A Canonical Correlation Analysis Based Method for Improving BSS of Two Related Data Sets

Juha Karhunen, Tele Hao, and Jarkko Ylipaavalniemi

Dept. of Information and Computer Science, Aalto University, School of Science.
P.O. Box 15400, FI-00076 Aalto, Espoo, Finland.
Email: `first.last@aalto.fi` URL: `http://ics.tkk.fi/en/`

Abstract. We consider an extension of ICA and BSS for separating mutually dependent and independent components from two related data sets. We propose a new method which first uses canonical correlation analysis for detecting subspaces of independent and dependent components. Different ICA and BSS methods can after this be used for final separation of these components. Our method has a sound theoretical basis, and it is straightforward to implement and computationally not demanding. Experimental results on synthetic and real-world fMRI data sets demonstrate its good performance.

1 Introduction

Various independent component analysis (ICA) and blind source separation (BSS) methods [1, 2] are nowadays well-known techniques for blind extraction of useful information from single vector-valued data $\mathbf{X} = [\mathbf{x}(1), \ldots, \mathbf{x}(N_x)]$ with many applications. The data model used in the basic linear ICA is simply

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

Thus each data vector $\mathbf{x}(t) = [x_1(t), x_2(t), \ldots, x_n(t)]^T$ is expressed as a linear combination of independent components or source signals $s_i(t)$, collected respectively to the source vector $\mathbf{s}(t) = [s_1(t), s_2(t), \ldots, s_n(t)]^T$. For simplicity, we first assume that both $\mathbf{x}(t)$ and $\mathbf{s}(t)$ are zero mean n-vectors, and that the mixing matrix $\mathbf{A}$ is a full-rank constant $n \times n$ matrix with column vectors $\mathbf{a}_i$, $i = 1, 2, \ldots, n$.

In standard linear ICA, the index t which usually denotes time or sample index is not important, because the order of the data vectors $\mathbf{x}(t)$ can be arbitrary. This holds if they are samples from some multivariate statistical distribution. However, the data vectors $\mathbf{x}(t)$ have often important underlying temporal structure. Alternative BSS methods have been developed for utilizing such temporal information. They usually utilize either temporal autocorrelations directly or smoothly changing nonstationarity of variances. The assumptions and applications domains of these three major categories of methods based on the simple model (1) vary somewhat [1, 2].

The most widely used standard ICA method is currently FastICA [1, 3] due to its efficient implementation and fast convergence which makes it applicable to higher dimensional problems, too. From the many methods using temporal autocorrelations, we have used the TDSEP method [4] which performs usually well. 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 [5], an approximate method called UbiBSS is developed which tries to utilize higher-order statistics, temporal autocorrelations, and nonstationarity of variances. We have used its Matlab code [6] in our experiments.

ICA and BSS have been generalized into many directions from the simple linear noiseless model (1) [1, 2]. We consider a generalization in which one tries to find out mutually dependent and independent components from two different but related data sets $\mathbf{X}$ and $\mathbf{Y} = [\mathbf{y}(1), \ldots, \mathbf{y}(N_y)]$. Data vectors $\mathbf{y}(t)$ have dimension m which can be different from dimension n of the data vectors $\mathbf{x}(t)$ in $\mathbf{X}$, but they obey a similar basic linear model

$$\mathbf{y}(t) = \mathbf{Br}(t) = \sum_{i=1}^{m} r_i(t)\mathbf{b}_i \qquad (2)$$

in which $\mathbf{r}(t)$ is m-vector and $\mathbf{B}$ $m \times m$ matrix.

This generalization of ICA and BSS has not been studied as much as several others, but some related work can be found in [7–9, 11–13]. In most of these methods the data model is more rectrictive than ours, assuming that in the data sets $\mathbf{X}$ and $\mathbf{Y}$ there exist pairs of sources which are mutually dependent, but these sources are independent of all the other sources in $\mathbf{X}$ and $\mathbf{Y}$. In particular, canonical correlation analysis (CCA) or its extension to multiple data sets is applied in [11, 13], but in a different way than we do. Due to space limitations, we do not discuss these related works in more detail here.

2 Our method

We apply canonical correlation analysis (CCA) to find the subspaces of dependent and independent sources in the two related data sets. CCA [14] is an old statistical technique which measures the linear relationships between two multidimensional datasets $\mathbf{X}$ and $\mathbf{Y}$ using their autocovariances and cross-covariances. CCA 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 canonical correlations and the respective basis vectors can be computed by solving a generalized eigenvalue problem as discussed in [14]. This solution simplifies considerably if the data vectors $\mathbf{x}$ and $\mathbf{y}$ are prewhitened [1]. It turns out that the basis vectors of CCA can then be determined from the singular value decomposition (SVD) of the cross-covariance matrix $\mathbf{C}_{\mathbf{xy}} = \mathrm{E}\{\mathbf{xy}^T\}$ of $\mathbf{x}$ and $\mathbf{y}$:

$$\mathbf{C}_{\mathbf{xy}} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T = \sum_{i=1}^{L} \rho_i \mathbf{u}_i \mathbf{v}_i^T \tag{3}$$

Note that the SVD of $\mathbf{C}_{\mathbf{yx}} = \mathrm{E}\{\mathbf{yx}^T\} = \mathbf{C}_{\mathbf{xy}}^T$ is quite similar and is obtained by transposing both sides of Eq. (3). There $\mathbf{U}$ and $\mathbf{V}$ are two orthogonal square matrices ($\mathbf{U}^T\mathbf{U} = \mathbf{I}$, $\mathbf{V}^T\mathbf{V} = \mathbf{I}$) containing as their column vectors the singular vectors $\mathbf{u}_i$ and $\mathbf{v}_j$. In our case, these singular vectors are the basis vectors 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

$$\boldsymbol{\Sigma} = \begin{bmatrix} \mathbf{D} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \tag{4}$$

consists of a diagonal matrix $\mathbf{D}$ containing the non-zero singular values appended with zero matrices so that the matrix $\boldsymbol{\Sigma}$ 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 L is the smaller one of the dimensions of the data vectors $\mathbf{x}$ and $\mathbf{y}$.

We first make the data vectors $\mathbf{x} \in \mathbf{X}$ zero mean if necessary. These data vectors are whitened separately:

$$\mathbf{v}_{\mathbf{x}} = \mathbf{V}_{\mathbf{x}}\mathbf{x}, \qquad \mathbf{v}_{\mathbf{y}} = \mathbf{V}_{\mathbf{y}}\mathbf{y} \tag{5}$$

We use standard principal component analysis (PCA) for whitening as discussed in [1]. After this we estimate the cross-covariance matrix $\mathbf{C}_{\mathbf{v_x v_y}}$ of the whitened data vectors $\mathbf{v}_{\mathbf{x}}$ and $\mathbf{v}_{\mathbf{y}}$ in standard manner:

$$\widehat{\mathbf{C}}_{\mathbf{v_x v_y}} = \frac{1}{N} \sum_{t=1}^{N} \mathbf{v}_{\mathbf{x}}(t)\mathbf{v}_{\mathbf{y}}^T(t) \tag{6}$$

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 the SVD of the estimated cross-covariance matrix $\widehat{\mathbf{C}}_{\mathbf{v_x v_y}}$ quite similarly as for $\mathbf{C}_{\mathbf{xy}}$ in (3). After inspecting the magnitudes of the singular values in the pseudodiagonal matrix $\boldsymbol{\Sigma}$, we divide the matrices $\mathbf{U}$ and $\mathbf{V}$ of singular vectors into two submatrices:

$$\mathbf{U} = [\mathbf{U}_1\,\mathbf{U}_2], \qquad \mathbf{V} = [\mathbf{V}_1\,\mathbf{V}_2] \tag{7}$$

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 small. The data are then mapped using these submatrices onto subspaces corresponding to the dependent and independent components by computing

$$\mathbf{U}_1^T\mathbf{X}, \quad \mathbf{U}_2^T\mathbf{X}, \quad \mathbf{V}_1^T\mathbf{Y}, \quad \mathbf{V}_2^T\mathbf{Y} \tag{8}$$

where $\mathbf{X} = [\mathbf{x}(1), \ldots, \mathbf{x}(N_x)]$ and $\mathbf{Y} = [\mathbf{y}(1), \ldots, \mathbf{y}(N_y)]$. 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 $\mathbf{Y}$. We are not aware that CCA would have used in this way in ICA and BSS previously.

Sometimes CCA alone used in this way is sufficient for coarse separation of sources, but in most cases CCA at least makes clear progress towards separation, providing signal-to-noise ratios of a few decibels. The preliminary separation results of CCA can often be improved by applying to the four mapped data sets defined in (8) some suitable ICA or BSS method. In principle at least it is possible to apply any kind of postprocessing here.

The somewhat surprising result than CCA alone can provide coarse separation can be justfied heuristically as follows. First, let us denote the separating matrices after the whitening step in (5) 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

$$\widehat{\mathbf{s}} = \mathbf{W}_\mathbf{x}^T\mathbf{V}_\mathbf{x}\mathbf{x} = \mathbf{W}_\mathbf{x}^T\mathbf{V}_\mathbf{x}\mathbf{A}\mathbf{s} = \mathbf{s} \tag{9}$$

where we have for simplicity assumed that the estimated sources $\widehat{\mathbf{s}}$ appear in the same order as the original sources $\mathbf{s}$. Assuming that there are as many linearly independent mixtures $\mathbf{x}$ and $\mathbf{W}_\mathbf{y}$ as sources $\mathbf{s}$, so that the mixing matrix $\mathbf{A}$ is a full-rank square matrix, we get from (9) by setting $\widehat{\mathbf{s}} = \mathbf{s}$

$$\mathbf{A} = (\mathbf{W}_\mathbf{x}^T\mathbf{V}_\mathbf{x})^{-1} = \mathbf{V}_\mathbf{x}^{-1}\mathbf{W}_\mathbf{x} \tag{10}$$

due to the orthogonality of the matrix $\mathbf{W}_\mathbf{x}$. Quite similarly, we get for the another mixing matrix $\mathbf{B}$ in (2) the equivalent result $\mathbf{B} = \mathbf{V}_\mathbf{y}^{-1}\mathbf{W}_\mathbf{y}$.

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

$$\mathbf{C}_{\mathbf{v}_\mathbf{x}\mathbf{v}_\mathbf{y}} = \mathrm{E}\{\mathbf{v}_\mathbf{x}\mathbf{v}_\mathbf{y}^T\} = \mathbf{V}_\mathbf{x}\mathrm{E}\{\mathbf{x}\mathbf{y}\}\mathbf{V}_\mathbf{y}^T = \mathbf{V}_\mathbf{x}\mathbf{A}\mathbf{Q}\mathbf{B}^T\mathbf{V}_\mathbf{y}^T \tag{11}$$

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 $\mathbf{A} = \mathbf{V}_\mathbf{x}^{-\mathbf{1}}\mathbf{W}_\mathbf{x}$ and $\mathbf{B} = \mathbf{V}_\mathbf{x}^{-\mathbf{1}}\mathbf{W}_\mathbf{y}$ into (11) yields finally

$$\mathbf{C}_{\mathbf{v}_\mathbf{x}\mathbf{v}_\mathbf{y}} = \mathbf{W}_\mathbf{x}\mathbf{Q}\mathbf{W}_\mathbf{y}^T \tag{12}$$

But this is exactly the same type of expansion as the SVD of the whitened cross-covariance matrix $\mathbf{C}_{\mathbf{v}_\mathbf{x}\mathbf{v}_\mathbf{y}}$ in (3), because the matrices $\mathbf{W}_\mathbf{x}$ and $\mathbf{W}_\mathbf{y}$ are orthogonal matrices and $\mathbf{Q}$ is a diagonal matrix. 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 the next section.

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. Our method combines both types of statistics. Our experimental results demonstrate that this often provide better results than using solely second-order or higher-statistics for separation. Dividing the separation problem into subproblems using the matrices in (8) may also help. Probably solving two lower dimensional subproblems is easier than solving a higher dimensional separation problem.

3 Experimental results

We have successfully tested our method with synthetical data sets, with data sets in which real-world sources have been mixed synthetically, and with real-world robot and fMRI (functional magnetic resonance imaging) data. Due to space limitations, we can show some quite selected results only here. More experimental results can be found in [16].

Consider first a set of 6 synthetical stochastic sources which have been purposedly designed so that they are very difficult to separate for most ICA and BSS methods. They are defined in the Matlab code [6] of the UniBSS method and explained in the respective paper [5]. Standard ICA methods based on non-Gaussianity should be able to separate only the two first sources. Methods based on temporal statistics should not able to separate any of them. Method utilizing smoothly changing variances are able to separate only the fifth and sixth source. Only the approximative UniBSS method [5] which utilizes all these properties is able to separate all these 6 sources.

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 [9]	25.7	9.8	9.4	23.1
Method in [13]	12.5	11.4	11.3	13.2

Table 1. Signal-to-noise ratios (dB) of different methods for the source signals 1-4 in the first data set $\mathbf{X}$.

We mixed the first three sources and the fifth one to form the first data set $\mathbf{X}$, and the second, third, fourth and sixth source to the second data set $\mathbf{Y}$. Thus in these data sets there are two completely dependent sources, while

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 [9]	9.8	9.4	9.5	9.5
Method in [13]	11.4	11.3	3.6	3.9

Table 2. Signal-to-noise ratios (dB) of different methods for the source signals 5-8 in the second data set $\mathbf{Y}$.

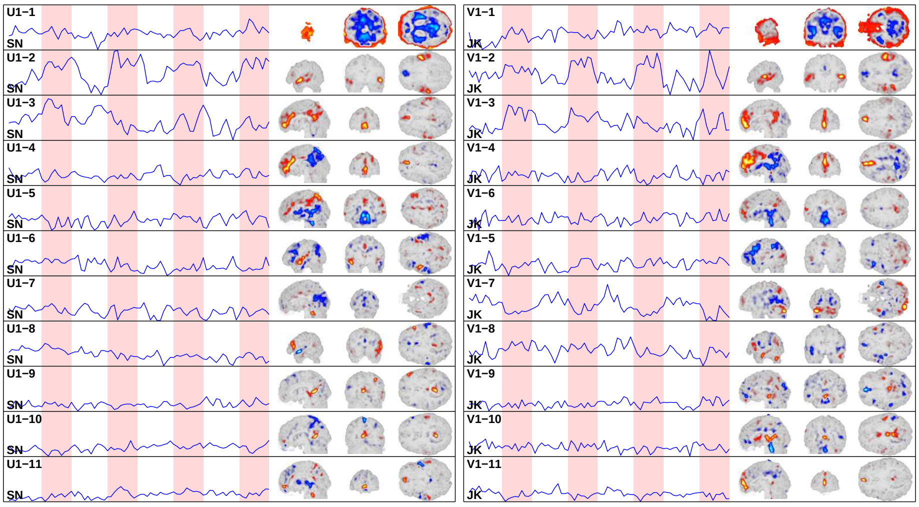
the remaining two sources in them are statistically independent of all the other sources. We used 5000 data vectors and source signal values ($t = 1, 2, \ldots, 5000$) for providing enough data to the UniBSS method [5]. The other tested methods, CCA, FastICA, TDSEP and their combinations require less samples, especially CCA. We computed the average signal-to-noise ratios of the estimated sources over 100 random realizations of the sources and the data sets $\mathbf{X}$ and $\mathbf{Y}$ because the results vary for single realizations. In each realization, the elements of the 4×4 mixing matrices were Gaussian random numbers.

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 [9] assumes that the dependent sources in the two data sets are active simultaneously. The second compared method [13] 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 [12].

The separation results for the four sources 1-4 contained in the first data set $\mathbf{X}$ are shown in Table 1, and for the 4 sources in the other data set $\mathbf{Y}$ in Table 2. For clarity, we have numbered these sources from 5 to 8. 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 8 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. The UniBSS method separates well all the sources. The results are qualitatively similar if the dependent and independent sources are selected otherwise from the 6 original sources.

Combining CCA with post-processing with 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 [9] and [13] provide clearly lower signal-to-noise ratios, failing for some sources. Using CCA combined with FastICA or TDESP methods is in practice often preferable over using the UniBSS method. The UniBSS method requires much more samples for reliable results. It may already converge to a separating solution but then deviates again farther away, and this can happen several times. The UniBSS method also requires different types of nonlinearities for sub-Gaussian and super-Gaussian sources. The FastICA and TDSEP methods don't suffer from this limitation.



(a) (b)

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 (a) the first and (b) the second dataset.

The usefulness of the method was tested with data from a functional magnetic resonance imaging (fMRI) study [10], 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) [15] for post-processing the results given by CCA, because it gave better results than FastICA.

Fig. 1 shows the results of applying our method to the two datasets and separating 11 components from the dependent subspaces **U1** and **V1**. The consistency of the components across the subjects is quite good. The first component shows a global hemodynamic contrast, that may also be related to artifacts originating from smoothing the data in the standard preprocessing. The activity of the second component is focused on the primary auditory cortices. The third and fourth components show both positively and negatively task-related activity around the anterior and posterior cingulate gyrus. These first results are promising and in good agreement with the the ones reported in [10]. Future tasks are extension of the method to multiple datasets for interpreting the found components more thoroughly, and a more extensive comparison with existing ICA and BSS methods using real-world data.

References

1. A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis.* Wiley, 2001.
2. P. Comon and C. Jutten (Eds.), *Handbook of Blind Source Separation: Independent Component Analysis and Applications*, Academic Press, 2010.
3. A. Hyvärinen et al., "The FastICA package for Matlab", Helsinki Univ. of Technology, Espoo, Finland, 2005. http://www.cis.hut.fi/projects/ica/fastica/ .
4. 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.
5. A. Hyvärinen, "A unifying model for blind separation of independent sources", *Signal Processing*, vol. 85, no. 7, pp. 1419–1427, 2005.
6. A. Hyvärinen, "Basic Matlab code for the unifying model for BSS". http://www.cs.helsinki.fi/u/ahyvarin/code/UniBSS.m, Univ. of Helsinki, Finland, 2003–2006.
7. 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.
8. M. Van Hulle,"Constrained subspace ICA based on mutual information optimization directly", *Neural Computation*, vol. 20, no. 4, 2008, pp. 964–973.
9. M. Gutmann and A. Hyvärinen, "Extracting coactivated features from multiple data sets", in *Proc. of the Int. Conf. on Artificial Neural Networks (ICANN2011)*, Espoo, Finland, June 2011, pp. 323–330.
10. J. Ylipaavalniemi et al., "Analyzing consistency of independent components: An fMRI illustration", *NeuroImage*, vol. 39, 2008, pp. 169–180.
11. 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.
12. 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), pp. 354–361, 2010.
13. Y.-Q. Li et al., "Joint blind source separation by multiset canonical correlation analysis", *IEEE Trans. on Signal Processing*, Vol. 57, No. 10, October 2009, pp. 3918–3928.
14. A. Rencher, *Methods of Multivariate Analysis, 2nd ed.*, Wiley, 2002.
15. L. Wiskott and T. Sejnowski, "Slow feature analysis: unsupervised learning of invariances", *Neural Computation*, Vol. 14, pp. 715–770, 2002.
16. J. Karhunen and T. Hao, "Finding dependent and independent components from two related data sets", in *Proc. of 2011 Int. J. Conf. on Neural Networks (IJCNN 2011)*, San Jose, CA, USA, July-August 2011, pp. 457–466.