

Learning Canonical Correlations

Hans Knutsson **Magnus Borga** **Tomas Landelius**
knutte@isy.liu.se magnus@isy.liu.se tc@isy.liu.se

Computer Vision Laboratory
Department of Electrical Engineering
Linköping University, S-581 83 Linköping, Sweden

Abstract

This paper presents a novel learning algorithm that finds the linear combination of one set of multi-dimensional variates that is the best predictor, and at the same time finds the linear combination of another set which is the most predictable. This relation is known as the *canonical correlation* and has the property of being invariant with respect to affine transformations of the two sets of variates. The algorithm successively finds all the canonical correlations beginning with the largest one.

1 Introduction

A common problem in neural networks and learning, incapacitating many theoretically promising algorithms, is the high dimensionality of the input-output space. As an example typical dimensionalities for systems having visual inputs far exceed acceptable limits. For this reason a priori restrictions must be invoked. A common restriction is to use only locally linear models. To obtain efficient systems the dimensionalities of the models should be as low as possible. The use of locally low-dimensional linear models will in most cases be adequate if the subdivision of the input and output spaces are made adaptively [2, 5].

An important problem is to find the best directions in the input- and output spaces for the local models. Algorithms like the Kohonen self organizing feature maps [4] and others that work with principal component analysis will find directions where the signal variances are high. This is, however, of little use in a response generating system. Such a system should find directions that efficiently represents signals that are *important* rather than signals that have large energy.

In general the input to a system comes from a set of different sensors and it is evident that the range of the signal values from a given sensor is unrelated to the importance of the received information. The same line of reasoning holds for the output which may consist of signals to a set of different effectuators. For this reason the *correlation* between input and output signals is interesting since this measure of input-output relation is independent of the signal variances. However, correlation alone is not necessarily meaningful. Only input-output pairs that are regarded as relevant should be entered in the correlation analysis.

If the system for each input-output pair is supplied with a reward signal the system learns the relationship between rewarded (i.e. relevant)

pairs. Such a system is a *reinforcement learning system*. In this paper we consider the case where we have a distribution of rewarded pairs of input and output signals. This is the distribution for which we are interested in finding an efficient representation by the use of low-dimensional linear models.

Relating only the projections of the input, \mathbf{x} , and output, \mathbf{y} , on two vectors, \mathbf{w}_x and \mathbf{w}_y , establishes a one-dimensional linear relation between the input and output. We wish to find the vectors that maximizes $\text{corr}(\mathbf{x}^T \mathbf{w}_x, \mathbf{y}^T \mathbf{w}_y)$, i.e. the correlation between the projections. This relation is known as *canonical correlation* [3]. It is a statistical method of finding the linear combination of one set of variables that is the best predictor, and at the same time finding the linear combination of an other set which is the most predictable.

In section 2 a brief review of the theory of canonical correlation is given. In section 3 we present an iterative learning rule, equation 7, that finds the directions and magnitudes of the canonical correlations. To illustrate the algorithm behaviour some experiments are presented and discussed in section 4.

2 Canonical Correlation

Consider two random variables, \mathbf{x} and \mathbf{y} , from a multi-normal distribution:

$$(1) \quad \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} \sim N \left(\begin{bmatrix} \mathbf{x}_0 \\ \mathbf{y}_0 \end{bmatrix}, \begin{bmatrix} \mathbf{C}_{xx} & \mathbf{C}_{xy} \\ \mathbf{C}_{yx} & \mathbf{C}_{yy} \end{bmatrix} \right),$$

where $\mathbf{C} = \begin{bmatrix} \mathbf{C}_{xx} & \mathbf{C}_{xy} \\ \mathbf{C}_{yx} & \mathbf{C}_{yy} \end{bmatrix}$ is the covariance matrix. \mathbf{C}_{xx} and \mathbf{C}_{yy} are non-singular matrices and $\mathbf{C}_{xy} = \mathbf{C}_{yx}^T$. Consider the linear combinations, $x = \mathbf{w}_x^T (\mathbf{x} - \mathbf{x}_0)$ and $y = \mathbf{w}_y^T (\mathbf{y} - \mathbf{y}_0)$, of the two variables respectively. The correlation between x and y is given by equation 2, see for example [1]:

$$(2) \quad \rho = \frac{\mathbf{w}_x^T \mathbf{C}_{xy} \mathbf{w}_y}{\sqrt{\mathbf{w}_x^T \mathbf{C}_{xx} \mathbf{w}_x \mathbf{w}_y^T \mathbf{C}_{yy} \mathbf{w}_y}}.$$

The directions of the partial derivatives of ρ with respect to \mathbf{w}_x and \mathbf{w}_y are given by:

$$(3) \quad \begin{cases} \frac{\partial \rho}{\partial \mathbf{w}_x} \hat{=} \mathbf{C}_{xy} \hat{\mathbf{w}}_y - \frac{\hat{\mathbf{w}}_x^T \mathbf{C}_{xy} \hat{\mathbf{w}}_y}{\hat{\mathbf{w}}_x^T \mathbf{C}_{xx} \hat{\mathbf{w}}_x} \mathbf{C}_{xx} \hat{\mathbf{w}}_x \\ \frac{\partial \rho}{\partial \mathbf{w}_y} \hat{=} \mathbf{C}_{yx} \hat{\mathbf{w}}_x - \frac{\hat{\mathbf{w}}_y^T \mathbf{C}_{yx} \hat{\mathbf{w}}_x}{\hat{\mathbf{w}}_y^T \mathbf{C}_{yy} \hat{\mathbf{w}}_y} \mathbf{C}_{yy} \hat{\mathbf{w}}_y \end{cases}$$

where ' $\hat{\cdot}$ ' indicates unit length and ' $\hat{=}$ ' means that the vectors, left and right, have the same directions. A complete description of the canonical correlations is given by:

$$(4) \quad \begin{bmatrix} \mathbf{C}_{xx} & [0] \\ [0] & \mathbf{C}_{yy} \end{bmatrix}^{-1} \begin{bmatrix} [0] & \mathbf{C}_{xy} \\ \mathbf{C}_{yx} & [0] \end{bmatrix} \begin{pmatrix} \hat{\mathbf{w}}_x \\ \hat{\mathbf{w}}_y \end{pmatrix} = \rho \begin{pmatrix} \lambda_x \hat{\mathbf{w}}_x \\ \lambda_y \hat{\mathbf{w}}_y \end{pmatrix}$$

where: $\rho, \lambda_x, \lambda_y > 0$ and $\lambda_x \lambda_y = 1$. Equation 4 can be rewritten as:

$$(5) \quad \begin{cases} \mathbf{C}_{xx}^{-1} \mathbf{C}_{xy} \hat{\mathbf{w}}_y = \rho \lambda_x \hat{\mathbf{w}}_x \\ \mathbf{C}_{yy}^{-1} \mathbf{C}_{yx} \hat{\mathbf{w}}_x = \rho \lambda_y \hat{\mathbf{w}}_y \end{cases}$$

Solving equation 5 gives N solutions $\{\rho_n, \hat{\mathbf{w}}_{xn}, \hat{\mathbf{w}}_{yn}\}$. N is the minimum of the input dimensionality and the output dimensionality. The linear combinations, $x_n = \hat{\mathbf{w}}_{xn}^T \mathbf{x}$ and $y_n = \hat{\mathbf{w}}_{yn}^T \mathbf{y}$, are termed *canonical variates* and the correlations, ρ_n , between these variates are termed the *canonical correlations* [3]. An important aspect in this context is that the canonical correlations are *invariant to affine transformations* of \mathbf{x} and \mathbf{y} . Also note that the canonical variates corresponding to the different roots of equation 5 are uncorrelated, implying that:

$$(6) \quad \begin{cases} \mathbf{w}_{xn}^T \mathbf{C}_{xx} \mathbf{w}_{xm} = 0 \\ \mathbf{w}_{yn}^T \mathbf{C}_{yy} \mathbf{w}_{ym} = 0 \end{cases} \quad \text{if } n \neq m$$

3 Learning Canonical Correlations

We have developed a novel learning algorithm that finds the canonical correlations and the corresponding canonical variates by an iterative method. The update rule for the vectors \mathbf{w}_x and \mathbf{w}_y is given by:

$$(7) \quad \begin{cases} \mathbf{w}_x \leftarrow \mathbf{w}_x + \alpha_x \mathbf{x} (\mathbf{y}^T \hat{\mathbf{w}}_y - \mathbf{x}^T \mathbf{w}_x) \\ \mathbf{w}_y \leftarrow \mathbf{w}_y + \alpha_y \mathbf{y} (\mathbf{x}^T \hat{\mathbf{w}}_x - \mathbf{y}^T \mathbf{w}_y) \end{cases}$$

where \mathbf{x} and \mathbf{y} both have the mean $\mathbf{0}$. To see that this rule finds the directions of the canonical correlation we look at the expected change, in one iteration, of the vectors, \mathbf{w}_x and \mathbf{w}_y :

$$\begin{cases} E\{\Delta \mathbf{w}_x\} = \alpha_x E\{\mathbf{x} \mathbf{y}^T \hat{\mathbf{w}}_y - \mathbf{x} \mathbf{x}^T \mathbf{w}_x\} = \alpha_x (\mathbf{C}_{xy} \hat{\mathbf{w}}_y - \|\mathbf{w}_x\| \mathbf{C}_{xx} \hat{\mathbf{w}}_x) \\ E\{\Delta \mathbf{w}_y\} = \alpha_y E\{\mathbf{y} \mathbf{x}^T \hat{\mathbf{w}}_x - \mathbf{y} \mathbf{y}^T \mathbf{w}_y\} = \alpha_y (\mathbf{C}_{yx} \hat{\mathbf{w}}_x - \|\mathbf{w}_y\| \mathbf{C}_{yy} \hat{\mathbf{w}}_y) \end{cases}$$

Identifying with equation 3 gives:

$$(8) \quad E\{\Delta \mathbf{w}_x\} \cong \frac{\partial \rho}{\partial \mathbf{w}_x} \quad \text{and} \quad E\{\Delta \mathbf{w}_y\} \cong \frac{\partial \rho}{\partial \mathbf{w}_y}$$

with

$$\|\mathbf{w}_x\| = \frac{\hat{\mathbf{w}}_x^T \mathbf{C}_{xy} \hat{\mathbf{w}}_y}{\hat{\mathbf{w}}_x^T \mathbf{C}_{xx} \hat{\mathbf{w}}_x} \quad \text{and} \quad \|\mathbf{w}_y\| = \frac{\hat{\mathbf{w}}_y^T \mathbf{C}_{yx} \hat{\mathbf{w}}_x}{\hat{\mathbf{w}}_y^T \mathbf{C}_{yy} \hat{\mathbf{w}}_y}$$

This shows that the expected changes of the vectors \mathbf{w}_x and \mathbf{w}_y are in the same directions as the gradient of the canonical correlation, ρ , which means that the learning rules in equation 7 on average is a gradient search on ρ . ρ , λ_x and λ_y are found as:

$$(9) \quad \rho = \sqrt{\|\mathbf{w}_x\| \|\mathbf{w}_y\|}; \quad \lambda_x = \lambda_y^{-1} = \sqrt{\frac{\|\mathbf{w}_x\|}{\|\mathbf{w}_y\|}}$$

3.1 Learning of successive canonical correlations

The learning rule maximizes the correlation and finds the directions, $\hat{\mathbf{w}}_{x1}$ and $\hat{\mathbf{w}}_{y1}$, corresponding to the largest correlation, ρ_1 . To find the second largest canonical correlation and the corresponding canonical variates of equation 5 we use the modified learning rule

$$(10) \quad \begin{cases} \mathbf{w}_x \leftarrow \mathbf{w}_x + \alpha_x \mathbf{x} ((\mathbf{y} - \mathbf{y}_1)^T \hat{\mathbf{w}}_y - \mathbf{x}^T \mathbf{w}_x) \\ \mathbf{w}_y \leftarrow \mathbf{w}_y + \alpha_y \mathbf{y} ((\mathbf{x} - \mathbf{x}_1)^T \hat{\mathbf{w}}_x - \mathbf{y}^T \mathbf{w}_y) \end{cases}$$

where

$$\mathbf{x}_1 = \frac{\mathbf{x}^T \hat{\mathbf{w}}_{x1} \mathbf{v}_{x1}}{\hat{\mathbf{w}}_{x1}^T \mathbf{v}_{x1}} \quad \text{and} \quad \mathbf{y}_1 = \frac{\mathbf{y}^T \hat{\mathbf{w}}_{y1} \mathbf{v}_{y1}}{\hat{\mathbf{w}}_{y1}^T \mathbf{v}_{y1}}.$$

\mathbf{v}_{x1} and \mathbf{v}_{y1} are estimates of $\mathbf{C}_{xx} \hat{\mathbf{w}}_{x1}$ and $\mathbf{C}_{yy} \hat{\mathbf{w}}_{y1}$ respectively and are estimated using the iterative rule:

$$(11) \quad \begin{cases} \mathbf{v}_{x1} \leftarrow \mathbf{v}_{x1} + \beta (\mathbf{x} \mathbf{x}^T \hat{\mathbf{w}}_{x1} - \mathbf{v}_{x1}) \\ \mathbf{v}_{y1} \leftarrow \mathbf{v}_{y1} + \beta (\mathbf{y} \mathbf{y}^T \hat{\mathbf{w}}_{y1} - \mathbf{v}_{y1}) \end{cases}$$

The expected change of \mathbf{w}_x and \mathbf{w}_y is then given by

$$(12) \quad \begin{cases} E\{\Delta \mathbf{w}_x\} = \alpha_x \left(\mathbf{C}_{xy} \left[\hat{\mathbf{w}}_y - \hat{\mathbf{w}}_{y1} \frac{\hat{\mathbf{w}}_{y1}^T \mathbf{C}_{yy} \hat{\mathbf{w}}_y}{\hat{\mathbf{w}}_{y1}^T \mathbf{C}_{yy} \hat{\mathbf{w}}_{y1}} \right] - \|\mathbf{w}_x\| \mathbf{C}_{xx} \hat{\mathbf{w}}_x \right) \\ E\{\Delta \mathbf{w}_y\} = \alpha_y \left(\mathbf{C}_{yx} \left[\hat{\mathbf{w}}_x - \hat{\mathbf{w}}_{x1} \frac{\hat{\mathbf{w}}_{x1}^T \mathbf{C}_{xx} \hat{\mathbf{w}}_x}{\hat{\mathbf{w}}_{x1}^T \mathbf{C}_{xx} \hat{\mathbf{w}}_{x1}} \right] - \|\mathbf{w}_y\| \mathbf{C}_{yy} \hat{\mathbf{w}}_y \right) \end{cases}$$

It can be seen that the parts of \mathbf{w}_x and \mathbf{w}_y parallel to $\mathbf{C}_{xx} \hat{\mathbf{w}}_{x1}$ and $\mathbf{C}_{yy} \hat{\mathbf{w}}_{y1}$ respectively will vanish ($\Delta \mathbf{w}_x^T \mathbf{w}_{x1} \leq 0 \quad \forall \mathbf{x}$ and $\Delta \mathbf{w}_y^T \mathbf{w}_{y1} \leq 0 \quad \forall \mathbf{y}$ in equation 10). In the subspaces orthogonal to $\mathbf{C}_{xx} \hat{\mathbf{w}}_{x1}$ and $\mathbf{C}_{yy} \hat{\mathbf{w}}_{y1}$ the learning rule will be equivalent to that given by equation 7. In this way the parts of the signals correlated with $\mathbf{w}_{x1}^T \mathbf{x}$ (and $\mathbf{w}_{y1}^T \mathbf{y}$) are disregarded leaving the rest unchanged. Consequently the algorithm finds the second largest correlation ρ_2 and the corresponding vectors \mathbf{w}_{x2} and \mathbf{w}_{y2} . Successive canonical correlations can be found by repeating the procedure.

4 Experiments

The experiments carried out are intended to display the behaviour of the algorithm. The results show that the presented algorithm, which has complexity $\mathcal{O}(N)$, has a performance comparable to what can be obtained by estimating the sample covariance matrices and calculating the eigenvectors and eigenvalues explicitly (complexity $\mathcal{O}(N^3)$). The latter will be referred to as the optimal solutions.

4.1 Adaptive update rate

Rather than tuning parameters to produce a nice result for a specific distribution we have used adaptive update factors and parameters producing similar behaviour for different distributions and different number of dimensions. Also note that the adaptability allows a system without a pre-specified time dependent update rate decay. The coefficients α_x and α_y were in the experiments calculated according to equation 13.

$$(13) \quad \begin{cases} E_x \leftarrow E_x + b (\|\mathbf{x}\mathbf{x}^T \mathbf{w}_x\| - E_x) \\ E_y \leftarrow E_y + b (\|\mathbf{y}\mathbf{y}^T \mathbf{w}_y\| - E_y) \\ \alpha_x = a \lambda_x E_x^{-1} \\ \alpha_y = a \lambda_y E_y^{-1} \end{cases}$$

4.2 Adaptive smoothing

To get a smooth and yet fast behaviour an adaptively time averaged set of vectors, \mathbf{w}_a was calculated. The update speed was made dependent on the consistency in the change of the original vectors \mathbf{w} according to equation 14.

$$(14) \quad \begin{cases} \Delta_x \leftarrow \Delta_x + d (\Delta \mathbf{w}_x - \Delta_x) \\ \Delta_y \leftarrow \Delta_y + d (\Delta \mathbf{w}_y - \Delta_y) \\ \mathbf{w}_{ax} \leftarrow \mathbf{w}_{ax} + c \|\Delta_x\| \|\mathbf{w}_x\|^{-1} (\mathbf{w}_x - \mathbf{w}_{ax}) \\ \mathbf{w}_{ay} \leftarrow \mathbf{w}_{ay} + c \|\Delta_y\| \|\mathbf{w}_y\|^{-1} (\mathbf{w}_y - \mathbf{w}_{ay}) \end{cases}$$

4.3 Results

The experiments have been carried out using a randomly chosen distribution of a 10-dimensional \mathbf{x} variable and a 5-dimensional \mathbf{y} variable. Two \mathbf{x} and two \mathbf{y} dimensions were partly correlated. The other 8 dimensions of \mathbf{x} and 3 dimensions of \mathbf{y} were uncorrelated. The variances in the 15 dimensions are in the same order of magnitude. The two canonical correlations for this distribution were 0.98 and 0.80. The parameters used in the experiments were $a = 0.1$, $b = 0.05$, $c = 0.01$, $d = 4$ and $\beta = 0.01$. 10 runs of 2000 iterations have been performed. For each run error measures were calculated. The errors shown in figure 1 are the averages over the 10 runs. The errors in directions for the vectors \mathbf{w}_{ax1} , \mathbf{w}_{ax2} , \mathbf{w}_{ay1} and \mathbf{w}_{ay2} were calculated as the angle between the vectors and the exact solutions, $\hat{\mathbf{e}}$ (known from the \mathbf{x} \mathbf{y} sample distribution), i.e.

$$Err[\hat{\mathbf{w}}] = \arccos(\hat{\mathbf{w}}_a^T \hat{\mathbf{e}})$$

These measures are drawn with a solid line in the four top diagrams. As a comparison the error for the optimal solution was calculated for each run as

$$Err[\hat{\mathbf{w}}_{opt}] = \arccos(\hat{\mathbf{w}}_{opt}^T \hat{\mathbf{e}})$$

where \mathbf{w}_{opt} were calculated by solving the eigenvalue equations for the actual sample covariance matrices. These errors are drawn with dotted lines in the same diagrams. Finally the errors in the estimations of canonical correlations were calculated as:

$$Err[\text{Corr}] = \left| \frac{\rho_n}{\rho_{en}} - 1 \right|$$

where ρ_{en} are the exact solutions. The results are plotted with solid lines in the bottom diagrams. Again the corresponding errors for the optimal solutions were calculated and drawn with dotted lines in the same diagrams.

It should be pointed out that using a significantly higher dimensionality was prohibited by the time required for computing the optimal solutions. Even for the low dimensionality used in the experiment obtaining the results for the optimal solutions required an order of magnitude more of computation time than the computations involved in the algorithm.

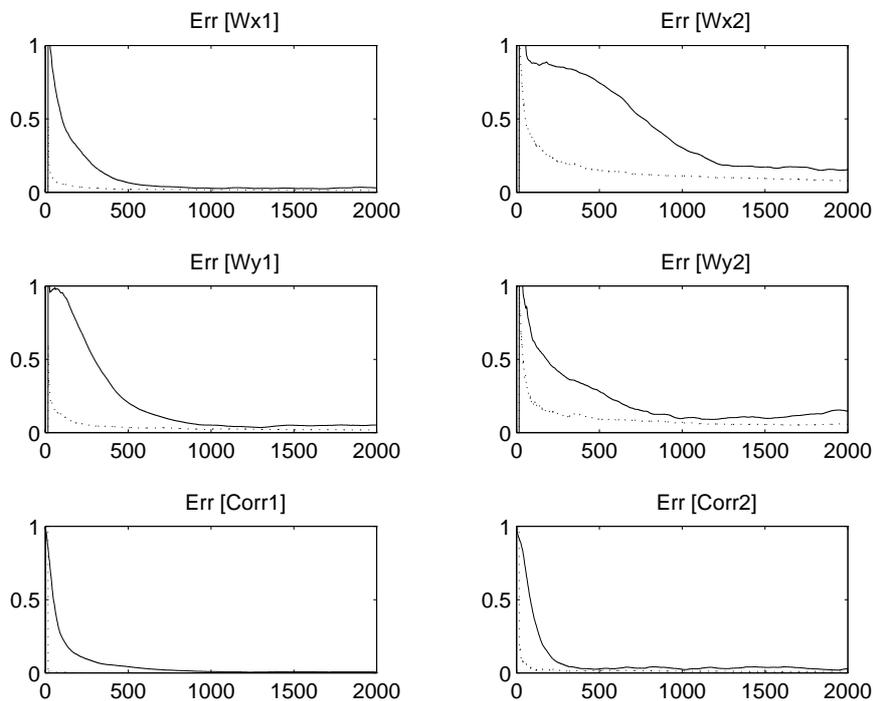


Figure 1: Error magnitudes averaged over 10 runs of the algorithm. The solid lines displays the differences between the algorithm and the exact values. The dotted lines shows the differences between the optimal solutions obtained by solving the eigenvector equations and the exact values, (see text for further explanation). The top row shows the error angles in radians for $\hat{\mathbf{w}}_{ax}$. The middle row shows the same errors for $\hat{\mathbf{w}}_{ay}$. The bottom row shows the relative error in the estimation of ρ . The left column shows results for the first canonical correlation and the right column shows the results for the second canonical correlation.

References

- [1] R. D. Bock. *Multivariate Statistical Methods in Behavioral Research*. McGraw-Hill series in psychology. McGraw-Hill, 1975.
- [2] M. Borga. Hierarchical reinforcement learning. In S. Gielen and B. Kappen, editors, *ICANN'93*, Amsterdam, September 1993. Springer-Verlag.
- [3] H. Hotelling. Relations between two sets of variables. *Biometrika*, 28:321–377, 1936.
- [4] T. Kohonen. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43:59–69, 1982.
- [5] T. Landelius and H. Knutsson. The learning tree, a new concept in learning. In *Proceedings of the 2:nd Int. Conf. on Adaptive and Learning Systems*. SPIE, April 1993.