On two Classes of Alternatives to Canonical Correlation Analysis, using Mutual Information and Oblique Projections

De Bie Tijl and De Moor Bart

ESAT-SCD
Katholieke Universiteit Leuven
Kasteelpark Arenberg 10
3001 Leuven
Belgium
http://www.esat.kuleuven.ac.be/sista-cosic-docarch
tijl.debie@esat.kuleuven.ac.be
tel.: +32-16-32 19 39
fax: +32-16-32 19 70

Abstract. Many of the widely applied signal processing algorithms are based on second order statistics. PCA (*principal component analysis*) identifies directions of large variance in data living in one single data space, CCA (*canonical correlation analysis*) searches highly correlated directions in data living in different (often 2) spaces. While the use of only these second order statistics is well justifiable for gaussian data, often higher order statistics are required to cover the needs of real life applications. Considerations based on statistical independency and information theory have lead to ICA (*independent component analysis*), which can be seen as an alternative to PCA. We motivate the need for an analogous alternative to CCA by explaining why CCA doesn't work appropriately for certain data sets. Firstly, a class of alternative versions of CCA based on mutual information will be presented in this paper. As a second alternative, we show how, under certain conditions, CCA may be adapted to overcome its weaknesses while still remaining based on second order statistics. A side result is a new and very efficient algorithm for calculating ICA under certain circumstances.[1]

keywords: canonical correlation analysis, independent component analysis, signal processing, mutual information

[1] Matrices will be denoted by capital letters; the i-th column of matrix M is denoted as M(:,i); vector variables are denoted by a regular lower case letter; samples of a vector variable, put next to each other in a matrix, are denoted by a boldface lower case letter corresponding to the variable.

1 Canonical Correlation Analysis and its Limitations

1.1 Canonical Correlation Analysis (CCA)

Canonical correlation analysis is used to find relations between two (or sometimes more) datasets $\mathbf{y_1}$ and $\mathbf{y_2}$ (here assumed to be uncentered), of which the k colums represent samples in an l_1- and l_2-dimensional space respectively [1]. More specifically, in each of the datasets, it searches for directions, called $D_1(:,1)$ and $D_2(:,1)$, so as to maximize the correlations between the orthogonal projections of the data $\mathbf{y_1}$ and $\mathbf{y_2}$ on $D_1(:,1)$ and $D_2(:,1)$ respectively.

$$
\begin{aligned}
(D_1(:,1), D_2(:,1)) &= \mathrm{argmax}_{v_1 \in \mathbb{R}^l_1, v_2 \in \mathbb{R}^l_2} \frac{v_1^T \cdot \mathbf{y_1}\mathbf{y_2}^T \cdot v_2}{\sqrt{(v_1^T \cdot \mathbf{y_1}\mathbf{y_1}^T \cdot v_1)(v_2^T \cdot \mathbf{y_2}\mathbf{y_2}^T \cdot v_2)}} \\
&= \mathrm{argmax}_{v_1 \in \mathbb{R}^l_1, v_2 \in \mathbb{R}^l_2} \frac{v_1^T \cdot \Sigma_{12} \cdot v_2}{\sqrt{(v_1^T \cdot \Sigma_{11} \cdot v_1)(v_2^T \cdot \Sigma_{22} \cdot v_2)}}
\end{aligned}
$$

where

$$
\begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^T & \Sigma_{22} \end{pmatrix} = \frac{1}{k} \begin{pmatrix} \mathbf{y_1} \\ \mathbf{y_2} \end{pmatrix} (\mathbf{y_1}^T \mathbf{y_2}^T) \tag{1}
$$

This gives the first canonical components $D_1(:,1)$ and $D_2(:,1)$. The following canonical components are found by maximizing the same cost function, under the additional constraint that the data projected on them must be uncorrelated with the projections on the previous canonical components. The number of canonical components is equal to the smallest dimension among the datasets: $\min(l_1, l_2)$. One can easily prove that they are found by solving the generalized eigenvalue problem

$$
\begin{gathered}
\begin{pmatrix} 0 & \Sigma_{12} \\ \Sigma_{21} & 0 \end{pmatrix} \cdot \begin{pmatrix} D_1 & D_1 \\ D_2 & -D_2 \end{pmatrix} \\
= \\
\begin{pmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{pmatrix} \begin{pmatrix} D_1 & D_1 \\ D_2 & -D_2 \end{pmatrix} \cdot \begin{pmatrix} \Lambda & 0 \\ 0 & -\Lambda \end{pmatrix}
\end{gathered} \tag{2}
$$

where Λ is a diagonal matrix containing the generalized eigenvalues.

1.2 Limitations

It is easy to see that CCA often leads to meaningless results for non-gaussian data. Suppose we have the following (noise free) generative model for the data, where we take dimensions of y_1, y_2 and x equal to l:

$$
\begin{cases} y_1 = M_1 \cdot x \\ y_2 = M_2 \cdot x \end{cases} \tag{3}
$$

where M_1 and M_2 are square mixing matrices and x is a vector containing independent signals. This means that the mutual information[2] between $x(i)$ and

[2] The mutual information between $x(i)$ and $x(j)$, with $\mathcal{H}(\S)$ the entropy of x, is defined as $\mathcal{MI}(x(i), x(j)) = \mathcal{H}(x(i)) + \mathcal{H}(x(j)) - \mathcal{H}(x(i), x(j))$.

$x(j)$ is zero for each i and j with $i \neq j$

$$\mathcal{MI}(x(i), x(j)) = 0$$

Applying CCA on a such a datasets generally leads to useless results, in the sense that the retrieved canonical components give no information on the underlying mixing model (figure 1).

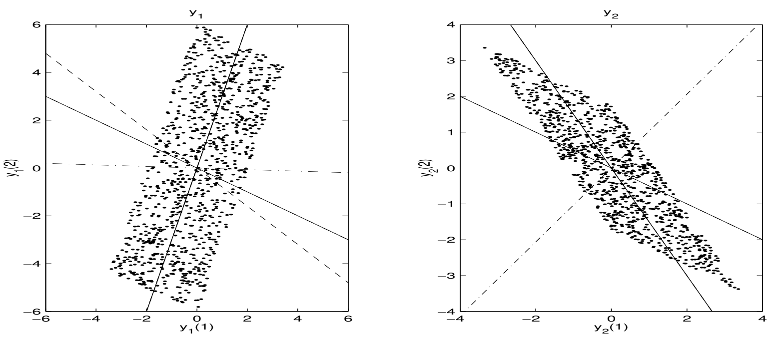

Fig. 1. Depicted in the figure are 1000 samples of a 2-dimensional y_1 on the left, and on the right the corresponding samples of y_2. The solid lines represent the directions of the independent components. The dash-dotted line in each plot shows the first canonical component, the dashed line the second. For calculating the first canonical component, the correlation between the *orthogonal* projections of y_1 and y_2 on the respective directions is calculated and maximized. The second canonical component is such that the orthogonal projections of the samples on it, is uncorrelated with the orthogonal projections on the first canonical component. It is clear that in order to find directions which give an indication on the underlying model, orthogonal projections are not suited. The reason for this is that the independent components are linearly combined as in formula (4), and the colums of M_1 and M_2 are not necessarily orthogonal.

We can pin down two reasons for this:

- only second order statistics are used, and
- projections of the data on the canonical components are orthogonal, whereas the colums of the mixing matrices are not necessarily orthogonal to each other.

1.3 Comparison with Principal Component Analysis and Independent Component Analysis

ICA is an analysis tool analogous to the well known PCA. It searches for a linear decomposition of a signal y in components x along $M(:, i)$, so that the components of x are independent of each other. Thus, it is assumed that the data is generated by the model:

$$y = M \cdot (x + n) + m$$

where the components of x are mutually independent, n and m represent noise, and M is the mixing matrix. [3] When PCA would be applied on such data, the x are usually not retrieved. Attempts to solve this in general using nonlinear PCA versions didn't succeed.

A lot of research has been done on algorithms to find the independent components ([6],[5]). Most of them consist in a dewhitening step performed first, followed by a step involving higher order statistics or nonlinear decorrelation. The first of these steps accounts for the fact that projections on the independent components are not orthogonal. The second one minimizes a mutual information like quantity between the projections.

As for PCA, nonlinear versions of CCA have been proposed (such as in [4],[3]), but in general, these are not suited either for determining independent components underlying two datasets. The analogy between PCA and CCA however suggests to try an analogous generalization of CCA to an ICA-like algorithm. We will thus propose two variants of CCA, which relate to CCA in the same way as ICA relates to PCA and both tend to fill the gap in the following table:

	based on second order statistics	**based on mutual information**
one signal space	PCA	ICA
more than one signal space	CCA	?
	algorithms use orthogonal projections	**algorithms use oblique projections**

2 A Class of Alternatives to Canonical Correlation Analysis

2.1 General Framework

ICA can be seen as a model-based variant of PCA. We will now formulate an analogous variant to CCA. We consider a model of the form:

$$\begin{cases} y_1 = M_1 \cdot (x + n_1) + m_1 \\ y_2 = M_2 \cdot (x + n_2) + m_2 \end{cases} \tag{4}$$

[3] The noise term n is often omitted, since in ICA no distinction between x and n can be made anyway. However, it is clear that noise on the independent component signals directly is often present. (For instance, noise on independent radio signals can be represented by n, while m stands for noise on the measurements performed by the antenna array.) Therefore, and since it will become relevant in the remainder of the text, we mention n as an additional signal here already.

where x_1 and x_2 are the independent component variables, M_1 and M_2 the mixing matrices, y_1 and y_2 the measurements, and n_1, n_2, m_1 and m_2 represent noise on the independent component variables, and measurement noise respectively. From now on, for convenience, we assume the dimensions of the data vectors to be all equal to l. Further, we assume that being $d = \left(x^T n_1^T n_2^T m_1^T m_2^T\right)^T$, $\mathcal{MI}\{d(i), d(j)\} = 0$ for all i and j.

Analogous to ICA, and assuming the noise is small enough, we can now search for demixing matrices transforming the variables y_1 and y_2 into $\widetilde{x}_1$ and $\widetilde{x}_2$ with $\mathcal{MI}\{\widetilde{x}_*(i), \widetilde{x}_*(j)\} = 0$ (or as small as possible).

For the problem at hand, however, we have an additional criterium that says that for every $\widetilde{x}_1(i)$, there must exist an $\widetilde{x}_2(j)$, so that $\widetilde{x}_1(i)$ and $\widetilde{x}_2(j)$ *correspond* with each other. Here, *correspond* should be defined more precisely.

These two criteria clearly overdetermine the problem. We will now propose several ways to cope with this.

Single space ICA followed by matching One possible solution is to perform ICA on $\mathbf{y_1}$ and on $\mathbf{y_2}$ seperately, This will generally yield two different estimations $\widetilde{\mathbf{x}}_\mathbf{1}$ and $\widetilde{\mathbf{x}}_\mathbf{2}$ of $\mathbf{x}$. Then, as a second step, one looks for pairwise correspondences between the elements of $\widetilde{\mathbf{x}}_\mathbf{1}$ and of $\widetilde{\mathbf{x}}_\mathbf{2}$. This can be done by scoring every pair with a mutual information like quantity, or with the square of the euclidian distance, and then optimizing the sum of this distance measure over all possible matchings. An alternative approach to this problem in which these two steps are performed more or less simultaneously, is described in [2]. We should note that the algorithm described in [2] is not restricted to signals parameterized by our linear model (5).

Rank reduction followed by ICA If the measurement noise is small enough as compared to the signals, we can first do a rank reduction on $\begin{pmatrix} \mathbf{y_1} \\ \mathbf{y_2} \end{pmatrix}$, leading to $\widetilde{\mathbf{y}}$, and then apply ICA on $\widetilde{\mathbf{y}}$, leading to only one estimate of $\mathbf{x}$. Note that the rank reduction is nothing else than performing PCA, and retaining the principal components containing most of the energy.

Double Space ICA followed by selection of independent components A similar approach consists of performing ICA directly on $\begin{pmatrix} \mathbf{y_1} \\ \mathbf{y_2} \end{pmatrix}$, followed by a selection of the independent components which are most prominently present in the measurements (in terms of energy content). Another selection criterium could be to choose the independent components which are the least gaussian (noise in particular and thus measurement noise is often close to gaussian).

2.2 Discussion

Each of these methods leads to similar results in many cases, except:

- The first method leads to two estimates of $\mathbf{x}$, whereas the other two only give one and usually better estimate.
- The second method is the fastest, and performs well in the case of low measurement noise levels.
- The third method can handle higher levels of measurement noise, as long as it has a gaussian distribution and the independent components $\mathbf{x}$ don't (which is likely in many cases).

3 A Second Class for a Special Case

The key element in the next algorithm is that certain oblique projections are invoked. Furthermore, we assume quite stringent conditions on the generative model of underlying the data, so that only correlation analysis needs to be applied, no use of higher order statistics has to be made.

3.1 Conditions

We assume in the algorithm that the data is generated in the following way:

$$\begin{cases} y_1 = M_1 \cdot (x + n_1) \\ y_2 = M_2 \cdot (x + n_2) \end{cases} \tag{5}$$

with lengths of vectors x, n_1 and n_2 equal to l, n_1 and n_2 being (not necessary gaussian) noise on x and with correlations

$$\frac{1}{k} \begin{pmatrix} \mathbf{x} \\ \mathbf{n_1} \\ \mathbf{n_2} \end{pmatrix} (\mathbf{x}^T \mathbf{n_1}^T \mathbf{n_2}^T) = \begin{pmatrix} \mathbb{I} & 0 & 0 \\ 0 & \mathit{\Gamma} & 0 \\ 0 & 0 & \mathit{\Gamma} \end{pmatrix} \tag{6}$$

with $\mathbb{I}$ the identity, $\mathit{\Gamma}$ diagonal, and where $\mathbf{x}$, $\mathbf{n_1}$ and $\mathbf{n_2}$ have zero mean.

In other words, we expect y_1 and y_2 to be generated by x, and each by an uncorrelated zero mean noise influence which is also uncorrelated with x.

3.2 Algorithm and Proof

We are now interested in retrieving the independent variables $\mathbf{x}$, as well as M_1 and M_2 giving the independent component directions.

Theorem 3.21 *We can find an estimate of the independent component directions, being the columns of M_1 and M_2, as the solutions of the eigenvalue problem*

$$\begin{pmatrix} 0 & \Sigma_{12} \cdot \Sigma_{22}^{-1} \\ \Sigma_{21} \cdot \Sigma_{11}^{-1} & 0 \end{pmatrix} \cdot \begin{pmatrix} M_1 & M_1 \\ M_2 & -M_2 \end{pmatrix} = \begin{pmatrix} M_1 & M_1 \\ M_2 & -M_2 \end{pmatrix} \cdot \begin{pmatrix} \Lambda & 0 \\ 0 & -\Lambda \end{pmatrix} \tag{7}$$

where $\Lambda = (I + \Gamma)^{-1}$ is diagonal.

We can see that as long as Γ has different values on its diagonal, the solution will be determined. Moreover, the relative values of the eigenvalues give an indication of the robustness of the eigenvalue problem: the closer to 1, the less robust.

Proof 3.21 *We 'll proof that*

$$\Sigma_{21} \cdot \Sigma_{11}^{-1} \cdot M_1 = M_2 \cdot \Lambda$$

the remainder of the proof is analogous.

$$\begin{aligned} \Sigma_{21} \cdot \Sigma_{11}^{-1} \cdot M_1 &= M_2 \cdot \Lambda \\ \Leftrightarrow M_2(\mathbf{x}+\mathbf{n_2})(\mathbf{x}+\mathbf{n_1})^T M_1^T \cdot [M_1(\mathbf{x}+\mathbf{n_1})(\mathbf{x}+\mathbf{n_1})^T M_1^T]^{-1} M_1 &= M_2 \cdot \Lambda \\ \Leftrightarrow (\mathbf{x}+\mathbf{n_2})(\mathbf{x}+\mathbf{n_1})^T \cdot [(\mathbf{x}+\mathbf{n_1})(\mathbf{x}+\mathbf{n_1})^T]^{-1} &= \Lambda \\ \Leftrightarrow \mathbf{x}\mathbf{x}^T \cdot [\mathbf{x}\mathbf{x}^T + \mathbf{n_1}\mathbf{n_1}^T]^{-1} &= \Lambda \\ \Leftrightarrow \mathbb{I} \cdot [\mathbb{I} + \Gamma]^{-1} &= \Lambda \end{aligned}$$

3.3 Corollary: a fast ICA algorithm

The goal of ICA is to separate a set of signals into its statistically independent components. In developing ICA methods, different approximations of mutual information have been made ([5],[6]). Essentially all go out from the following model underlying the data y: $y = M \cdot x + m$, where the mutual information between $x(i)$ and $x(j)$ is zero for each i and j with $i \neq j$: $\mathcal{MI}(x(i), x(j)) = 0$, and with m being gaussian noise.

However, when we have the measurements as in equation (5) at our disposition with negligible m_1 and m_2, we can apply the higher described algorithm for retrieving x, and thus the independent components. This algorithm is much more efficient than any of the previously described algorithms in literature. As higher described (subsection 3.2), a quality factor of the result can be returned, and if then appearing necessary, a standard ICA algorithm can still be performed.

4 Conclusions

We have shown that CCA in some cases leads to meaningless results. Adaptations of the CCA algorithm were made, firstly by making use of existing ICA algorithms, and secondly through requiring stronger assumptions on the dataset.

A side result is the fact that under certain assumptions on the data, use of correlation analysis can be enough for carrying out an independent component analysis. In the past, it was generally thought that second order statistics is indispensable for doing ICA.

5 Acknowledgments

Tijl De Bie is a research assistant with the Fund for Scientific Research - Flanders (FWO-Vlaanderen). Dr. Bart De Moor is a full professor at the Katholieke Universiteit Leuven, Belgium. Our research is supported by grants from several funding agencies and sources: **Research Council KUL**: Concerted Research Action GOA-Mefisto 666 (Mathematical Engineering), IDO (IOTA Oncology, Genetic networks), several PhD/postdoc & fellow grants; **Flemish Government**: Fund for Scientific Research Flanders (several PhD/postdoc grants, projects G.0256.97 (subspace), G.0115.01 (bio-i and microarrays), G.0240.99 (multilinear algebra), G.0197.02 (power islands), G.0407.02 (support vector machines), research communities ICCoS, ANMMM), AWI (Bil. Int. Collaboration Hungary/ Poland), IWT (Soft4s (softsensors), STWW-Genprom (gene promotor prediction), GBOU-McKnow (Knowledge management algorithms), Eureka-Impact (MPC-control), Eureka-FLiTE (flutter modeling), several PhD grants); **Belgian Federal Government**: DWTC (IUAP IV-02 (1996-2001) and IUAP V-22 (2002-2006): Dynamical Systems and Control: Computation, Identification & Modelling), Program Sustainable Development PODO-II (CP/40: Sustainibility effects of Traffic Management Systems); **Direct contract research**: Verhaert, Electrabel, Elia, Data4s, IPCOS;

References

1. H. Hotelling, "Relations between two sets of variables", Biometrika, 28, 321-327, 1936.
2. S.Akaho, Y.Kiuchi, and S. Umeyama, "MICA: Multimodal independent component analysis", Proc. of Int. Joint Conf. on Neural Networks (IJCNN'99), pp.927-932, 1999.
3. Thomas Melzer, Michael Reiter, Horst Bischof, "Kernel Canonical Correlation Analysis", PRIP-TR-65,2001.
4. Tony Van Gestel, Johan A.K. Suykens, Jos De Brabanter, Bart De Moor, Joos Vandewalle, "Kernel Canonical Correlation Analysis and Least Squares Support Vector Machines", Proc. International Conference on Artificial Neural Networks (ICANN 2001), pp. 381-386, Vienna, Austria, August 21-25, 2001.
5. Lieven De Lathauwer, Bart De Moor, Joos Vandewalle, "An Introduction to Independent Component Analysis", Journal of Chemometrics, Special Issue on Multi-Way Analysis, Vol. 14, No. 3, May-June 2000, pp. 123-149.
6. Aapo Hyvärinen, Juha Karhunen, Erkki Oja, "Independent Component Analysis", John Wiley & Sons, 2001.