A Canonical Correlation Approach to Blind Source Separation

Magnus Borga, Hans Knutsson

Dept. of Biomedical Engineering
Linköping University
University Hospital
SE-581 85 Linköping
Sweden

June 5, 2001

Abstract

A method based on canonical correlation analysis (CCA) for solving the blind source (BSS) problem is presented. In contrast to independent components analysis (ICA), the proposed method utilises the autocorrelation in the source signals. This makes the BSS problem easier to solve than if only the statistical distribution of the sample values is considered. Experiments show that the method is much more computationally efficient than ICA. The proposed method can be seen as a generalization of the maximum autocorrelation factors (MAF) method.

1 Introduction

The blind source separation (BSS) problem is formulated as follows: Given an unknown linear mixture of a set of unknown, statistically independent signals, the goal is to estimate the mixing matrix and recover the original signals. The mixture can be written as:

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t), \tag{1}$$

where $\mathbf{A}$ is the unknown mixing matrix and the components in $\mathbf{s}$ are the statistically independent source signals. The goal is then to estimate a matrix $\mathbf{W}$ such that

$$\mathbf{y}(t) = \mathbf{W}\mathbf{x}(t) \tag{2}$$

is equal to the unknown source signals in $\mathbf{s}$ (except for scalings and permutations).

In recent years, methods based on information theoretical measures such as mutual information has gained a lot of interest. Such methods try to minimize

the statistical dependence between the reconstructed signals. Measuring statistical dependence is, however, in general not possible and other related objectives has to be used as approximations.

One such method that have gained much attention is independent components analysis (ICA). It was introduced by Comon [3] and many articles has been published on the subject. Instead of minimizing the statistical dependence between the reconstructed signals, ICA tries to make the signals as non-Gaussian as possible. In many practical situations this also reduces the statistical dependence and ICA often give a very good result.

Here we present an alternative method that generates *uncorrelated components*. To look for uncorrelated components means that we use a weaker condition than statistical independence. There are several ways of choosing $\mathbf{W}$ so that the components in $\mathbf{y}$ are uncorrelated, but most of them will not give statistically independent components. One of the most well known methods of generating uncorrelated components i principal component analysis (PCA). PCA is, however, in general unlikely to solve the BSS problem.

A weakness with PCA, which also applies to ICA, is that temporal correlations (or spatial correlations in images) are not taken into account. In both PCA and ICA the samples in time (or spatial position) may be rearranged arbitrarily and the method will still give the same solution. This may seem as a strength, but the fact is that almost all the information in the signal is thrown away. This makes the problem more difficult then it needs to be. As an example, suppose that two images are added together. It is perhaps difficult, but not at all impossible, to see what the original images look like. Now imagine that all pixels in images are permuted randomly. It would now be impossible to see what the original images look like. In practice most, if not all, signals have a certain auto-correlation. This means that $s_i(t)$ and $s_i(t+1)$ are *correlated* while $s_i(t)$ and $s_j(t)$ are *uncorrelated* (since they are statistically independent). This autocorrelation can be used to solve the BSS problem.

Hence, if we want to use correlation instead of statistical dependence as a criterion for solving the BSS problem we must find uncorrelated components that, in addition, have maximum spatial or temporal correlation within each component. This can be achieved by canonical correlation analysis (CCA). A similar approach is called maximum autocorrelation factors (MAF) [8] and has successfully been used to analyse multi-spectral satellite images [7]. While MAF maximizes the autocorrelation for a pre-specified displacement vector, our approach automatically selects the optimum weighted combination of displacement vectors. Here we also have a more direct connection to CCA, which is an established statistical technique. There is a similarity between ICA and CCA in that both can be explained in terms of *mutual information*.

In the following section we give a brief description of CCA. In Section 3, we use the CCA formalism to describe the MAF approach and show under what condition it will solve the BSS problem. In Section 4 we introduce a generalization of the MAF method and show that the condition for solving the BSS problem in this case is weaker than for the MAF approach. Experimental results are presented in Section 5, where the proposed method is compared to

ICA computationally as well as regarding the results.

2 Canonical correlation analysis

CCA finds two sets of basis vectors, one in each signal space, such that the correlation matrix between the signals described in the new basis is a diagonal matrix. A subset of the vectors containing the N first pairs defines a linear rank-N relation between the sets that is optimal in a correlation sense. It has been shown that finding the canonical correlations is equivalent to maximizing the mutual information between the sets if the underlying distributions are elliptically symmetric [6].

Consider two multi-dimensional random variables, $\mathbf{a}$ and $\mathbf{b}$. Consider the linear combinations, $a = \mathbf{w}_a^T(\mathbf{a} - \bar{\mathbf{a}})$ and $b = \mathbf{w}_b^T(\mathbf{b} - \bar{\mathbf{b}})$, of the two variables respectively. The correlation between a and b is given by

$$\rho = \frac{\mathbf{w}_a^T \mathbf{C}_{ab} \mathbf{w}_b}{\sqrt{\mathbf{w}_a^T \mathbf{C}_{aa} \mathbf{w}_a \mathbf{w}_b^T \mathbf{C}_{bb} \mathbf{w}_b}}. \tag{3}$$

where $\mathbf{C}_{aa}$ and $\mathbf{C}_{bb}$ are the nonsingular within-set covariance matrices and $\mathbf{C}_{ab}$ is the between-sets covariance matrix. The maximum of ρ with respect to $\mathbf{w}_a$ and $\mathbf{w}_b$ is the largest *canonical correlation*. A complete description of the canonical correlations is given by:

$$\begin{bmatrix} \mathbf{C}_{aa} & [\,0\,] \\ [\,0\,] & \mathbf{C}_{bb} \end{bmatrix}^{-1} \begin{bmatrix} [\,0\,] & \mathbf{C}_{ab} \\ \mathbf{C}_{ba} & [\,0\,] \end{bmatrix} \begin{pmatrix} \hat{\mathbf{w}}_a \\ \hat{\mathbf{w}}_b \end{pmatrix} = \rho \begin{pmatrix} \lambda_a \hat{\mathbf{w}}_a \\ \lambda_b \hat{\mathbf{w}}_b \end{pmatrix} \tag{4}$$

where: $\rho,\ \lambda_a,\ \lambda_b\ > 0$ and $\lambda_a \lambda_b = 1$. Equation (4) can be rewritten as:

$$\begin{cases} \mathbf{C}_{aa}^{-1} \mathbf{C}_{ab} \hat{\mathbf{w}}_b = \rho \lambda_a \hat{\mathbf{w}}_a \\ \mathbf{C}_{bb}^{-1} \mathbf{C}_{ba} \hat{\mathbf{w}}_a = \rho \lambda_b \hat{\mathbf{w}}_b \end{cases} \tag{5}$$

Solving eq. (5) gives N solutions $\{\rho_n, \hat{\mathbf{w}}_{an}, \hat{\mathbf{w}}_{bn}\},\ n = \{1..N\}$. N is the minimum of the dimensionalities of $\mathbf{a}$ and $\mathbf{b}$. ρ_n are the *canonical correlations* [4]. More details can be found in [1, 2].

3 BSS by maximum autocorrelation

Consider the case where the source signals in $\mathbf{s}$ is a set of one-dimensional signals, e.g. time signals, $\mathbf{s}(t)$. The simplest way of using CCA for separating the mixed signals $\mathbf{x}(t)$ is to find the linear combination of $\mathbf{x}(t)$ that correlates most with a linear combination of $\mathbf{x}(t+1)$, i.e. the next sample. In other words, we let

$$\mathbf{a}(t) = \mathbf{x}(t) \text{ and } \mathbf{b}(t) = \mathbf{x}(t+1). \tag{6}$$

The first canonical correlation component will then give a linear combination $y_1(t)$ of the mixed signals with maximum autocorrelation $r_{y_1}(1)$. The second

canonical correlation component will give a new linear combination $y_2(t)$ with maximum autocorrelation $r_{y_2}(1)$ under the constraint that it is uncorrelated to the first component, etc.

In the case of a two-dimensional source signals, e.g. images, $\mathbf{s}(m,n)$, it is not that obvious how to choose $\mathbf{a}$ and $\mathbf{b}$. In the MAF approach

$$\mathbf{a}(t) = \mathbf{x}(m,n) \text{ and } \mathbf{b}(t) = \mathbf{x}(m+\Delta m, n+\Delta n) \tag{7}$$

where $(\Delta m, \Delta n)$ must be chosen a priori by the user [7]. A CCA on this data would maximize the auto correlation $r(\Delta m, \Delta n)$. A slightly more general approach would be to choose

$$\mathbf{a}(t) = \mathbf{x}(m,n) \text{ and } \mathbf{b}(t) = \mathbf{q}(\mathbf{x}(m,n)) \tag{8}$$

where $\mathbf{q}(\mathbf{x}(m,n))$ is the result of a convolution between $\mathbf{x}(m,n)$ and a filter that is close to circular symmetric and which have a zero in the center of the kernel. This would remove the directional bias in eq. (7). However, the selection of the filter kernel is still to be decided by the user. Another potential disadvantage with the latter method is that the correlation is maximized between a single pixel and a low-pass filtered version of its surrounding. This means that we try to maximize the correlation between two different frequency bands, which is probably not an optimal approach in general.

The method described above chooses the components so that the autocorrelation is maximized. Now, why is that a relevant criterion? Intuitively, it can be explained as follows: If we have two uncorrelated signals, $u(t)$ with high autocorrelation and $v(t)$ with low autocorrelation, the sum $x(t)$ of these signals will have less autocorrelation than the original signal $u(t)$. We can see it as if the addition of $v(t)$ has corrupted $u(t)$, making it harder to predict.

More formally we can say that the sum of uncorrelated signals will have an autocorrelation function that is less than or equal to the maximum of the autocorrelation functions of the individual signals, i.e.

$$r_x(d) \leq \max\{r_{s_i}(d)\} \tag{9}$$

with equality only when all the autocorrelation functions are equal, i.e.

$$r_{s_i}(d) = r_{s_j}(d) \quad \forall i,j.$$

$r_x(d)$ is the autocorrelation function of the sum of the individual signals s_i. To prove eq. (9), we look at the special case of two source signals. The generalization to more than two signals is straight forward.

Consider two uncorrelated signals $u(t)$ and $v(t)$ with zero mean. The autocovariance function of the sum $x(t)$ of the two signals is

$$\begin{aligned} c_x(d) &= \mathcal{F}^{-1}\left[\mathcal{F}[x]\mathcal{F}[x']\right] = \mathcal{F}^{-1}\left[\mathcal{F}[u+v]\mathcal{F}[u'+v']\right] \\ &= \mathcal{F}^{-1}\left[(\mathcal{F}[u]+\mathcal{F}[v])(\mathcal{F}[u]\mathcal{F}[v'])\right] \\ &= \mathcal{F}^{-1}[\mathcal{F}[u]\mathcal{F}[u'] + \mathcal{F}[v]\mathcal{F}[v'] + \underbrace{\mathcal{F}[u]\mathcal{F}[v']}_{=0} + \underbrace{\mathcal{F}[v]\mathcal{F}[u']}_{=0}] \\ &= c_u(d) + c_v(d) \end{aligned} \tag{10}$$

where $\mathcal{F}$ is the Fourier transform and $x'(t) = x(-t)$. The autocorrelation function (i.e. the normalized autocovariance function) is then

$$r_x(d) = \frac{c_x(d)}{c_x(0)} = \frac{c_u(d) + c_v(d)}{c_u(0) + c_v(0)} = \frac{c_u(0)r_u(d) + c_v(0)r_v(d)}{c_u(0) + c_v(0)} \tag{11}$$

which cannot be larger than the maximum of $r_u(d)$ and $r_v(d)$. In fact, eq. (11) shows that the autocorrelation function of the sum is the weighted average between the autocorrelation functions of the sources. The weighting coefficients $c_u(0)$ and $c_v(0)$ are estimates of the variances of $u(t)$ and $v(t)$ respectively.

The conclusion is that if the sources are uncorrelated and have different autocorrelations $r(d)$ for a certain d, then the CCA between the mixture $x(t)$ and $x(t+d)$ will indeed find the original source signals, except for scalings and permutations.

4 BSS by maximum canonical correlation

A problem with the maximum autocorrelation approach is that it maximizes the autocorrelation for a specific distance d, which must be specified a priori. If the autocorrelation functions for the source signals are equal for that particular choice of d, the method will fail.

A more general and robust method is to let the CCA choose, not only the optimal combination of channels, but also optimal filters within each channel. This means that instead of maximizing the correlation between a sample and a neighbouring sample, the CCA now maximizes the correlation between a sample and a linear combination of a neighbouring (but not overlapping) region of the signal. Again, consider the case where the source signals is a set $\mathbf{s}(t)$ of one-dimensional signals $s_i(t)$. We then let

$$\mathbf{a}(t) = \mathbf{x}(t) \quad \text{and} \quad \mathbf{b}(t) = (\mathbf{x}(t+1),\ ...,\ \mathbf{x}(t+N))^T \tag{12}$$

This is an auto-regressive model, i.e. the CCA will find mutually uncorrelated components where the components are optimal with respect to auto-regression error. This is more general than the method in eq. (6). A simple example will show that: Consider the case where $s_1(t) = \sin(t\pi/2)$. The autocorrelation $r(1)$ for such a signal is

$$r(1) = E\left[\sin\left(\frac{t\pi}{2}\right)\sin\left(\frac{t\pi}{2} + \frac{\pi}{2}\right)\right] = E\left[\sin\left(\frac{t\pi}{2}\right)\cos\left(\frac{t\pi}{2}\right)\right] = 0,$$

i.e. the method in eq. (6) would not work for $d = 1$. The auto-regressive model in eq. (12) with $N \geq 2$ would however find a high correlation (-1) between $s(t)$ and $s(t+2)$.

In order to see how the canonical correlation approach can find one of the source signals, we look again at the special case of two uncorrelated signals $u(t)$ and $v(t)$ such that their autocorrelation functions $r_u(d)$ and $r_v(d)$ differ for at

least one value of $d \leq N$. Consider the two signals

$$a(t) = \mathbf{w}_a^T \mathbf{a}(t) \quad \text{and} \quad b(t) = \sum_{i=1}^{N} z_i a(t+i).$$

Here we have assumed that the optimal de-mixing vector for each channel is equal for all samples in time. The coefficients z_i are the coefficients weighting together the different samples in time. The correlation between these two signals is

$$\begin{aligned} \rho(a,b) &= \frac{E[a(t)b(t)]}{\sqrt{E[a^2(t)]E[b^2(t)]}} = \frac{E\left[\sum_{i=1}^{N} z_i a(t)a(t+i)\right]}{\sigma_a^2} \\ &= \frac{\sum_{i=1}^{N} z_i E\left[a(t)a(t+i)\right]}{\sigma_a^2} = \frac{\sum_{i=1}^{N} z_i c_a(i)}{\sigma_a^2} = \frac{\mathbf{z}^T \mathbf{c}_a}{\sigma_a^2} \end{aligned} \tag{13}$$

The coefficient vector $\mathbf{z}$ that maximizes this correlation is the solution to the auto-regression problem of $a(t+i)$ on $a(t)$. This solution is given by

$$\mathbf{z} = \mathbf{C}_a^{-1} \mathbf{c}_a \tag{14}$$

where $\mathbf{C}_a$ is the autocovariance matrix, i.e.

$$C_a(i,j) = \frac{1}{N} \sum_{k=1}^{N} a(i)a(j). \tag{15}$$

This means that the correlation in eq. (13) can be written as

$$\rho(a,b) = \frac{\left(\mathbf{C}_a^{-1}\mathbf{r}_a\right)^T \mathbf{c}_a}{\sigma_a^2} = \frac{\mathbf{c}_a^T \mathbf{C}_a^{-1} \mathbf{c}_a}{\sigma_a^2} = \mathbf{r}_a^T \mathbf{R}_a^{-1} \mathbf{r}_a \tag{16}$$

where $\mathbf{R}_a = \mathbf{C}_a/\sigma_a^2$ and $\mathbf{r}_a = [r_a(1), \ldots, r_a(N)]^T$. From eq. (11) we see that we can write the autocorrelation function $\mathbf{r}_a$ as

$$\mathbf{r}_a = c\mathbf{r}_u + (1-c)\mathbf{r}_v$$

where $\mathbf{r}_u$ and $\mathbf{r}_v$ are the autocorrelation functions for the original signals $u(t)$ and $v(t)$ respectively. This means that we can write the correlation as

$$\begin{aligned} \rho(a,b) &= (c\mathbf{r}_u + (1-c)\mathbf{r}_v)^T \mathbf{R}_a^{-1} (c\mathbf{r}_u + (1-c)\mathbf{r}_v) \\ &= c^2 \mathbf{r}_u^T \mathbf{R}_a^{-1} \mathbf{r}_u + (1-c)^2 \mathbf{r}_v^T \mathbf{R}_a^{-1} \mathbf{r}_v + 2c(1-c)\mathbf{r}_u \mathbf{R}_a^{-1} \mathbf{r}_v \end{aligned} \tag{17}$$

since $\mathbf{R}_a^{-1}$ is symmetric. To find the maximum of the correlation we look at the second derivative of the expression in eq. (17) with respect to the mixing factor c:

$$\frac{\partial^2 \rho}{\partial c^2} = 2(\mathbf{r}_u^T \mathbf{R}_a^{-1} \mathbf{r}_u + \mathbf{r}_v^T \mathbf{R}_a^{-1} \mathbf{r}_v - 2\mathbf{r}_u^T \mathbf{R}_a^{-1} \mathbf{r}_v) \tag{18}$$

From eq. (15) we see that $\mathbf{C}_a$ is positive definite and, hence, so is $\mathbf{R}_a^{-1}$. Therefor we can write

$$\mathbf{r}^T\mathbf{R}_a^{-1}\mathbf{r} = \mathbf{r'}^T\mathbf{r'} \quad \text{whre} \quad \mathbf{r'} = \mathbf{R}^{-1/2}\mathbf{r}.$$

This means that

$$\frac{\partial^2\rho}{\partial c^2} = 2(\mathbf{r'}_u^T\mathbf{r'}_u + \mathbf{r'}_v^T\mathbf{r'}_v - 2\mathbf{r'}_u^T\mathbf{r'}_v) = 2\|\mathbf{r'}_u - \mathbf{r'}_v\|^2 > 0 \tag{19}$$

if $\mathbf{r}_u \neq \mathbf{r}_v$, which means that the correlation has it's maximum for $c = 0$ or $c = 1$, i.e. when $a(t) = u(t)$ or $a(t) = v(t)$. Since the CCA maximizes the correlation, we can make the conclusion that it will not choose a mixture of the source signals, since that would give less correlation than if one of the source signals is chosen.

5 Experimental results

In this section we present some experimental results illustrating the performance of the proposed method. The experiments are done using Matlab on a Sun Ultra 10. For the comparison with ICA, the FastICA algorithm [5] has been used with default parameter settings. The FastICA algorithm is claimed to be "10-100 times faster than conventional gradient descent methods for ICA".

The first experiment illustrates that the CCA approach and ICA give qualitatively the same results. Figure 1 shows the source signals, which are a sine wave function and Gaussian noise (top left) and a mixture of the sources (top right). The lower panel shows the results form the CCA approach and ICA respectively. In this experiment, only one time step has been used for the CCA, i. e. the maximum autocorrelation approach was used.

The second experiment illustrates the difference in computational efficiency between the CCA approach and FastICA. The mixed signals in this case are five EEG channels (Figure 2). We can see that the results from CCA (Figure 3) and ICA (Figure 4) are qualitatively similar. Also in this experiment the maximum autocorrelation approach was used. The computational cost, however, differed significantly. The FastICA algorithm needed 27 Mflops, and 3.8 seconds, while the CCA approach needed 6.8 Mflops, and 0.39 seconds on the Ultra 10 work station.

The final experiment illustrates the difference between the MAF approach, i.e. CCA with only one time step in the **b**-variable, and the more general CCA approach with more than one time step. Figure 5 shows the result when the two source signals were $\sin(t\pi/2)$ and white noise respectively. The one-time step CCA (MAF) fails to recover the source signals, wile the two-time steps CCA manage to recover the sources.

6 Summary and discussion

In many practical situations, the ICA algorithm makes the BSS problem unnecessarily difficult. By only considering the statistical distribution of the sample

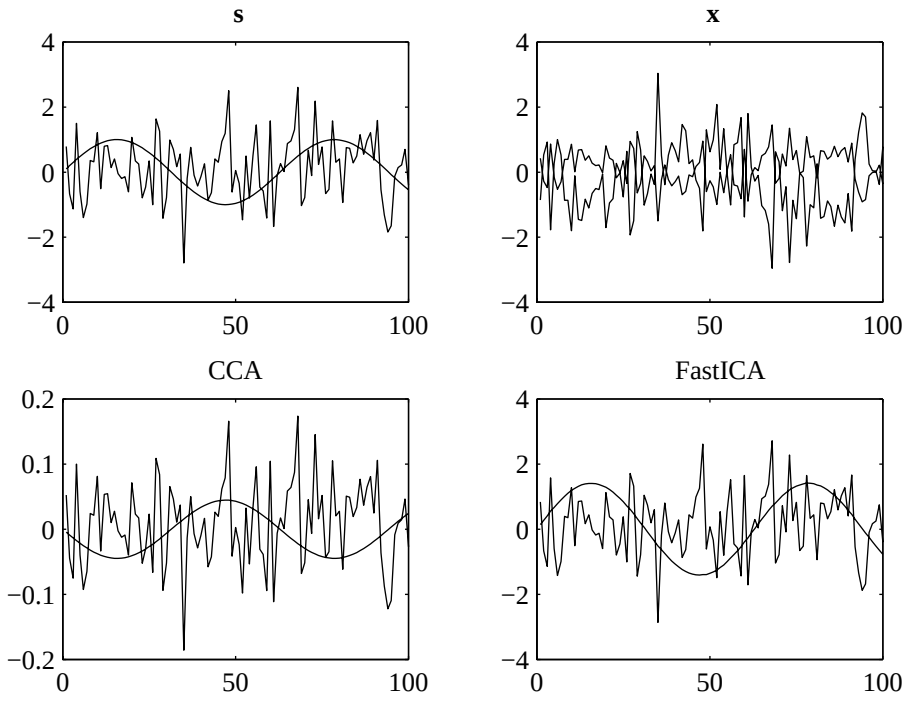

Figure 1: A simple example showing the result of the proposed method compared to the "FastICA" algorithm.

values, ignoring the temporal or spatial relations within the source signals, relevant information is discarded. Most, if not all, natural signals have temporal or spatial relations, causing an autocorrelation in the signal. Canonical correlation analysis can solve the BSS problem with much less computational effort than the ICA algorithm by utilising the autocorrelation functions of the source signals. A necessary condition for the CCA approach to work is that the source signals have different autocorrelation functions.

The method of Maximum autocorrelation factors (MAF) is a special case of the CCA approach, maximizing the autocorrelation in the reconstructed source signals for a particular distance d. In this paper a more general approach has been presented that adaptively finds the optimum filter to be applied to maximize the correlation within the signal. Hence, the method does not depend upon the a priori choice of d.

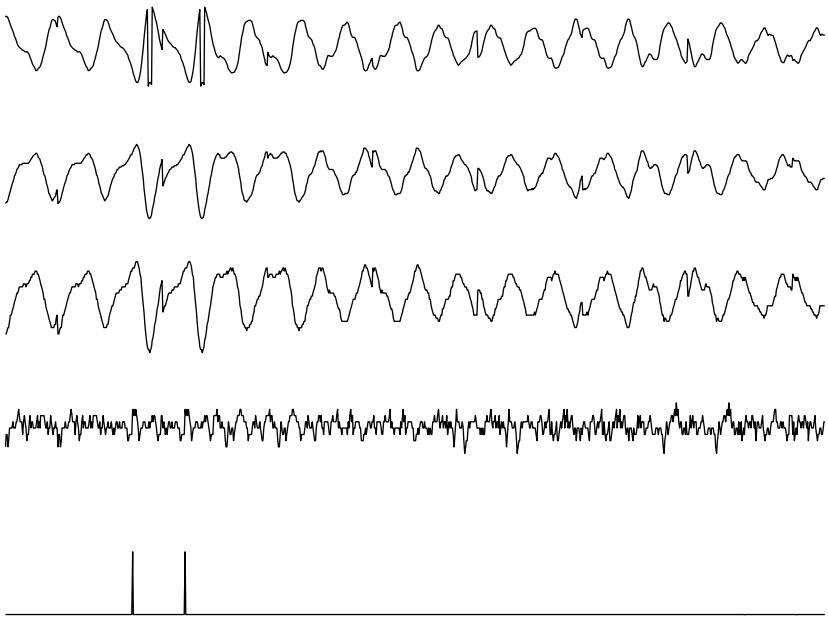

Figure 2: Original EEG signals.

References

[1] T. W. Anderson. *An Introduction to Multivariate Statistical Analysis*. John Wiley & Sons, second edition, 1984.

[2] M. Borga. *Learning Multidimensional Signal Processing*. PhD thesis, Linköping University, Sweden, SE-581 83 Linköping, Sweden, 1998. Dissertation No 531, ISBN 91-7219-202-X, http://people.imt.liu.se/∼magnus/.

[3] P. Comon. Independent component analysis, a new concept? *Signal Processing*, 36(3):287–314, April 1994.

[4] H. Hotelling. Relations between two sets of variates. *Biometrika*, 28:321–377, 1936.

[5] Aapo Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 10(3):626–634, 1999.

[6] J. Kay. Feature discovery under contextual supervision using mutual information. In *International Joint Conference on Neural Networks*, volume 4, pages 79–84. IEEE, 1992.

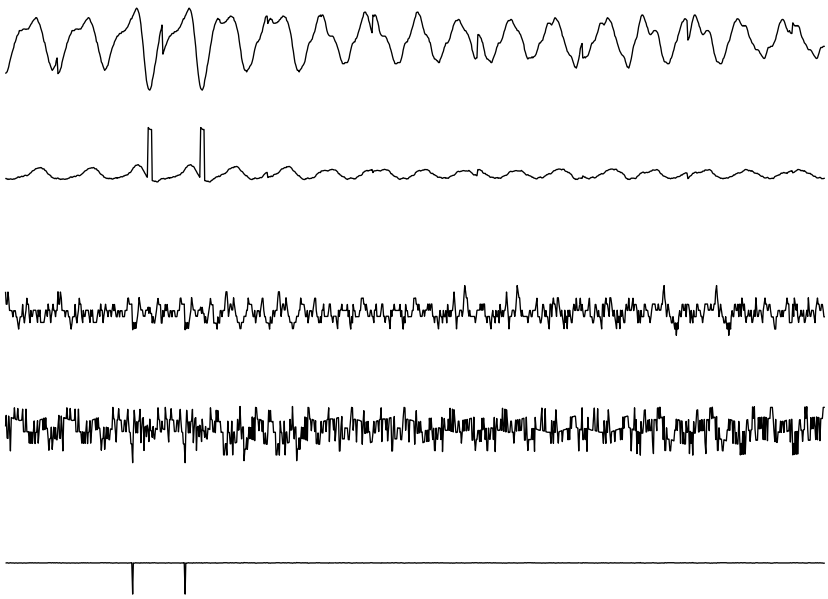

Figure 3: EEG signals decomposed using CCA.

[7] A. A. Nielsen, K Conradsen, and J. Simpson. Mulivariate alteration detection (MAD) and MAF postprocessing in multispectral, bitemporal image data: New approaches to change detection studies. *Remote sens. environ.*, 64:1–19, 1998.

[8] P. Switzer and A. A. Green. Min/max autocorrelation factors for multivariate spatial imagery. Technical Report 6, Department of Statistics, Stanford University, 1984.

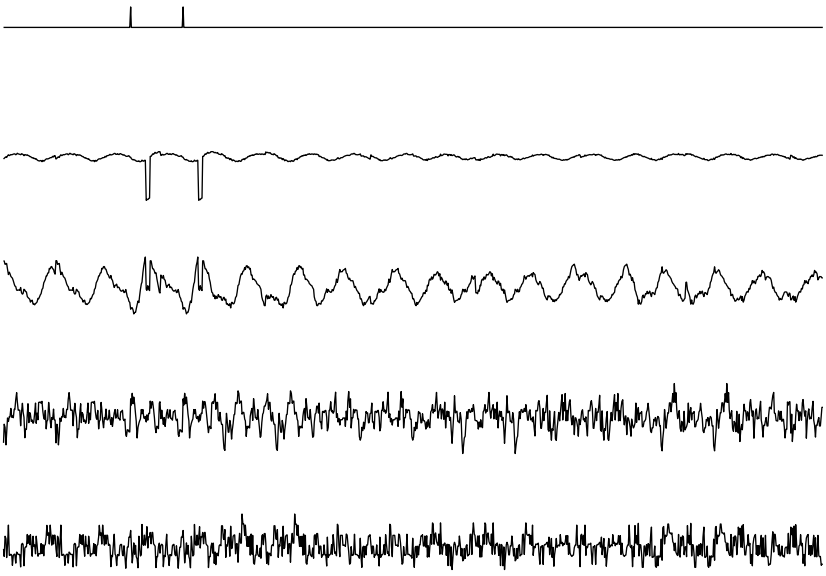

Figure 4: EEG signals decomposed using ICA.

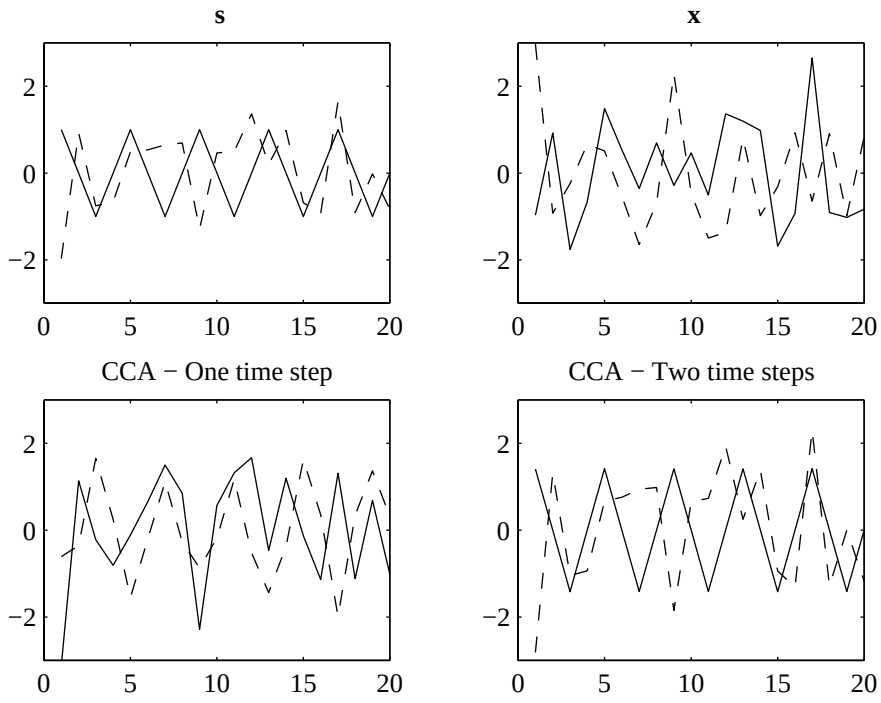

Figure 5: The results using one time step (MAF) or two time steps in the CCA method.