ICA USING KERNEL CANONICAL CORRELATION ANALYSIS

Colin Fyfe and Pei Ling Lai

Applied Computational Intelligence Research Unit,
The University of Paisley, Scotland.

ABSTRACT

We derive a new method based on kernels for performing Canonical Correlation Analysis. We show that the method can be used to extract individual sinusoids from a linear mixture of sinusoids and that this is also possible when the number of mixtures is less than the number of signals, when there is a nonlinear mixture of the signals and when the mixture is time varying. In the last case, the nature of the time varying mixture matrix is revealed by some of the lower order canonical correlations when we use a nonlinear kernel.

1. INTRODUCTION

We consider it to be interesting that we have two ears. This is not totally explained by the fact that organisms exhibit symmetry since we have only one primary organ for transmitting information (and indeed this organ doubles as an energy input device). We therefore consider a method of extracting information from two data sets each of which contains a mixture of signals as our ears do. We have previously [2] introduced an artificial neural network method of performing the statistical technique of Canonical Correlation Analysis. Canonical Correlation Analysis is a statistical technique used when we have two data sets which we believe have some underlying correlation. We have extended our neural method to enable nonlinear correlations to be found [4]. In this paper, we investigate Kernel Canonical Correlation Analysis (KCCA) which is the linear operation of Canonical Correlation Analysis performed in a feature space formed from a nonlinear mapping of input space. The most frequently reported linear operation involving unsupervised learning in feature space has been Kernel Principal Component Analysis (KPCA) e.g. [8, 7, 6, 5]. We apply a similar method to perform Kernel Canonical Correlation Analysis and show its ability to extract Independent Components from linear, nonlinear and time- varying mixtures of signals.

2. CANONICAL CORRELATION ANALYSIS

Consider two sets of input data; $\mathbf{x}_1$ and $\mathbf{x}_2$. Then in classical CCA, we attempt to find that linear combination of the variables which give us maximum correlation between the combinations. Let

$$\begin{aligned}\mathbf{y}_1 &= \mathbf{w}_1\mathbf{x}_1 = \sum_j w_{1j}x_{1j} \\ \mathbf{y}_2 &= \mathbf{w}_2\mathbf{x}_2 = \sum_j w_{2j}x_{2j}\end{aligned}$$

where we have used $\mathbf{x}_{ij}$ as the j^{th} element of $\mathbf{x}_1$.

Then we wish to find those values of $\mathbf{w}_1$ and $\mathbf{w}_2$ which maximise the correlation between $\mathbf{y}_1$ and $\mathbf{y}_2$. If the relation between $\mathbf{y}_1$ and $\mathbf{y}_2$ is believe to be causal, we may view the process as one of finding the best predictor of the set $\mathbf{x}_2$ by the set $\mathbf{x}_1$ and similarly of finding the best predictable criterion in the set $\mathbf{x}_2$ from the set $\mathbf{x}_1$ data set.

Then the standard statistical method (see [3]) lies in defining

$$\begin{aligned}\Sigma_{11} &= E\{(\mathbf{x}_1-\mu_1)(\mathbf{x}_1-\mu_1)^T\} \\ \Sigma_{22} &= E\{(\mathbf{x}_2-\mu_2)(\mathbf{x}_2-\mu_2)^T\} \\ \Sigma_{12} &= E\{(\mathbf{x}_1-\mu_1)(\mathbf{x}_2-\mu_2)^T\} \\ \text{and } K &= \Sigma_{11}^{-\frac{1}{2}}\Sigma_{12}\Sigma_{22}^{-\frac{1}{2}} \qquad (1)\end{aligned}$$

where T denotes the transpose of a vector. We then perform a Singular Value Decomposition of K to get

$$K = (\alpha_1, \alpha_2, ..., \alpha_k)D(\beta_1, \beta_2, ..., \beta_k)^T \qquad (2)$$

where α_i and β_i are the standardised eigenvectors of KK^T and K^TK respectively and D is the diagonal matrix of eigenvalues.

Then the first canonical correlation vectors (those which give greatest correlation) are given by

$$\mathbf{w}_1 = \Sigma_{11}^{-\frac{1}{2}}\alpha_1 \qquad (3)$$

$$\mathbf{w}_2 = \Sigma_{22}^{-\frac{1}{2}}\beta_1 \qquad (4)$$

with subsequent canonical correlation vectors defined in terms of the subsequent eigenvectors, α_i and β_i.

2.1. Kernel Canonical Correlation Analysis

Consider mapping the input data to a high dimensional (perhaps infinite dimensional) feature space, F. Now,

$$\begin{aligned}\Sigma_{11} &= E\{(\phi(\mathbf{x}_1)-\mu_1)(\phi(\mathbf{x}_1)-\mu_1)^T\}\\ \Sigma_{22} &= E\{(\phi(\mathbf{x}_2)-\mu_2)(\phi(\mathbf{x}_2)-\mu_2)^T\}\\ \Sigma_{12} &= E\{(\phi(\mathbf{x}_1)-\mu_1)(\phi(\mathbf{x}_2)-\mu_2)^T\}\end{aligned}$$

where now $\mu_i = E(\phi(\mathbf{x}_i))$ for $i = 1, 2$. Let us assume for the moment that the data has been centred in feature space (we actually will use the same trick as [7] to centre the data later). Then we define

$$\begin{aligned}\Sigma_{11} &= E\{\phi(\mathbf{x}_1)\phi(\mathbf{x}_1)^T\}\\ \Sigma_{22} &= E\{\phi(\mathbf{x}_2)\phi(\mathbf{x}_2)^T\}\\ \Sigma_{12} &= E\{\phi(\mathbf{x}_1)\phi(\mathbf{x}_2)^T\}\end{aligned}$$

and we wish to find those values $\mathbf{w}_1$ and $\mathbf{w}_2$ which will maximise $\mathbf{w}_1^T\Sigma_{12}\mathbf{w}_2$ subject to the constraints $\mathbf{w}_1^T\Sigma_{11}\mathbf{w}_1 = 1$ and $\mathbf{w}_2^T\Sigma_{22}\mathbf{w}_2 = 1$.

In practise we will approximate Σ_{12} with $\frac{1}{n}\sum_i \phi(\mathbf{x}_{1i})\phi(\mathbf{x}_{2i})$, the sample average.

At this stage we can see the similarity with nonlinear CCA [2]: if we consider an instantaneous hill-climbing algorithm, we would derive precisely our NL-CCA algorithm for the particular nonlinearity involved.

Now $\mathbf{w}_1$ and $\mathbf{w}_2$ exist in the feature space which is spanned by $\{\phi(\mathbf{x}_{11}), \phi(\mathbf{x}_{12}), ..., \phi(\mathbf{x}_{1n}), \phi(\mathbf{x}_{21}), ..., \phi(\mathbf{x}_{2n})\}$ and therefore can be expressed as

$$\begin{aligned}\mathbf{w}_1 &= \sum_{i=1}^{n}\alpha_{1i}\phi(\mathbf{x}_{1i}) + \sum_{i=1}^{n}\alpha_{2i}\phi(\mathbf{x}_{2i})\\ \mathbf{w}_2 &= \sum_{i=1}^{n}\beta_{1i}\phi(\mathbf{x}_{1i}) + \sum_{i=1}^{n}\beta_{2i}\phi(\mathbf{x}_{2i})\end{aligned}$$

With some abuse of the notation we will use $\mathbf{x}_i$ to be the i^{th} instance from the set of data i.e. from either the set of values of $\mathbf{x}_1$ or from those of $\mathbf{x}_2$ and write

$$\begin{aligned}\mathbf{w}_1 &= \sum_{i=1}^{2n}\alpha_i\phi(\mathbf{x}_i)\\ \mathbf{w}_2 &= \sum_{i=1}^{2n}\beta_i\phi(\mathbf{x}_i)\end{aligned}$$

Therefore substituting this in the criteria we wish to optimise, we get

$$(\mathbf{w}_1^T\Sigma_{12}\mathbf{w}_2) = \frac{1}{n}\sum_{k,i}\alpha_k.\phi^T(\mathbf{x}_k)\phi(\mathbf{x}_{1i})\sum_l \beta_l\phi^T(\mathbf{x}_{2i})\phi(\mathbf{x}_l) \tag{5}$$

where the sums over i are to find the sample means over the data set. Similarly with the constraints and so

$$\begin{aligned}\mathbf{w}_1^T\Sigma_{11}\mathbf{w}_1 &= \frac{1}{n}\sum_{k,i}\alpha_k.\phi^T(\mathbf{x}_k)\phi(\mathbf{x}_{1i}).\sum_l \alpha_l\phi^T(\mathbf{x}_{1i})\phi(\mathbf{x}_l)\\ \mathbf{w}_2^T\Sigma_{22}\mathbf{w}_2 &= \frac{1}{n}\sum_{k,i}\beta_k.\phi^T(\mathbf{x}_k)\phi(\mathbf{x}_{2i}).\sum_l \beta_l\phi^T(\mathbf{x}_{2i})\phi(\mathbf{x}_l)\end{aligned}$$

Using $(K_1)_{ij} = \phi^T(\mathbf{x}_i)\phi(\mathbf{x}_{1j})$ and $(K_2)_{ij} = \phi^T(\mathbf{x}_i)\phi(\mathbf{x}_{2j})$ we then have that we require to maximise $\alpha^T K_1 K_2^T\beta$ subject to the constraints $\alpha^T K_1 K_1^T\alpha = 1$ and $\beta^T K_2 K_2^T\beta = 1$. Therefore if we define $\Gamma_{11} = K_1K_1^T$, $\Gamma_{22} = K_2K_2^T$ and $\Gamma_{12} = K_1K_2^T$ we solve the problem in the usual way: by forming matrix $K = \Gamma_{11}^{-\frac{1}{2}}\Gamma_{12}\Gamma_{22}^{-\frac{1}{2}}$ and performing a singular value decomposition on it as before to get

$$K = (\gamma_1, \gamma_2, ..., \gamma_k)D(\theta_1, \theta_2, ..., \theta_k)^T \tag{6}$$

where γ_i and θ_i are again the standardised eigenvectors of KK^T and K^TK respectively and D is the diagonal matrix of eigenvalues [1]

Then the first canonical correlation vectors in feature space are given by

$$\alpha_1 = \Gamma_{11}^{-\frac{1}{2}}\gamma_1 \tag{7}$$

$$\beta_1 = \Gamma_{22}^{-\frac{1}{2}}\theta_1 \tag{8}$$

with subsequent canonical correlation vectors defined in terms of the subsequent eigenvectors, γ_i and θ_i.

Now for any new values $\mathbf{x}_1$, we may calculate

$$\mathbf{w}_1.\phi(\mathbf{x}_1) = \sum_i \alpha_i\phi(\mathbf{x}_i)\phi(\mathbf{x}_1) = \sum_i \alpha_i K_1(\mathbf{x}_i, \mathbf{x}_1) \tag{9}$$

which then requires to be centered as before. We see that we are again performing a dot product in feature space (it is actually calculated in the subspace formed from projections of $\mathbf{x}_i$).

The optimal weight vectors are vectors in a feature space which we may never determine. We are simply going to calculate the appropriate matrices using the kernel trick - e.g. we may use Gaussian kernels so that

$$K_1(\mathbf{x}_{1i}\mathbf{x}_{1j}) = exp(-(\mathbf{x}_{1i} - \mathbf{x}_{1j})^2) \tag{10}$$

which gives us a means of calculating K_{11} without ever having had to calculate $\phi(\mathbf{x}_{1i})$ or $\phi(\mathbf{x}_{1j})$ explicitly.

[1] This optimisation is applicable for all symmetric matrices (Theorem A.9.2, [3]).

3. SIMULATIONS

We repeat the ICA simulation in [1] in which a mixture of two sines is separated. The data set we use comprises three sinusoids in noise mixed linearly with a random mixing matrix; we keep the number of mixtures to two so that we do not stray too far from biological plausibility and so the second line of Figure 1 has only two images. We use 80 samples and linear kernels. Each sample contains the mixture at time t and the mixture at times t-1,..., t-9 so that a vector of 10 samples from the mixture is used at any one time. The underlying signals are shown in the top line of Figure 1 and the correlation filtered data in the bottom line of that Figure. We see that the three sinusoids have been separated with great accuracy. Similar results have been achieved with Gaussian and sigmoid kernels. There is a little beating in the higher frequency sinusoids which is not apparent when we separate two signals from two mixtures. This particular mixture was chosen since there is very little of the signal s_2 in the first mixture. This case is very readily treated with the technique of Minor Component Analysis ([4]) but is much the most difficult case for KCCA.

3.0.1. Nonlinear and Time-varying Mixtures

We now consider one nonlinear mixture and one linear mixture of two sinusoids such as

$$\begin{aligned} \text{Let } m_1 &= s_1 + s_2 \\ m_2 &= (s_1 - s_2) * s_2 \end{aligned}$$

The results of linear KCCA are shown in Figure 2. Again the signals are in the top line, the mixtures in the second line and the filtered data from the correlation process in the bottom line. We see that the two sinusoids have been found but we have to sound a note of caution: the projections shown are on the second and seventh CCA directions. Each sinusoid appears in two filters (one π radians out of phase with the other) but even so, the second signal was not found until the seventh and eighth correlation directions. This finding of the signals in the lower order filters is even more pronounced when we use Gaussian kernels; Figure 3 shows the filtered data found by the nineth and tenth canonical correlation filters.

The extraction of the individual sines becomes progressively harder the more nonlinear both mixtures become.

Non stationary mixtures are shown in Figure 4. Now we create mixtures, m_1 and m_2 using

$$\begin{aligned} m_1 &= 0.95 \sin(t/17) s_1 + 0.61 \sin(t/17) s_2 \\ m_2 &= 0.23 \sin(t/17) s_1 + 0.49 \sin(t/17) s_2 \end{aligned}$$

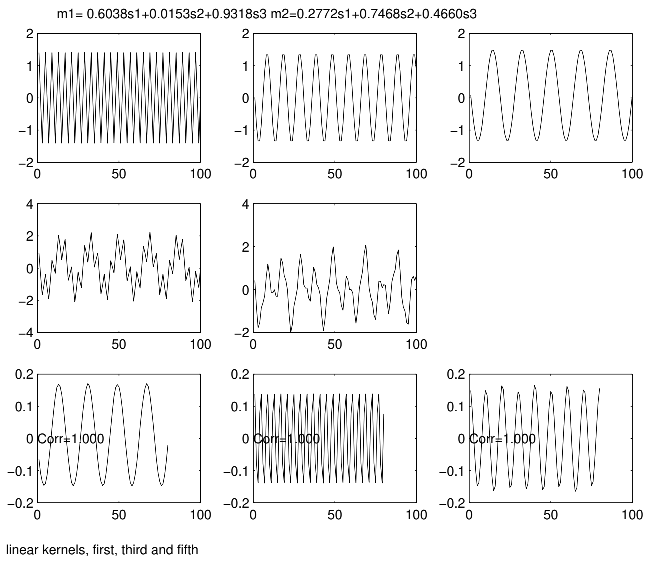

Figure 1: The top line shows the three underlying sinusoids. The second line shows the two linear mixtures of these sinusoids presented to the algorithm. The bottom line shows three filters produced.

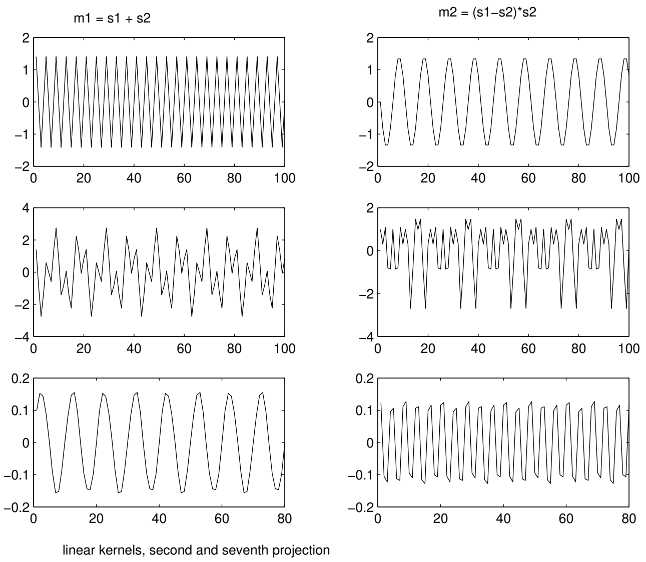

Figure 2: The top line shows the underlying signals, the second line the data and the third line two kernel correlation projections when linear kernels are used.

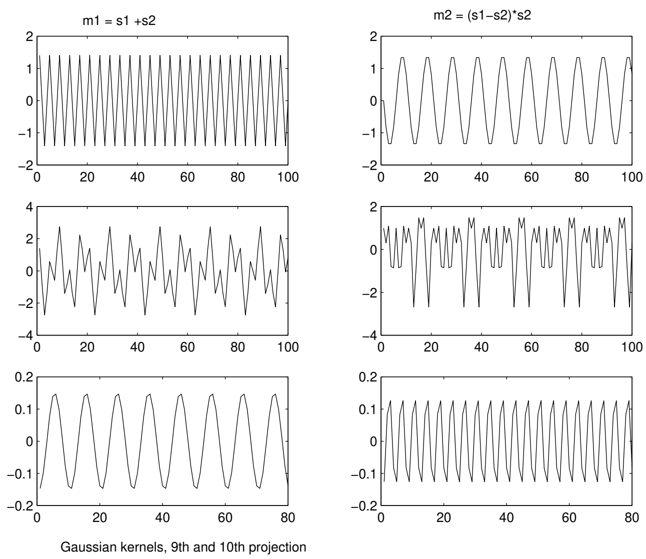

Figure 3: The top line shows the underlying signals, the second line the data and the third line two kernel correlation projections when rbf kernels are used.

where t is a parameter denoting time. The underlying frequencies are extracted by linear kernels (Figure 4).

The rbf kernel, however, (Figure 5) is also able to extract the mixing frequency in its 9th filter. This also happens when we use alternate sines and cosines (and with different frequencies) in our mixing matrix. The linear kernels also finds the individual sines but does not find the mixing frequencies.

4. CONCLUSION

We have shown that we may extract signals from linear mixtures of signals using a kernel implementation of Canonical Correlation Analysis. Interestingly we have used linear kernels in the simulation discussed in this paper and shown that we may extract signals when we have fewer mixtures than signals. We have also shown some success in identifying signals from nonlinear mixtures and in time varying mixtures.

It is an open research problem as to which kernel is appropriate in this problem with particular nonlinear and nonstationary mixtures of signals and the particular values of the parameter set (e.g. width of Gaussian) required for optimal signal extraction is somewhat of a black art. However we have shown that nonlinear kernels (e.g. Gaussian) are able to extract information from a data set which is not extractable using linear kernels.

The standard statistical method of performing CCA will extract sinusoids from stationary mixtures but with less accuracy than the Kernel method described here. We believe that it is the kernel method's ability to describe their feature space with an overcomplete basis (in terms of $\phi(\mathbf{x}_i), \forall i$, which gives this method its power to find interesting projections. Ongoing work will investigate the nature of the feature spaces and the minimal set of vectors needed to span them.

5. REFERENCES

[1] Juha Karhunen and Jyrki Joutsensalo. Representation and separation of signals using nonlinear pca type learning. *Neural Networks*, 7(1):113–127, 1994.

[2] P.L. Lai and C Fyfe. A neural network implementation of canonical correlation analysis. *Neural Networks*, 12(10):1391–1397, Dec. 1999.

[3] K. V. Mardia, J.T. Kent, and J.M. Bibby. *Multivariate Analysis*. Academic Press, 1979.

[4] D. Charles P. L. Lai and C.Fyfe. *Developments in Artificial Neural Network Theory : Independent Component Analysis and Blind Source Sepa-*

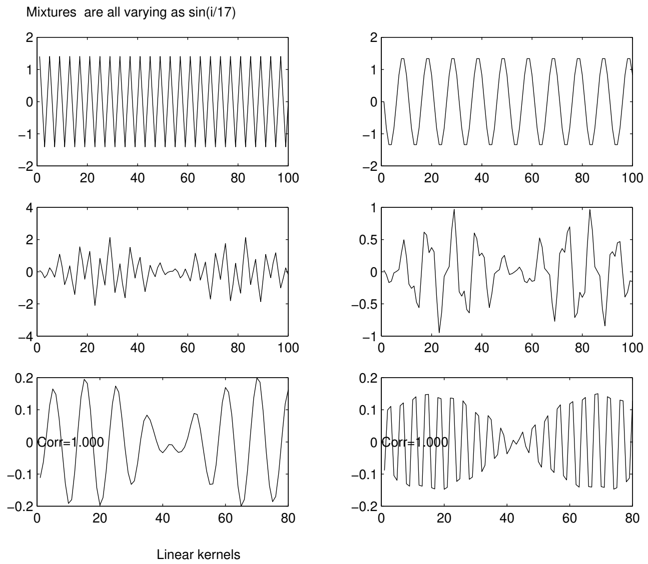

Figure 4: The mixtures are time varying (see text) but nevertheless the linear kernel method extracts the underlying frequencies. Each figure in the bottom line shows the amplitude of the signals varying as they do in the mixtures.

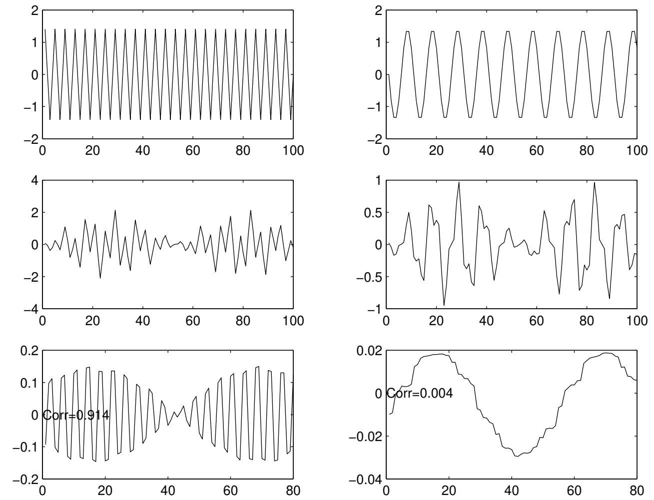

Figure 5: The Gaussian kernel also finds the signals but also finds the underlying slowly changing mixing process in its seventh projection.

ration, chapter Seeking Independence using Biologically Inspired Artificial Neural Networks. Springer-Verlag, 2000.

[5] S. Romdhani, S. Gong, and A. Psarrou. A multi-view nonlinear active shape model using kernel pca. In *BMVC99*, 1999.

[6] B. Scholkopf, S. Mika, C. Burges, P. Knirsch, K.-R. Muller, G. Ratsch, and A. J. Smola. Input space vs feature space in kernel-based methods. *IEEE Transactions on Neural Networks*, 10:1000–1017, 1999.

[7] B. Scholkopf, A. Smola, and K.-R. Muller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10:1299–1319, 1998.

[8] A. J. Smola, O. L. Mangasarian, and B Scholkopf. Sparse kernel feature analysis. Technical Report 99-04, University of Wiscosin Madison, 1999.