matrix”. http://en.wikipedia.org/wiki/ .