A Unified Approach to PCA, PLS, MLR and CCA

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

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

Abstract

This paper presents a novel algorithm for analysis of stochastic processes. The algorithm can be used to find the required solutions in the cases of principal component analysis (PCA), partial least squares (PLS), canonical correlation analysis (CCA) or multiple linear regression (MLR). The algorithm is iterative and sequential in its structure and uses on-line stochastic approximation to reach an equilibrium point. A quotient between two quadratic forms is used as an energy function and it is shown that the equilibrium points constitute solutions to the generalized eigenproblem.

Keywords: Generalized eigenproblem, stochastic approximation, on-line algorithm, system learning, self-adaptation, principal components, partial least squares, canonical correlation, linear regression, reduced rank, mutual information, independent components.

1 Introduction

The ability to perform dimensionality reduction is crucial to systems exposed to high dimensional data e.g. images, image sequences [10], and even scalar signals where relations between a high number of different time instances need to be considered [6]. This can for example be done by projecting the data onto new basis vectors that span a subspace of lower dimensionality. Without detailed prior knowledge, a suitable basis can only be found using an adaptive approach [17, 18]. For signals with high dimensionality, d, an iterative algorithm for finding this basis must not exhibit a memory requirement nor a computational cost significantly exceeding $\mathcal{O}(d)$ per iteration. The employment of traditional techniques, involving matrix multiplications (having memory requirements of order $\mathcal{O}(d^2)$ and computational costs of order $\mathcal{O}(d^3)$), quickly become infeasible when signal space dimensionality increases.

The criteria for an appropriate the new basis is, of course, dependent on the application. One way of approaching this problem is to project the data on the subspace of maximum data *variation*, i.e. the subspace spanned by the largest *principal components*. There are a number of applications in signal processing where the largest eigenvalue and the corresponding eigenvector of input data correlation- or covariance matrices play an important role, e.g. image coding.

In applications where relations between two sets of data, e.g. process input and output, are considered an analysis can be done by finding the subspaces in the input and the output spaces for which the data *covariation* is maximized. These subspaces turn out to be the ones accompanying the largest singular values of the between sets covariance matrix [19].

In general, however, the input to a system comes from a set of different sensors and it is evident that the range (or variance) 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. In these cases the covariances between signals are not relevant. Here, *correlation* between input and output signals is a more appropriate target for analysis since this measure of input-output relations is invariant to the signal magnitudes.

Finally, when the goal is to predict a signal as well as possible in a least square error sense, the directions must be chosen so that this error measure is minimized. This corresponds to a

low-rank approximation of *multiple linear regression* also known as *reduced rank regression* [14] or as *redundancy analysis* [23].

An important problem with direct relation to the situations discussed above is the *generalized eigenproblem* or two-matrix eigenproblem [3, 9, 22]:

$$\mathbf{A}\hat{\mathbf{e}} = \lambda\mathbf{B}\hat{\mathbf{e}} \quad \text{or} \quad \mathbf{B}^{-1}\mathbf{A}\hat{\mathbf{e}} = \lambda\hat{\mathbf{e}}. \tag{1}$$

The next section will describe the generalized eigenproblem in some detail and show its relation to an energy function called the *Rayleigh quotient*. It is shown that four important problems emerges as special cases of the generalized eigenproblem: principal component analysis (PCA), partial least squares (PLS), canonical correlation analysis (CCA) and multiple linear regression (MLR). These analysis methods corresponds to finding the subspaces of maximum variance, maximum covariance, maximum correlation and minimum square error respectively.

In section 3 we present an iterative, $\mathcal{O}(d)$ algorithm that solves the generalized eigenproblem by a gradient search on the Rayleigh quotient. The solutions are found in a successive order beginning with the largest eigenvalue and corresponding eigenvector. It is shown how to apply this algorithm to obtain the required solutions in the special cases of PCA, PLS, CCA and MLR.

2 The generalized eigenproblem

When dealing with many scientific and engineering problems, some version of the generalized eigenproblem needs to be solved along the way.

In mechanics, the eigenvalues often corresponds to modes of vibration. In this paper, however, we will consider the case where the matrices $\mathbf{A}$ and $\mathbf{B}$ consist of components which are expectation values from stochastic processes. Furthermore, both matrices will be hermitian and, in addition, $\mathbf{B}$ will be positive definite.

The generalized eigenproblem is closely related to the the problem of finding the extremum points of a ratio of *quadratic forms*

$$r = \frac{\mathbf{w}^T\mathbf{A}\mathbf{w}}{\mathbf{w}^T\mathbf{B}\mathbf{w}} \tag{2}$$

where both $\mathbf{A}$ and $\mathbf{B}$ are hermitian and $\mathbf{B}$ is positive definite, i.e. a *metric* matrix. This ratio is known as the *Rayleigh quotient* and its critical points, i.e. the points of zero derivatives, will correspond to the eigensystem of the generalized eigenproblem. To see this, let us look at the gradient of r:

$$\frac{\partial r}{\partial \mathbf{w}} = \frac{2}{\mathbf{w}^T\mathbf{B}\mathbf{w}}(\mathbf{A}\mathbf{w} - r\mathbf{B}\mathbf{w}) = \alpha(\mathbf{A}\hat{\mathbf{w}} - r\mathbf{B}\hat{\mathbf{w}}), \tag{3}$$

where $\alpha = \alpha(\mathbf{w})$ is a positive scalar and "$\hat{}$" denotes a vector of unit length. Setting the gradient to $\mathbf{0}$ gives

$$\mathbf{A}\hat{\mathbf{w}} = r\mathbf{B}\hat{\mathbf{w}} \quad \text{or} \quad \mathbf{B}^{-1}\mathbf{A}\hat{\mathbf{w}} = r\hat{\mathbf{w}} \tag{4}$$

which is recognized as the generalized eigenproblem, eq. 1. The solutions r_i and $\hat{\mathbf{w}}_i$ are the eigenvalues and eigenvectors respectively of the matrix $\mathbf{B}^{-1}\mathbf{A}$. This means that the extremum points (i.e. points of zero derivative) of the Rayleigh quotient $r(\mathbf{w})$ are solutions to the corresponding generalized eigenproblem so that the eigenvalue is the extremum value of the quotient and the eigenvector is the corresponding parameter vector $\mathbf{w}$ of the quotient. As an ilustration, the Rayleigh quotient is plotted to the left in in figure 1 for two matrices $\mathbf{A}$ and $\mathbf{B}$. The quotient is plotted as the radius in different directions $\hat{\mathbf{w}}$. Note that the quotient is invariant to the norm of $\mathbf{w}$. The two eigenvalues are shown as circles with their radii corresponding to the eigenvalues. It can be seen that the eigenvectors $\mathbf{e}_1$ and $\mathbf{e}_2$ of the generalized eigenproblem coincides with the maximum and minimum values of the Rayleigh quotient. To the right in the same figure, the gradient of the Rayleigh quotient is illustrated as a function of the direction of $\mathbf{w}$. Note that the gradient is orthogonal to $\mathbf{w}$ (see equation 3). This means that a small change of $\mathbf{w}$ in the direction of the gradient can be seen as a rotation of $\mathbf{w}$. The arrows indicate the direction of this orientation and the radii of the 'blobs' correspons to the magnitude of the gradient. The figure shows that the directions of

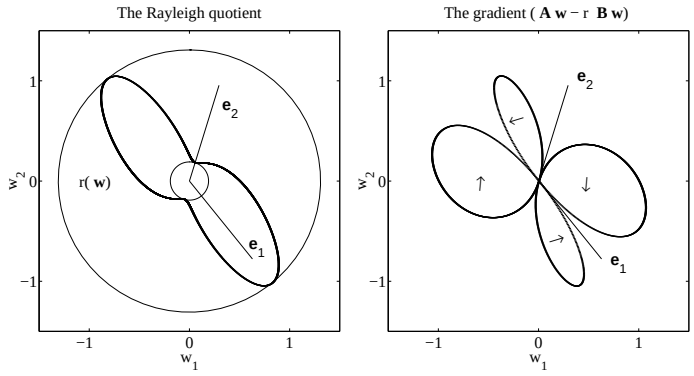

Figure 1: **Left:** *The Rayleigh quotient* $r(\mathbf{w})$ *between two matrices* $\mathbf{A}$ *and* $\mathbf{B}$*. The curve is plotted as* $r\hat{\mathbf{w}}$*. The eigenvectors of* $\mathbf{B}^{-1}\mathbf{A}$ *are marked as reference. The corresponding eigenvalues are marked as the radii of the two circles. Note that the quotient is invariant to the norm of* $\mathbf{w}$*.* **Right:** *The gradient of* r*. The arrows indicate he direction of this orientation and the radii of the 'blobs' correspons to the magnitude of the gradient.*

zero gradient coincides with the eigenvectors and that the gradient points towards the eigenvector coresponding to the largest eigenvalue.

If the eigenvalues r_i are distinct (i.e. $r_i \neq r_j$ for $i \neq j$), the different eigenvectors are orthogonal in the metrics $\mathbf{A}$ and $\mathbf{B}$ which means that

$$\hat{\mathbf{w}}_i^T \mathbf{B} \hat{\mathbf{w}}_j = \begin{cases} 0 & \text{for} \quad i \neq j \\ \beta_i > 0 & \text{for} \quad i = j \end{cases} \quad \text{and} \quad \hat{\mathbf{w}}_i^T \mathbf{A} \hat{\mathbf{w}}_j = \begin{cases} 0 & \text{for} \quad i \neq j \\ r_i \beta_i & \text{for} \quad i = j \end{cases} \tag{5}$$

(see proof 6.1). This means that the $\mathbf{w}_i$s are linearly independent (see proof 6.2). Since an n-dimensional space gives n eigenvectors which are linearly independent, hence, $\{\mathbf{w}_1, \ldots, \mathbf{w}_n\}$ constitutes a *base* and any $\mathbf{w}$ can be expressed as a linear combination of the eigenvectors. Now, it can be proved (see proof 6.3) that the function r is bounded by the largest and smallest eigenvalue, i.e.

$$r_n \leq r \leq r_1 \tag{6}$$

which means that there exists a global maximum and that this maximum is r_1.

To investigate if there are any other local maxima, we look at the second derivative, or the *Hessian* $\mathbf{H}$, of r for the solutions of the eigenproblem,

$$\mathbf{H}_i = \left. \frac{\partial^2 r}{\partial \mathbf{w}^2} \right|_{\mathbf{w}=\hat{\mathbf{w}}_i} = \frac{2}{\hat{\mathbf{w}}_i^T \mathbf{B} \hat{\mathbf{w}}_i} (\mathbf{A} - r_i \mathbf{B}) \tag{7}$$

(see proof 6.4). It can be shown (see proof 6.5) that the Hessian $\mathbf{H}_i$ have got positive eigenvalues for $i > 1$, i.e. there exits vectors $\mathbf{w}$ such that

$$\mathbf{w}^T \mathbf{H}_i \mathbf{w} > 0 \quad \forall\, i > 1 \tag{8}$$

This means that for all solutions to the eigenproblem except for the largest root, there exists a direction in which r increases. In other words, all extremum points of the function r are saddle points except for the global minimum and maximum points. Since the two-dimensional example in figure 1 only has two eigenvalues, as illustrated in the figure, they correspond to the maximum and minimum values of r.

We will now show that finding the directions of maximum variance, maximum covariance, maximum correlation and minimum square error can be seen as special cases of the generalized eigenproblem.

2.1 Direction of maximum data variation

For a set of random numbers $\{x_k\}$ with zero mean, the variance is defined as $E\{xx\}$. Now let us turn to a set of random vectors, with zero mean. In this case we consider the covariance matrix, defined by:

$$\mathbf{C}_{xx} = E\{\mathbf{x}\mathbf{x}^T\} \tag{9}$$

By the direction of maximum data variation we mean the direction $\hat{\mathbf{w}}$ with the property that the linear combination $x = \hat{\mathbf{w}}^T\mathbf{x}$ posses maximum variance. Hence, finding this direction is hence equivalent to finding the maximum of

$$\rho = E\{xx\} = E\{\hat{\mathbf{w}}^T\mathbf{x}\hat{\mathbf{w}}^T\mathbf{x}\} = \hat{\mathbf{w}}^T E\{\mathbf{x}\mathbf{x}^T\}\hat{\mathbf{w}} = \frac{\mathbf{w}^T\mathbf{C}_{xx}\mathbf{w}}{\mathbf{w}^T\mathbf{w}}. \tag{10}$$

This problem is a special case of that presented in eq. 2 with

$$\mathbf{A} = \mathbf{C}_{xx} \quad \text{and} \quad \mathbf{B} = \mathbf{I}. \tag{11}$$

Since the covariance matrix is symmetric, it is possible to expand it in its eigenvalues and orthogonal eigenvectors as:

$$\mathbf{C}_{xx} = E\{\mathbf{x}\mathbf{x}^T\} = \sum \lambda_i \ \hat{\mathbf{e}}_i\hat{\mathbf{e}}_i^T \tag{12}$$

where λ_i and $\hat{\mathbf{e}}_i$ are the eigenvalues and orthogonal eigenvectors respectively. This is known as principal component analysis (PCA). Hence, the problem of maximizing the variance, ρ, can be seen as the problem of finding the largest eigenvalue, λ_1, and its corresponding eigenvector since:

$$\lambda_1 = \hat{\mathbf{e}}_1^T\mathbf{C}_{xx}\hat{\mathbf{e}}_1 = \max\frac{\mathbf{w}^T\mathbf{C}_{xx}\mathbf{w}}{\mathbf{w}^T\mathbf{w}} = \max\rho. \tag{13}$$

It is also worth noting that it is possible to find the direction and magnitude of maximum data variation to the inverse of the covariance matrix. In this case we simply identify the matrices in eq. 2 as $\mathbf{A} = \mathbf{I}$ and $\mathbf{B} = \mathbf{C}_{xx}$.

2.2 Directions of maximum data covariation

Given two sets of random numbers with zero mean, $\{x_k\}$ and $\{y_k\}$, their covariance is defined as $E\{xy\} = E\{yx\}$. If we consider the multivariate case, we can define the between sets covariance matrix according to:

$$\mathbf{C}_{xy} = E\{\mathbf{x}\mathbf{y}^T\} \tag{14}$$

This time we look at the *two* directions of maximal data covariation, by which we mean the directions, $\hat{\mathbf{w}}_x$ and $\hat{\mathbf{w}}_y$, such that the linear combinations $x = \hat{\mathbf{w}}_x^T\mathbf{x}$ and $y = \hat{\mathbf{w}}_y^T\mathbf{y}$ gives maximum covariance. This means that we want to maximize the following function:

$$\rho = E\{xy\} = E\{\hat{\mathbf{w}}_x^T\mathbf{x}\hat{\mathbf{w}}_y^T\mathbf{y}\} = \hat{\mathbf{w}}_x^T E\{\mathbf{x}\mathbf{y}^T\}\hat{\mathbf{w}}_y = \frac{\mathbf{w}_x^T\mathbf{C}_{xy}\mathbf{w}_y}{\sqrt{\mathbf{w}_x^T\mathbf{w}_x\mathbf{w}_y^T\mathbf{w}_y}}. \tag{15}$$

Note that, for each ρ, a corresponding value $-\rho$ is obtained by rotating $\mathbf{w}_x$ or $\mathbf{w}_y$ 180°. For this reason, we obtain the maximum *magnitude* of ρ by finding the largest positive value.

This function cannot be written as a Rayleigh quotient. However, the critical points of this function coincides with the critical points of a Rayleigh quotient with proper choices of $\mathbf{A}$ and $\mathbf{B}$. To see this, we calculate the derivatives of this function with respect to the vectors $\mathbf{w}_x$ and $\mathbf{w}_y$ (see proof 6.6):

$$\begin{cases} \frac{\partial\rho}{\partial\mathbf{w}_x} &= \frac{1}{\|\mathbf{w}_x\|}(\mathbf{C}_{xy}\hat{\mathbf{w}}_y - \rho\hat{\mathbf{w}}_x) \\ \frac{\partial\rho}{\partial\mathbf{w}_y} &= \frac{1}{\|\mathbf{w}_y\|}(\mathbf{C}_{yx}\hat{\mathbf{w}}_x - \rho\hat{\mathbf{w}}_y). \end{cases} \tag{16}$$

Setting these expressions to zero and solving for $\mathbf{w}_x$ and $\mathbf{w}_y$ results in

$$\begin{cases} \mathbf{C}_{xy}\mathbf{C}_{yx}\hat{\mathbf{w}}_x &= \rho^2\hat{\mathbf{w}}_x \\ \mathbf{C}_{yx}\mathbf{C}_{xy}\hat{\mathbf{w}}_y &= \rho^2\hat{\mathbf{w}}_y. \end{cases} \tag{17}$$

This is exactly the same result as that obtained after a gradient search on r in eq. 2 if the matrices $\mathbf{A}$ and $\mathbf{B}$ and the vector $\mathbf{w}$ are chosen according to:

$$\mathbf{A} = \begin{pmatrix} \mathbf{0} & \mathbf{C}_{xy} \\ \mathbf{C}_{yx} & \mathbf{0} \end{pmatrix} \quad , \quad \mathbf{B} = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \quad \text{, and} \quad \mathbf{w} = \begin{pmatrix} \mu_x\hat{\mathbf{w}}_x \\ \mu_y\hat{\mathbf{w}}_y \end{pmatrix}. \tag{18}$$

This is easily verified by insertion of the expressions above into eq. 4 which results in

$$\begin{cases} \mathbf{C}_{xy}\hat{\mathbf{w}}_y &= r\frac{\mu_x}{\mu_y}\hat{\mathbf{w}}_x \\ \mathbf{C}_{yx}\hat{\mathbf{w}}_x &= r\frac{\mu_y}{\mu_x}\hat{\mathbf{w}}_y \end{cases} \tag{19}$$

and then solving for $\mathbf{w}_x$ and $\mathbf{w}_y$ which gives equation 17 with $r^2 = \rho^2$. Hence, the problem of finding the direction and magnitude of the largest data covariation can be seen as maximizing a special case of eq. 2 with the appropriate choice of matrices.

The between sets covariance matrix can be expanded by means of singular value decomposition (SVD) where the two sets of vectors $\{\hat{\mathbf{e}}_{xi}\}$ and $\{\hat{\mathbf{e}}_{yi}\}$ are mutually orthogonal:

$$\mathbf{C}_{xy} = E_{xy}\{\mathbf{x}\mathbf{y}^T\} = \sum \lambda_i \ \hat{\mathbf{e}}_{xi}\hat{\mathbf{e}}_{yi}^T \tag{20}$$

where the positive numbers, λ_i, are referred to as the singular values. Since the basis vectors are orthogonal, we see that the problem of maximizing the quotient in eq. 15 is equivalent to finding the largest singular value:

$$\lambda_1 = \hat{\mathbf{e}}_{x1}^T\mathbf{C}_{xy}\hat{\mathbf{e}}_{y1} = \max \frac{\mathbf{w}_x^T\mathbf{C}_{xy}\mathbf{w}_y}{\sqrt{\mathbf{w}_x^T\mathbf{w}_x\mathbf{w}_y^T\mathbf{w}_y}} = \max \rho. \tag{21}$$

The SVD of a between-sets covariance matrix has a direct relation to the method of partial least squares (PLS) [13, 25].

2.3 Directions of maximum data correlation

Again, consider two random variables $\mathbf{x}$ and $\mathbf{y}$ with zero mean and stemming from a multi-normal distribution with

$$\mathbf{C} = \begin{pmatrix} \mathbf{C}_{xx} & \mathbf{C}_{xy} \\ \mathbf{C}_{yx} & \mathbf{C}_{yy} \end{pmatrix} = E\left\{ \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix}^T \right\} \tag{22}$$

as the covariance matrix. Consider the linear combinations $x = \hat{\mathbf{w}}_x^T\mathbf{x}$ and $y = \hat{\mathbf{w}}_y^T\mathbf{y}$ of the two variables respectively. The correlation[1] between x and y is defined as $E\{xy\}/\sqrt{E\{xx\}E\{yy\}}$. This means that the function we want to maximize can be written as

$$\rho = \frac{E\{xy\}}{\sqrt{E\{xx\}E\{yy\}}} = \frac{E\{\hat{\mathbf{w}}_x^T\mathbf{x}\mathbf{y}^T\hat{\mathbf{w}}_y\}}{\sqrt{E\{\hat{\mathbf{w}}_x^T\mathbf{x}\mathbf{x}^T\hat{\mathbf{w}}_x\}E\{\hat{\mathbf{w}}_y^T\mathbf{y}\mathbf{y}^T\hat{\mathbf{w}}_y\}}} = \frac{\mathbf{w}_x^T\mathbf{C}_{xy}\mathbf{w}_y}{\sqrt{\mathbf{w}_x^T\mathbf{C}_{xx}\mathbf{w}_x\mathbf{w}_y^T\mathbf{C}_{yy}\mathbf{w}_y}}. \tag{23}$$

Also in this case, as ρ changes sign if $\mathbf{w}_x$ or $\mathbf{w}_y$ is rotated 180°, it is sufficient to find the positive values.

Just like equation 15, this function cannot be written as a Rayleigh quotient. But also in this case, we can show that the critical points of this function coincides with the critical points of a

[1] The term correlation is some times inappropriately used to denote the second order *origin* moment (Σx^2) as opposed to *variance* which is the second order *central* moment ($\Sigma[x-x_0]^2$). The definition used here can be found in textbooks in mathematical statistics. It can loosely be described as the covariance between two variables normalized with the geometric mean of the variables' variances.

Rayleigh quotient with proper choices of $\mathbf{A}$ and $\mathbf{B}$. The partial derivatives of ρ with respect to $\mathbf{w}_x$ and $\mathbf{w}_y$ are (see proof 6.7)

$$\begin{cases} \frac{\partial \rho}{\partial \mathbf{w}_x} &= \frac{a}{\|\mathbf{w}_x\|}\left(\mathbf{C}_{xy}\hat{\mathbf{w}}_y - \frac{\hat{\mathbf{w}}_x^T\mathbf{C}_{xy}\hat{\mathbf{w}}_y}{\hat{\mathbf{w}}_x^T\mathbf{C}_{xx}\hat{\mathbf{w}}_x}\mathbf{C}_{xx}\hat{\mathbf{w}}_x\right) \\ \frac{\partial \rho}{\partial \mathbf{w}_y} &= \frac{a}{\|\mathbf{w}_y\|}\left(\mathbf{C}_{yx}\hat{\mathbf{w}}_x - \frac{\hat{\mathbf{w}}_y^T\mathbf{C}_{yx}\hat{\mathbf{w}}_x}{\hat{\mathbf{w}}_y^T\mathbf{C}_{yy}\hat{\mathbf{w}}_y}\mathbf{C}_{yy}\hat{\mathbf{w}}_y\right) \end{cases} \tag{24}$$

where a is a positive scalar. Setting the derivatives to zero gives the equation system

$$\begin{cases} \mathbf{C}_{xy}\hat{\mathbf{w}}_y = \rho\lambda_x\mathbf{C}_{xx}\hat{\mathbf{w}}_x \\ \mathbf{C}_{yx}\hat{\mathbf{w}}_x = \rho\lambda_y\mathbf{C}_{yy}\hat{\mathbf{w}}_y \end{cases} \tag{25}$$

where

$$\lambda_x = \lambda_y^{-1} = \sqrt{\frac{\hat{\mathbf{w}}_y^T\mathbf{C}_{yy}\hat{\mathbf{w}}_y}{\hat{\mathbf{w}}_x^T\mathbf{C}_{xx}\hat{\mathbf{w}}_x}}. \tag{26}$$

λ_x is the ratio between the standard deviation of y and the standard deviation of x and vice versa. The λ's can be interpreted as a scaling factor between the linear combinations. Rewriting equation system 25 gives (see proof 6.9)

$$\begin{cases} \mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}\mathbf{C}_{yy}^{-1}\mathbf{C}_{yx}\hat{\mathbf{w}}_x = \rho^2\hat{\mathbf{w}}_x \\ \mathbf{C}_{yy}^{-1}\mathbf{C}_{yx}\mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}\hat{\mathbf{w}}_y = \rho^2\hat{\mathbf{w}}_y. \end{cases} \tag{27}$$

Hence, $\hat{\mathbf{w}}_x$ and $\hat{\mathbf{w}}_y$ are found as the eigenvectors to $\mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}\mathbf{C}_{yy}^{-1}\mathbf{C}_{yx}$ and $\mathbf{C}_{yy}^{-1}\mathbf{C}_{yx}\mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}$ respectively. The corresponding eigenvalues ρ^2 are the squared *canonical correlations* [4, 5, 24, 12, 16]. The eigenvectors corresponding to the largest eigenvalue ρ_1^2 are the vectors $\hat{\mathbf{w}}_{x1}$ and $\hat{\mathbf{w}}_{y1}$ that maximizes the correlation between the *canonical variates* $x_1 = \hat{\mathbf{w}}_{x1}^T\mathbf{x}$ and $y_1 = \hat{\mathbf{w}}_{y1}^T\mathbf{y}$.

Now, if we let

$$\mathbf{A} = \begin{pmatrix} \mathbf{0} & \mathbf{C}_{xy} \\ \mathbf{C}_{yx} & \mathbf{0} \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} \mathbf{C}_{xx} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{yy} \end{pmatrix}, \quad \text{and} \quad \mathbf{w} = \begin{pmatrix} \mathbf{w}_x \\ \mathbf{w}_y \end{pmatrix} = \begin{pmatrix} \mu_x\hat{\mathbf{w}}_x \\ \mu_y\hat{\mathbf{w}}_y \end{pmatrix} \tag{28}$$

we can write equation 4 as

$$\begin{cases} \mathbf{C}_{xy}\hat{\mathbf{w}}_y = r\frac{\mu_x}{\mu_y}\mathbf{C}_{xx}\hat{\mathbf{w}}_x \\ \mathbf{C}_{yx}\hat{\mathbf{w}}_x = r\frac{\mu_y}{\mu_x}\mathbf{C}_{yy}\hat{\mathbf{w}}_y \end{cases} \tag{29}$$

which we recognize as equation 25 for $\rho\lambda_x = r\frac{\mu_x}{\mu_y}$ and $\rho\lambda_y = r\frac{\mu_y}{\mu_x}$. If we solve for $\mathbf{w}_x$ and $\mathbf{w}_y$ in eq. 29, we will end up in eq. 27 with $r^2 = \rho^2$. This shows that we obtain the equations for the canonical correlations as the result of a maximizing the energy function r.

An important property of canonical correlations is that they are invariant with respect to affine transformations of $\mathbf{x}$ and $\mathbf{y}$. An affine transformation is given by a translation of the origin followed by a linear transformation. The translation of the origin of $\mathbf{x}$ or $\mathbf{y}$ has no effect on ρ since it leaves the covariance matrix $\mathbf{C}$ unaffected. Invariance with respect to scalings of $\mathbf{x}$ and $\mathbf{y}$ follows directly from equation 23. For invariance with respect to other linear transformations see proof 6.10.

2.4 Directions for minimum square error

Again, consider two random variables $\mathbf{x}$ and $\mathbf{y}$ with zero mean and stemming from a multi-normal distribution with covariance as in equation 22. In this case, we want to minimize the square error

$$\begin{aligned} \epsilon^2 &= E\{\|\mathbf{y} - \beta\hat{\mathbf{w}}_y\hat{\mathbf{w}}_x^T\mathbf{x}\|^2\} \\ &= E\{\mathbf{y}^T\mathbf{y} - 2\beta\mathbf{y}^T\hat{\mathbf{w}}_y\hat{\mathbf{w}}_x^T\mathbf{x} + \beta^2\hat{\mathbf{w}}_x^T\mathbf{x}\mathbf{x}^T\hat{\mathbf{w}}_x\hat{\mathbf{w}}_y^T\} \\ &= E\{\mathbf{y}^T\mathbf{y}\} - 2\beta\hat{\mathbf{w}}_y^T\mathbf{C}_{yx}\hat{\mathbf{w}}_x + \beta^2\hat{\mathbf{w}}_x^T\mathbf{C}_{xx}\hat{\mathbf{w}}_x, \end{aligned} \tag{30}$$

i.e. a rank-one approximation of the MLR of $\mathbf{y}$ onto $\mathbf{x}$ based on minimum square error. The problem is to find not only the regression coefficient β, but also the optimal basis $\hat{\mathbf{w}}_x$ and $\hat{\mathbf{w}}_y$. To get an expression for β, we calculate the derivative

$$\frac{\partial \epsilon^2}{\partial \beta} = 2\left(\beta \hat{\mathbf{w}}_x^T \mathbf{C}_{xx} \hat{\mathbf{w}}_x - \hat{\mathbf{w}}_y^T \mathbf{C}_{yx} \hat{\mathbf{w}}_x\right) = 0, \tag{31}$$

which gives

$$\beta = \frac{\hat{\mathbf{w}}_y^T \mathbf{C}_{yx} \hat{\mathbf{w}}_x}{\hat{\mathbf{w}}_x^T \mathbf{C}_{xx} \hat{\mathbf{w}}_x}. \tag{32}$$

By inserting this expression into eq. 30 we get

$$\epsilon^2 = E\{\mathbf{y}^T \mathbf{y}\} - \frac{(\hat{\mathbf{w}}_y^T \mathbf{C}_{yx} \hat{\mathbf{w}}_x)^2}{\hat{\mathbf{w}}_x^T \mathbf{C}_{xx} \hat{\mathbf{w}}_x}. \tag{33}$$

Since ϵ^2 cannot be negative and the left term is independent of the parameters, we can minimize ϵ^2 by maximizing the quotient to the right in eq. 33, i.e. maximizing the quotient

$$\rho = \frac{\hat{\mathbf{w}}_x^T \mathbf{C}_{xy} \hat{\mathbf{w}}_y}{\sqrt{\hat{\mathbf{w}}_x^T \mathbf{C}_{xx} \hat{\mathbf{w}}_x}} = \frac{\mathbf{w}_x^T \mathbf{C}_{xy} \mathbf{w}_y}{\sqrt{\mathbf{w}_x^T \mathbf{C}_{xx} \mathbf{w}_x \mathbf{w}_y^T \mathbf{w}_y}}. \tag{34}$$

Note that if $\mathbf{w}_x$ and $\mathbf{w}_y$ minimizes ϵ^2, the negation of one or both of these vectors will give the same minimum. Hence, it is sufficient to maximize the positive root. The square of this quotient, i.e. ρ^2, is also known as the *redundancy index* [21] in the rank one case.

As in the two previous cases, while this function cannot not be written as a Rayleigh quotient, we can show that its critical points coincides with the critical points of a Rayleigh quotient with proper choices of $\mathbf{A}$ and $\mathbf{B}$. The partial derivatives of ρ with respect to $\mathbf{w}_x$ and $\mathbf{w}_y$ are (see proof 6.8)

$$\begin{cases} \frac{\partial \rho}{\partial \mathbf{w}_x} &= \frac{a}{\|\mathbf{w}_x\|}\left(\mathbf{C}_{xy} \hat{\mathbf{w}}_y - \beta \mathbf{C}_{xx} \hat{\mathbf{w}}_x\right) \\ \frac{\partial \rho}{\partial \mathbf{w}_y} &= \frac{a}{\|\mathbf{w}_x\|}\left(\mathbf{C}_{yx} \hat{\mathbf{w}}_x - \frac{\rho^2}{\beta} \hat{\mathbf{w}}_y\right). \end{cases} \tag{35}$$

Setting the derivatives to zero gives the equation system

$$\begin{cases} \mathbf{C}_{xy} \hat{\mathbf{w}}_y = \beta \mathbf{C}_{xx} \hat{\mathbf{w}}_x \\ \mathbf{C}_{yx} \hat{\mathbf{w}}_x = \frac{\rho^2}{\beta} \hat{\mathbf{w}}_y, \end{cases} \tag{36}$$

which gives

$$\begin{cases} \mathbf{C}_{xx}^{-1} \mathbf{C}_{xy} \mathbf{C}_{yx} \hat{\mathbf{w}}_x = \rho^2 \hat{\mathbf{w}}_x \\ \mathbf{C}_{yx} \mathbf{C}_{xx}^{-1} \mathbf{C}_{xy} \hat{\mathbf{w}}_y = \rho^2 \hat{\mathbf{w}}_y. \end{cases} \tag{37}$$

Now, if we let

$$\mathbf{A} = \begin{pmatrix} \mathbf{0} & \mathbf{C}_{xy} \\ \mathbf{C}_{yx} & \mathbf{0} \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} \mathbf{C}_{xx} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}, \quad \text{and} \quad \mathbf{w} = \begin{pmatrix} \mathbf{w}_x \\ \mathbf{w}_y \end{pmatrix} = \begin{pmatrix} \mu_x \hat{\mathbf{w}}_x \\ \mu_y \hat{\mathbf{w}}_y \end{pmatrix} \tag{38}$$

we can write equation 4 as

$$\begin{cases} \mathbf{C}_{xy} \hat{\mathbf{w}}_y = r \frac{\mu_x}{\mu_y} \mathbf{C}_{xx} \hat{\mathbf{w}}_x \\ \mathbf{C}_{yx} \hat{\mathbf{w}}_x = r \frac{\mu_y}{\mu_x} \hat{\mathbf{w}}_y \end{cases} \tag{39}$$

which we recognize as equation 36 for $\beta = r\frac{\mu_x}{\mu_y}$ and $\frac{\rho^2}{\beta} = r\frac{\mu_y}{\mu_x}$. If we solve for $\mathbf{w}_x$ and $\mathbf{w}_y$ in eq. 39 we will end up in eq. 37 with $r^2 = \rho^2$. This shows that we minimize the square error in eq. 30 as a result of maximizing the energy function r in eq. 2 for the proper choice of regression coefficient β.

It should be noted that the regression coefficient β defined in eq. 32 is valid for any choice of $\hat{\mathbf{w}}_x$ and $\hat{\mathbf{w}}_y$. In particular, if we use the directions of maximum variance, β is the regression coefficient for *principal components regression* (PCR) and for the directions of maximum covariance, β is the regression coefficient for PLS regression.

2.5 Examples

To see how these four different special cases of the generalized eigenproblem may differ, the solutions for the same data is plotted in figure 2. The data is two-dimensional in X and Y and randomly distributed with zero mean. The top row shows the eigenvectors in the X-space for the CCA, MLR, PLS and PCA respectively. The bottom row shows the solutions in the Y-space. Note that all solutions except the two solutions for CCA and the X-solution for MLR are orthogonal. Figure 3 shows the correlation, mean square error, covariance and variance of the data projected onto the first eigenvectors for each method. It can be seen that: The correlation is maximized for the CCA solution; The mean square error is minimized for the MLR solution. The covariance is maximized for the PLS solution. The variance is maximized for the PCA solution.

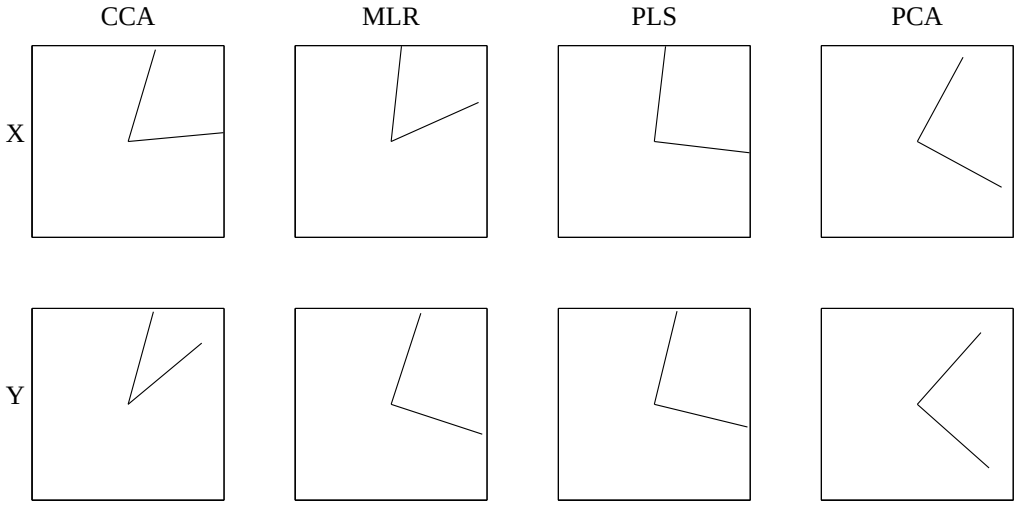

Figure 2: *Examples of eigenvectors using CCA, MLR, PLS and CCA on the same sets of data.*

3 The algorithm

We will now show that we can find the solutions to the generalized eigenproblem and, hence, perform PCA, PLS, CCA or MLR by doing a gradient search on the Rayleigh quotient.

Finding the largest eigenvalue In the previous section, it was shown that the only stable critical point of the Rayleigh quotient is the global maximum (eq. 8). This means that it should be possible to find the largest eigenvalue of the generalized eigenproblem and its corresponding eigenvector by performing a gradient search on the energy function r. This can be done with an iterative algorithm:

$$\mathbf{w}(t+1) = \mathbf{w}(t) + \Delta\mathbf{w}(t), \tag{40}$$

where the update vector $\Delta\mathbf{w}$, on average, lies in the direction of the gradient:

$$E\{\Delta\mathbf{w}\} = \beta\frac{\partial r}{\partial \mathbf{w}} = \alpha(\mathbf{A}\hat{\mathbf{w}} - r\mathbf{B}\hat{\mathbf{w}}) \tag{41}$$

where α and β are positive numbers. α is the gain controlling how far, in the direction of the gradient, the vector estimate is updated at each iteration. This gain could be constant as well as data or time dependent.

In all four cases treated in this article, $\mathbf{A}$ has got at least one positive eigenvalue, i.e. there exist an $r > 0$. We can then use an update rule such that

$$E\{\Delta\mathbf{w}\} = \alpha(\mathbf{A}\hat{\mathbf{w}} - \mathbf{B}\mathbf{w}) \tag{42}$$

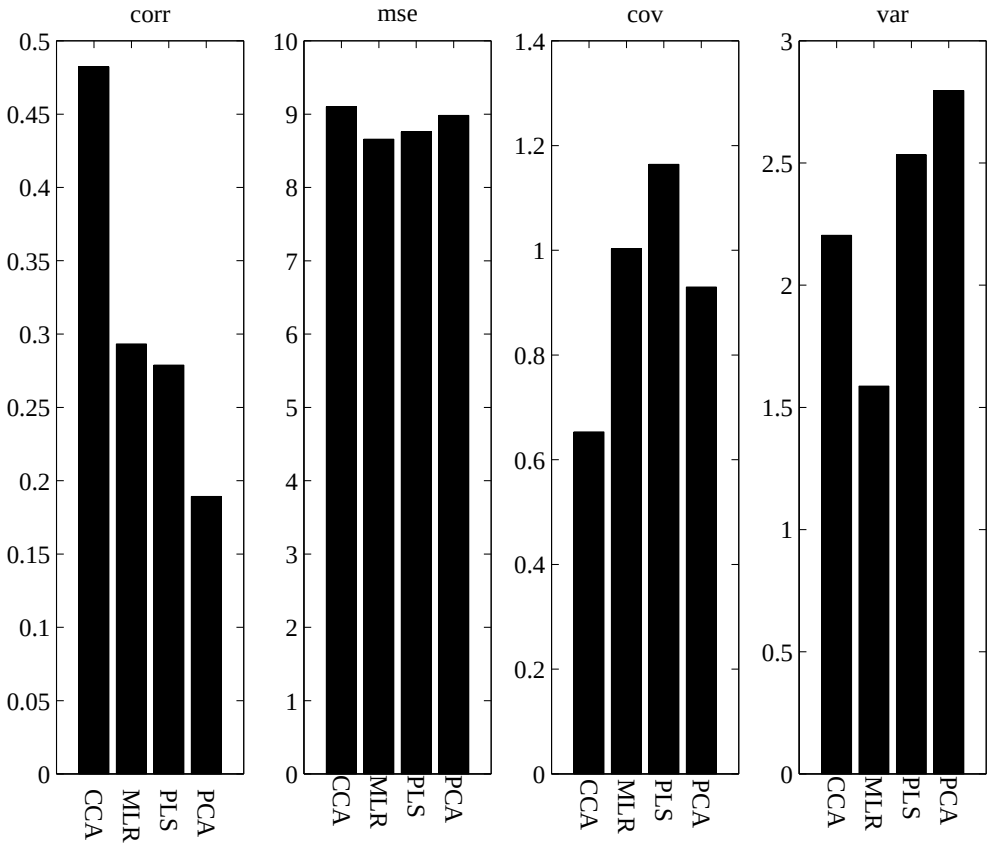

Figure 3: *The correlation, mean square error, covariance and variance when using the first pair of vectors for each method. The correlation is maximized for the CCA solution; The mean square error is minimized for the MLR solution. The covariance is maximized for the PLS solution. The variance is maximized for the PCA solution. (See section 2.5)*

to find the positive eigenvalues. Here, the length of the vector will represent the corresponding eigenvalue, i.e. $\|\mathbf{w}\| = r$. To see this, consider a choise of $\mathbf{w}$ that gives $r < 0$. Then we have $\mathbf{w}^T\Delta\mathbf{w} < 0$ since $\mathbf{w}^T\mathbf{A}\mathbf{w} < 0$. This means that $\|\mathbf{w}\|$ will decrease until r becomes positive.

The function $\mathbf{A}\hat{\mathbf{w}} - \mathbf{B}\mathbf{w}$ is illustrated in figure 4 together with the Rayleigh quotient plotted to the left in figure 1.

Finding successive eigenvalues Since the learning rule defined in eq. 41 maximizes the Rayleigh quotient in eq. 2, it will find the largest eigenvalue λ_1 and a corresponding eigenvector $\hat{\mathbf{w}}_1 = \pm\hat{\mathbf{e}}_1$ of eq. 1. The question naturally arises if, and how, the algorithm can be modified to find the successive eigenvalues and vectors, i.e. the successive solutions to the eigenvalue equation 1.

Let $\mathbf{G}$ denote the $n \times n$ matrix $\mathbf{B}^{-1}\mathbf{A}$. Then the n equations for the n eigenvalues solving the eigenproblem in eq. 1 can be written as

$$\mathbf{GE} = \mathbf{ED} \quad \Rightarrow \quad \mathbf{G} = \mathbf{EDE}^{-1} = \sum \lambda_i \hat{\mathbf{e}}_i \mathbf{f}_i^T, \tag{43}$$

where the eigenvalues and vectors constitute the matrices $\mathbf{D}$ and $\mathbf{E}$ respectively:

$$\mathbf{D} = \begin{pmatrix} \lambda_1 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \lambda_n \end{pmatrix}, \quad \mathbf{E} = \begin{pmatrix} | & & | \\ \hat{\mathbf{e}}_1 & \cdots & \hat{\mathbf{e}}_n \\ | & & | \end{pmatrix}, \quad \mathbf{E}^{-1} = \begin{pmatrix} - & \mathbf{f}_1^T & - \\ & \vdots & \\ - & \mathbf{f}_n^T & - \end{pmatrix}. \tag{44}$$

The vectors, $\mathbf{f}_i$, appearing in the rows of the inverse of the matrix containing the eigenvectors are the *dual vectors* to the eigenvectors $\hat{\mathbf{e}}_i$, which means that

$$\mathbf{f}_i^T \hat{\mathbf{e}}_j = \delta_{ij}. \tag{45}$$

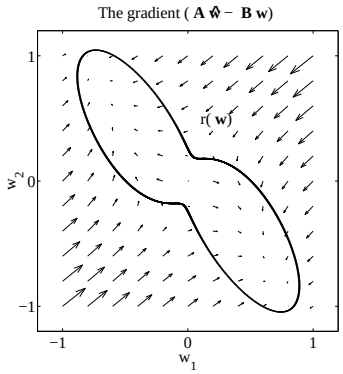

Figure 4: *The function* $\mathbf{A}\hat{\mathbf{w}} - \mathbf{B}\mathbf{w}$*, for the same matrices* $\mathbf{A}$ *and* $\mathbf{B}$ *as in figure 1, plotted for different* $\mathbf{w}$*. The Rayleigh quotient is plotted as reference.*

$\{\mathbf{f}_i\}$ are also called the *left* eigenvectors of $\mathbf{G}$ and $\{\hat{\mathbf{e}}_i\}$, $\{\hat{\mathbf{f}}_i\}$ are said to be *biorthogonal*. From eq. 5 we know that the eigenvectors $\hat{\mathbf{e}}_i$ are both $\mathbf{A}$ and $\mathbf{B}$ orthogonal, i.e. that

$$\hat{\mathbf{e}}_i^T \mathbf{A} \hat{\mathbf{e}}_j = 0 \quad \text{and} \quad \hat{\mathbf{e}}_i^T \mathbf{B} \hat{\mathbf{e}}_j = 0 \quad \text{for} \quad i \neq j. \tag{46}$$

Hence we can use this result to find the dual vectors $\mathbf{f}_i$ possessing the property in eq. 45, e.g. by choosing them according to:

$$\mathbf{f}_i = \frac{\mathbf{B}\hat{\mathbf{e}}_i}{\hat{\mathbf{e}}_i^T \mathbf{B} \hat{\mathbf{e}}_i}. \tag{47}$$

Now, if $\hat{\mathbf{e}}_1$ is the eigenvector corresponding to the largest eigenvalue of $\mathbf{G}$, the new matrix

$$\mathbf{H} = \mathbf{G} - \lambda_1 \hat{\mathbf{e}}_1 \mathbf{f}_1^T \tag{48}$$

will have the same eigenvectors and eigenvalues as $\mathbf{G}$ except for the eigenvalue corresponding to $\hat{\mathbf{e}}_1$, which now becomes 0 (see proof 6.11). This means that the eigenvector corresponding to the largest eigenvalue of $\mathbf{H}$ is the same as the one corresponding to the second largest eigenvalue of $\mathbf{G}$.

Since the algorithm will first find the vector $\hat{\mathbf{w}}_1 = \hat{\mathbf{e}}_1$, we only need to estimate the dual vector $\mathbf{f}_1$ in order to subtract the correct outer product from $\mathbf{G}$ and remove its largest eigenvalue. In our case this is a little bit tricky since we do not generate $\mathbf{G}$ directly. Instead we must modify its two components $\mathbf{A}$ and $\mathbf{B}$ in order to produce the desired subtraction. Hence we want two modified components, $\mathbf{A}'$ and $\mathbf{B}'$, with the following property:

$$\mathbf{B}'^{-1}\mathbf{A}' = \mathbf{B}^{-1}\mathbf{A} - \lambda_1 \hat{\mathbf{e}}_1 \mathbf{f}_1^T. \tag{49}$$

A simple solution is obtained if we only modify one of the matrices and keep the other matrix fixed:

$$\mathbf{B}' = \mathbf{B} \quad \text{and} \quad \mathbf{A}' = \mathbf{A} - \lambda_1 \mathbf{B} \hat{\mathbf{e}}_1 \mathbf{f}_1^T. \tag{50}$$

This modification can be accomplished if we estimate a vector $\mathbf{u}_1 = \lambda_1 \mathbf{B}\hat{\mathbf{e}}_1 = \mathbf{B}\mathbf{w}_1$ iteratively as:

$$\mathbf{u}_1(t+1) = \mathbf{u}_1(t) + \Delta\mathbf{u}_1(t) \tag{51}$$

where

$$E\{\Delta\mathbf{u}_1\} = \alpha \left[r\mathbf{B}\hat{\mathbf{w}}_1 - \mathbf{u}_1 \right]. \tag{52}$$

Once this estimate has converged, we can use $\mathbf{u}_1 = \lambda_1 \mathbf{B}\hat{\mathbf{e}}_1$ to express the outer product in eq. 50:

$$\lambda_1 \mathbf{B}\hat{\mathbf{e}}_1 \mathbf{f}_1^T = \frac{\lambda_1 \mathbf{B}\hat{\mathbf{e}}_1 \hat{\mathbf{e}}_1^T \mathbf{B}^T}{\hat{\mathbf{e}}_1^T \mathbf{B} \hat{\mathbf{e}}_1^T} = \frac{\mathbf{u}_1 \mathbf{u}_1^T}{\hat{\mathbf{e}}_1^T \mathbf{u}_1}. \tag{53}$$

We can now estimate $\mathbf{A}'$ and, hence, get a modified version of the learning algorithm in eq. 41 which finds the second eigenvalue and the corresponding eigenvector to the generalized eigenproblem:

$$E\{\Delta\mathbf{w}\} = \alpha \left[\mathbf{A}^{'}\hat{\mathbf{w}} - r\mathbf{B}\hat{\mathbf{w}}\right] = \alpha \left[\left(\mathbf{A} - \frac{\mathbf{u}_1\mathbf{u}_1^T}{\hat{\mathbf{w}}_1^T\mathbf{u}_1}\right)\hat{\mathbf{w}} - r\mathbf{B}\hat{\mathbf{w}}\right]. \tag{54}$$

The vector $\mathbf{w}_1$ is the solution first produced by the algorithm, i.e. the largest eigenvalue and the corresponding eigenvector.

This scheme can of course be repeated to find the third eigenvalue by subtracting the second solution in the same way and so on. Note that this method does not put any demands on the range of $\mathbf{B}$ in contrast to exact solutions involving matrix inversion.

It is, of course, possible to enhance the proposed update rules and also take second order derivatives into account. This would include estimating the inverse of the Hessian and using this matrix to modify the update direction. Such procedures are, for the batch or off-line case, known as Gauss-Newton methods [7]. In this paper, however, we will not emphasize on speed and convergence rates. Instead we are interested in the structure of the algorithm and how different special cases of the generalized eigenproblem is reflected in the structure of the update rule.

In the following four sub-sections it will be shown how this iterative algorithm can be applied to the four important problems described in the previous section.

3.1 PCA

Finding the largest principal component We can find the direction of maximum data variation by a stochastic gradient search according to eq. 42 with $\mathbf{A}$ and $\mathbf{B}$ defined according to eq. 11:

$$E\{\Delta\mathbf{w}\} = \gamma\frac{\partial\rho}{\partial\mathbf{w}} = \alpha\left[\mathbf{C}_{xx}\hat{\mathbf{w}} - \rho\hat{\mathbf{w}}\right] = \alpha\, E\{\mathbf{x}\mathbf{x}^T\hat{\mathbf{w}} - \rho\hat{\mathbf{w}}\} \tag{55}$$

This leads to a novel unsupervised Hebbian learning algorithm that finds both the direction of maximum data variation and the variance of the data in that direction. The update rule for this algorithm is given by

$$\Delta\mathbf{w} = \alpha\,(\mathbf{x}\mathbf{x}^T\hat{\mathbf{w}} - \mathbf{w}), \tag{56}$$

where the length of the vector represents the estimated variance, i.e. $\|\mathbf{w}\| = \rho$. (Note that ρ in this case is allways positive.)

Note that this algorithm finds both the direction of maximal data variation as well as how much the data varies along that direction. Often algorithms for PCA only finds the direction of maximal data variation. If one is also interested in the variation along this direction, another algorithm need to be employed. This is the case for the well known PCA algorithm presented by Oja [20].

Finding successive principal components In order to find successive principal components, we we recall that $\mathbf{A} = \mathbf{C}_{xx}$ and $\mathbf{B} = \mathbf{I}$. Hence we have the matrix $\mathbf{G} = \mathbf{B}^{-1}\mathbf{A} = \mathbf{C}_{xx}$ which is symmetric and has orthogonal eigenvectors. This means that the dual vectors and the eigenvectors become indistinguishable and that we need not estimate any other vector than $\mathbf{w}$ itself. The outer product in eq. 50 then becomes:

$$\lambda_1\mathbf{B}\hat{\mathbf{e}}_1\mathbf{f}_1^T = \lambda_1\mathbf{I}\hat{\mathbf{e}}_1\hat{\mathbf{e}}_1^T = \mathbf{w}_1\hat{\mathbf{w}}_1^T. \tag{57}$$

From this we see that the modified learning rule for finding the second eigenvalue can be written as

$$E\{\Delta\mathbf{w}\} = \alpha\left[\mathbf{A}^{'}\hat{\mathbf{w}} - \mathbf{B}\mathbf{w}\right] = \alpha\left[(\mathbf{C}_{xx} - \mathbf{w}_1\hat{\mathbf{w}}_1^T)\hat{\mathbf{w}} - \mathbf{w}\right], \tag{58}$$

A stochastic approximation of this rule is achieved if we at each time step update the vector $\mathbf{w}$ by

$$\Delta\mathbf{w} = \alpha\left[(\mathbf{x}\mathbf{x}^T - \mathbf{w}_1\hat{\mathbf{w}}_1^T)\hat{\mathbf{w}} - \mathbf{w}\right]. \tag{59}$$

As mentioned in section 2.1, it is possible to perform a PCA on the inverse of the covariance matrix by choosing $\mathbf{A} = \mathbf{I}$ and $\mathbf{B} = \mathbf{C}_{xx}$. The learning rule associated with this behavior then becomes:

$$\Delta \mathbf{w} = \alpha \ (\hat{\mathbf{w}} - \mathbf{x}\mathbf{x}^T\mathbf{w}). \tag{60}$$

3.2 PLS

Finding the largest singular value If we want to find the directions of maximum data covariance, we define the matrices $\mathbf{A}$ and $\mathbf{B}$ according to eq. 18. Since we want to update $\mathbf{w}$, on average, in direction of the gradient, the update rule in eq. 42 gives:

$$E\{\Delta \mathbf{w}\} = \gamma \frac{\partial r}{\partial \mathbf{w}} = \alpha \left[\begin{pmatrix} \mathbf{0} & \mathbf{C}_{xy} \\ \mathbf{C}_{yx} & \mathbf{0} \end{pmatrix} \hat{\mathbf{w}} - r \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \hat{\mathbf{w}} \right]. \tag{61}$$

This behavior is accomplished if we at each time step update the vector $\mathbf{w}$ with

$$\Delta \mathbf{w} = \alpha \left[\begin{pmatrix} \mathbf{0} & \mathbf{x}\mathbf{y}^T \\ \mathbf{y}\mathbf{x}^T & \mathbf{0} \end{pmatrix} \hat{\mathbf{w}} - \mathbf{w} \right] \tag{62}$$

where the length of the vector at convergence represents the covariance, i.e. $\|\mathbf{w}\| = r = \rho$. This can be done since we know that it is sufficient to search for positive values of ρ.

Finding successive singular values Also in this case, the special structure of the $\mathbf{A}$ and $\mathbf{B}$ matrices will simplify the procedure for finding the subsequent directions with maximum data covariance. We have

$$\mathbf{A} = \begin{pmatrix} \mathbf{0} & \mathbf{C}_{xy} \\ \mathbf{C}_{yx} & \mathbf{0} \end{pmatrix} \quad \text{and} \quad \mathbf{B} = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}, \tag{63}$$

which again means that the compound matrix $\mathbf{G} = \mathbf{B}^{-1}\mathbf{A} = \mathbf{A}$ will be symmetric and have orthogonal eigenvectors, which are identical to their dual vectors. The outer product for modification of the matrix $\mathbf{A}$ in eq. 50 becomes identical to the one presented in the previous section:

$$\lambda_1 \mathbf{B}\hat{\mathbf{e}}_1\mathbf{f}_1^T = \lambda_1 \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \hat{\mathbf{e}}_1\hat{\mathbf{e}}_1^T = \mathbf{w}_1\hat{\mathbf{w}}_1^T. \tag{64}$$

A modified learning rule for finding the second eigenvalue can thus be written as

$$E\{\Delta \mathbf{w}\} = \alpha \ \left[\mathbf{A}'\hat{\mathbf{w}} - \mathbf{B}\mathbf{w}\right] = \ \alpha \ \left[\left(\begin{pmatrix} \mathbf{0} & \mathbf{C}_{xy} \\ \mathbf{C}_{yx} & \mathbf{0} \end{pmatrix} - \mathbf{w}_1\hat{\mathbf{w}}_1^T \right) \hat{\mathbf{w}} - \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \mathbf{w} \right]. \tag{65}$$

A stochastic approximation of this rule is achieved if we at each time step update the vector $\mathbf{w}$ by

$$\Delta \mathbf{w} = \alpha \ \left[\left(\begin{pmatrix} \mathbf{0} & \mathbf{x}\mathbf{y}^T \\ \mathbf{y}\mathbf{x}^T & \mathbf{0} \end{pmatrix} - \mathbf{w}_1\hat{\mathbf{w}}_1^T \right) \hat{\mathbf{w}} - \mathbf{w} \right]. \tag{66}$$

3.3 CCA

Finding the largest canonical correlation Again, the algorithm in eq. 42 for solving the generalized eigenproblem can be used for the stochastic gradient search. With the matrices $\mathbf{A}$ and $\mathbf{B}$ and the vector $\mathbf{w}$ as in eq. 28, we obtain the update direction as:

$$E\{\Delta \mathbf{w}\} = \gamma \frac{\partial r}{\mathbf{w}} = \alpha \left[\begin{pmatrix} \mathbf{0} & \mathbf{C}_{xy} \\ \mathbf{C}_{yx} & \mathbf{0} \end{pmatrix} \hat{\mathbf{w}} - r \begin{pmatrix} \mathbf{C}_{xx} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{yy} \end{pmatrix} \hat{\mathbf{w}} \right]. \tag{67}$$

This behavior is accomplished if we at each time step update the vector $\mathbf{w}$ with

$$\Delta \mathbf{w} = \alpha \left[\begin{pmatrix} \mathbf{0} & \mathbf{x}\mathbf{y}^T \\ \mathbf{y}\mathbf{x}^T & \mathbf{0} \end{pmatrix} \hat{\mathbf{w}} - \begin{pmatrix} \mathbf{x}\mathbf{x}^T & \mathbf{0} \\ \mathbf{0} & \mathbf{y}\mathbf{y}^T \end{pmatrix} \mathbf{w} \right]. \tag{68}$$

Since we will have $\|\mathbf{w}\| = r = \rho$ when the algorithm converges, the length of the vector represents the correlation between the variates.

Finding successive canonical correlations In the two previous cases it was easy to cancel out an eigenvalue because the matrix $\mathbf{G}$ was symmetric. This is not the case for canonical correlation. Here, we have

$$\mathbf{A} = \begin{pmatrix} \mathbf{0} & \mathbf{C}_{xy} \\ \mathbf{C}_{yx} & \mathbf{0} \end{pmatrix} \quad \text{and} \quad \mathbf{B} = \begin{pmatrix} \mathbf{C}_{xx} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{yy} \end{pmatrix}, \tag{69}$$

which gives us the non-symmetric matrix

$$\mathbf{G} = \mathbf{B}^{-1}\mathbf{A} = \begin{pmatrix} \mathbf{C}_{xx}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{yy}^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{0} & \mathbf{C}_{xy} \\ \mathbf{C}_{yx} & \mathbf{0} \end{pmatrix} = \begin{pmatrix} \mathbf{0} & \mathbf{C}_{xx}^{-1}\mathbf{C}_{xy} \\ \mathbf{C}_{yy}^{-1}\mathbf{C}_{yx} & \mathbf{0} \end{pmatrix}. \tag{70}$$

Because of this, we need to estimate the dual vector $\mathbf{f}_1$ corresponding to the eigenvector $\hat{\mathbf{e}}_1$, or rather the vector $\mathbf{u}_1 = \lambda_1 \mathbf{B}\hat{\mathbf{e}}_1$ as described in eq. 52:

$$E\{\Delta\mathbf{u}_1\} = \alpha \ [\mathbf{B}\mathbf{w}_1 - \mathbf{u}_1] = \alpha \ \left[\begin{pmatrix} \mathbf{C}_{xx} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{yy} \end{pmatrix} \mathbf{w}_1 - \mathbf{u}_1 \right]. \tag{71}$$

A stochastic approximation for this rule is given by

$$\Delta\mathbf{u}_1 = \alpha \ \left[\begin{pmatrix} \mathbf{x}\mathbf{x}^T & \mathbf{0} \\ \mathbf{0} & \mathbf{y}\mathbf{y}^T \end{pmatrix} \mathbf{w}_1 - \mathbf{u}_1 \right]. \tag{72}$$

With this estimate, the outer product in eq. 50 can be used to modify the matrix $\mathbf{A}$:

$$\mathbf{A}' = \mathbf{A} - \lambda_1 \mathbf{B}\hat{\mathbf{e}}_1 \mathbf{f}_1^T = \mathbf{A} - \frac{\mathbf{u}_1\mathbf{u}_1^T}{\hat{\mathbf{w}}_1^T\mathbf{u}_1}. \tag{73}$$

A modified version of the learning algorithm in eq. 42 which finds the second largest canonical correlations and its corresponding directions can be written on the following form:

$$E\{\Delta\mathbf{w}\} = \alpha \ \left[\mathbf{A}'\hat{\mathbf{w}} - \mathbf{B}\mathbf{w}\right] = \alpha \ \left[\left(\begin{pmatrix} \mathbf{0} & \mathbf{C}_{xy} \\ \mathbf{C}_{yx} & \mathbf{0} \end{pmatrix} - \frac{\mathbf{u}_1\mathbf{u}_1^T}{\hat{\mathbf{w}}_1^T\mathbf{u}_1} \right) \hat{\mathbf{w}} - \begin{pmatrix} \mathbf{C}_{xx} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{yy} \end{pmatrix} \mathbf{w} \right]. \tag{74}$$

Again to get a stochastic approximation of this rule, we perform the update at each time step according to:

$$\Delta\mathbf{w} = \alpha \ \left[\left(\begin{pmatrix} \mathbf{0} & \mathbf{x}\mathbf{y}^T \\ \mathbf{y}\mathbf{x}^T & \mathbf{0} \end{pmatrix} - \frac{\mathbf{u}_1\mathbf{u}_1^T}{\hat{\mathbf{w}}_1^T\mathbf{u}_1} \right) \hat{\mathbf{w}} - \begin{pmatrix} \mathbf{x}\mathbf{x}^T & \mathbf{0} \\ \mathbf{0} & \mathbf{y}\mathbf{y}^T \end{pmatrix} \mathbf{w} \right]. \tag{75}$$

Note that this algorithm simultaneously finds both the directions of canonical correlations *and* the canonical correlations ρ_i in contrast to the algorithm proposed by Kay [15], which only finds the directions.

3.4 MLR

Finding the directions for minimum square error Also here, the algorithm in eq. 42 can be used for a stochastic gradient search. With the $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{w}$ according to eq. 38, we get the update direction as:

$$E\{\Delta\mathbf{w}\} = \gamma\frac{\partial r}{\partial \mathbf{w}} = \alpha \left[\begin{pmatrix} \mathbf{0} & \mathbf{C}_{xy} \\ \mathbf{C}_{yx} & \mathbf{0} \end{pmatrix} \hat{\mathbf{w}} - r \begin{pmatrix} \mathbf{C}_{xx} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \hat{\mathbf{w}} \right]. \tag{76}$$

This behavior is accomplished if we at each time step update the vector $\mathbf{w}$ with

$$\Delta\mathbf{w} = \alpha \left[\begin{pmatrix} \mathbf{0} & \mathbf{x}\mathbf{y}^T \\ \mathbf{y}\mathbf{x}^T & \mathbf{0} \end{pmatrix} \hat{\mathbf{w}} - \begin{pmatrix} \mathbf{x}\mathbf{x}^T & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \mathbf{w} \right]. \tag{77}$$

Since we have $\|\mathbf{w}\| = r = \rho$ when the algorithm converges, we get the regression coefficient as $\beta = \|\mathbf{w}\|\frac{\mu_x}{\mu_y}$

Finding successive directions for minimum square error Also in this case we must use the dual vectors to cancel out the detected eigenvalues. Here, we have

$$\mathbf{A} = \begin{pmatrix} \mathbf{0} & \mathbf{C}_{xy} \\ \mathbf{C}_{yx} & \mathbf{0} \end{pmatrix} \quad \text{and} \quad \mathbf{B} = \begin{pmatrix} \mathbf{C}_{xx} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}, \tag{78}$$

which gives us the non-symmetric matrix $\mathbf{G}$ as

$$\mathbf{G} = \mathbf{B}^{-1}\mathbf{A} = \begin{pmatrix} \mathbf{C}_{xx}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{0} & \mathbf{C}_{xy} \\ \mathbf{C}_{yx} & \mathbf{0} \end{pmatrix} = \begin{pmatrix} \mathbf{0} & \mathbf{C}_{xx}^{-1}\mathbf{C}_{xy} \\ \mathbf{C}_{yx} & \mathbf{0} \end{pmatrix}. \tag{79}$$

Because of this, we need to estimate the dual vector $\mathbf{f}_1$ corresponding to the eigenvector $\hat{\mathbf{e}}_1$, or rather the vector $\mathbf{u}_1 = \lambda_1 \mathbf{B}\hat{\mathbf{e}}_1$ as described in eq. 52:

$$E\{\Delta\mathbf{u}_1\} = \alpha \ [\mathbf{B}\mathbf{w}_1 - \mathbf{u}_1] = \alpha \ \left[\begin{pmatrix} \mathbf{C}_{xx} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \mathbf{w}_1 - \mathbf{u}_1 \right]. \tag{80}$$

A stochastic approximation for this rule is given by

$$\Delta\mathbf{u}_1 = \alpha \ \left[\begin{pmatrix} \mathbf{x}\mathbf{x}^T & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \mathbf{w}_1 - \mathbf{u}_1 \right]. \tag{81}$$

With this estimate, the outer product in eq. 50 can be used to modify the matrix $\mathbf{A}$:

$$\mathbf{A}' = \mathbf{A} - \lambda_1 \mathbf{B}\hat{\mathbf{e}}_1 \mathbf{f}_1^T = \mathbf{A} - \frac{\mathbf{u}_1 \mathbf{u}_1^T}{\hat{\mathbf{w}}_1^T \mathbf{u}_1}. \tag{82}$$

A modified version of the learning algorithm in eq. 42 which finds the successive directions of minimum square error and their corresponding regression coefficient can be written on the following form:

$$E\{\Delta\mathbf{w}\} = \alpha \ \left[\mathbf{A}'\hat{\mathbf{w}} - \mathbf{B}\mathbf{w}\right] = \alpha \ \left[\left(\begin{pmatrix} \mathbf{0} & \mathbf{C}_{xy} \\ \mathbf{C}_{yx} & \mathbf{0} \end{pmatrix} - \frac{\mathbf{u}_1 \mathbf{u}_1^T}{\hat{\mathbf{w}}_1^T \mathbf{u}_1} \right) \hat{\mathbf{w}} - \begin{pmatrix} \mathbf{C}_{xx} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \mathbf{w} \right]. \tag{83}$$

Again to get a stochastic approximation of this rule, we perform the update at each time step according to:

$$\Delta\mathbf{w} = \alpha \ \left[\left(\begin{pmatrix} \mathbf{0} & \mathbf{x}\mathbf{y}^T \\ \mathbf{y}\mathbf{x}^T & \mathbf{0} \end{pmatrix} - \frac{\mathbf{u}_1 \mathbf{u}_1^T}{\hat{\mathbf{w}}_1^T \mathbf{u}_1} \right) \hat{\mathbf{w}} - \begin{pmatrix} \mathbf{x}\mathbf{x}^T & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \mathbf{w} \right]. \tag{84}$$

We can see that, in this case, the $\mathbf{w}_y$s are orthogonal but not necessarily the $\mathbf{w}_x$s. The orthogonality of the $\mathbf{w}_y$s is easily explained by the Cartesian separability of the square error; when the error in one direction is minimized, no more can be done in that direction to reduce the error. This shows that we can use this method for successively building up a low-rank approximation of MLR by adding a sufficient number of solutions, i.e.

$$\tilde{\mathbf{y}} = \sum_{i=1}^{k} \beta_i \hat{\mathbf{w}}_{yi} \hat{\mathbf{w}}_{xi}^T \mathbf{x} \tag{85}$$

where $\tilde{\mathbf{y}}$ is the estimated $\mathbf{y}$ and k is the rank. It may be pointed out that if all solutions are used, we obtain the well known Wiener filter.

4 Experiments

The memory requirement as well as the computational cost per iteration for the presented algorithm is of order $\mathcal{O}(md)$. This enables experiments in signal spaces having dimensionalities which would be impossible to handle using traditional techniques involving matrix multiplications (having memory requirements of order $\mathcal{O}(d^2)$ and computational costs of order $\mathcal{O}(d^3)$).

This section presents some experiments with the algorithm for analysis of stochastic processes. First, the algorithm is employed to perform PCA, PLS, CCA, and MLR. Here the dimensionality

of the signal space is kept reasonably low in order to make a comparison with the performance of an *optimal*, in the sense of maximum likelihood (ML), deterministic solution which is calculated for each iteration, based on the data accumulated so far.

Second, the algorithm is applied to a process in a high-dimensional (1000-dim.) signal space. In this case, the gain sequence is made data dependent and the output from the algorithm is post-filtered in order meet requirements for quick convergence together with algorithm robustness.

In all experiments the error in magnitude and angle were calculated relative the correct answer $\mathbf{w}_c$. The same error measures were used for the output from the algorithm as well as for the optimal ML estimate:

$$\epsilon_m(\mathbf{w}) = \|\mathbf{w}_o\| - \|\mathbf{w}\| \tag{86}$$

$$\epsilon_a(\mathbf{w}) = \arccos(\hat{\mathbf{w}}^T \hat{\mathbf{w}}_o). \tag{87}$$

4.1 Comparisons to optimal solutions

The test data for these four experiments was generated from a 30-dimensional Gaussian distribution such that the eigenvalues of the generalized eigenproblem decreased exponentially, from 0.9:

$$\lambda_i = 0.9 \left(\frac{2}{3}\right)^{i-1}.$$

The two largest eigenvalues (0.9 and 0.6) and the corresponding eigenvectors were simultaneously searched for. In the PLS, CCA and MLR experiments, the dimensionalities of signal vector belonging to the $\mathbf{x}$ and $\mathbf{y}$ part of the signal were 20 and 10 respectively.

The average angular and magnitude errors were calculated based on 10 different runs. This computation was made for each iteration, both for the algorithm and for the ML solution. The results are plotted in figures 5, 6, 7 and 8 for PCA, PLS, CCA and MLR respectively. The errors of the algorithm are drawn with solid lines and the errors of the ML solution are drawn with dotted lines. The vertical bars show the standard deviations. Note that the angular error is always positive and, hence, does not have have a symmetrical distribution. However, for simplicity, the standard deviation indicators have been placed symmetrically around the mean. The first 30 iterations were omitted to avoid singular matrices when calculating matrix inverses for the ML solutions.

No attempt was made to find an optimal set of parameters for the algorithm. Instead the experiments and comparisons were carried out only to display the behavior of the algorithm and show that it is robust and converges to the correct solutions. Initially, the estimate was assigned a small random vector. A constant gain factor of $\alpha = 0.001$ was used throughout all four experiments.

4.2 Performance in high dimensional signal spaces

The purpose of the methods presented in this paper is dimensionality reduction in high-dimensional signal spaces. We have previously shown that the proposed algorithm have the computational capacity to handle such signals. This experiment illustrates that the algorithm also behaves well in practice for high-dimensional signals. The dimensionality of $\mathbf{x}$ is 800 and the dimensionality of $\mathbf{y}$ is 200, so the total dimensionality of the signal space is 1000. The object in this experiment is CCA.

In the previous experiment, the algorithm was used in its basic form with constant update rates set by hand. In this experiment, however, a more sophisticated version of the algorithm is used where the update rate is adaptive and the vectors are averaged over time. The details of this extension to the algorithm are numerous and beyond the scope of this paper. Here, we will only give a brief explanation of the basic structure of the extended algorithm.

Adaptability is necessary for a system without a pre-specified (time dependent) update rate α. Here, the adaptive update rate is dependent on the energy of the signal projected onto the vector as well as the consistency of the change of the vector.

The averaged vectors $\mathbf{w_a}$ are calculated as

$$\mathbf{w_a} \Leftarrow \mathbf{w_a} + \gamma \ (\mathbf{w} - \mathbf{w_a}) \tag{88}$$

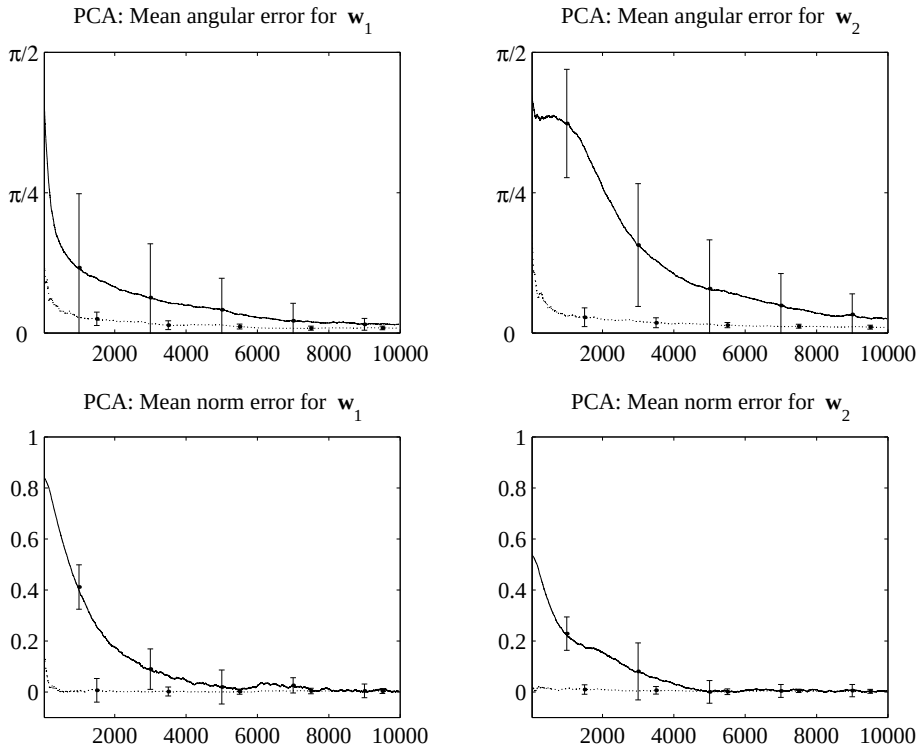

Figure 5: *Results for the PCA case*

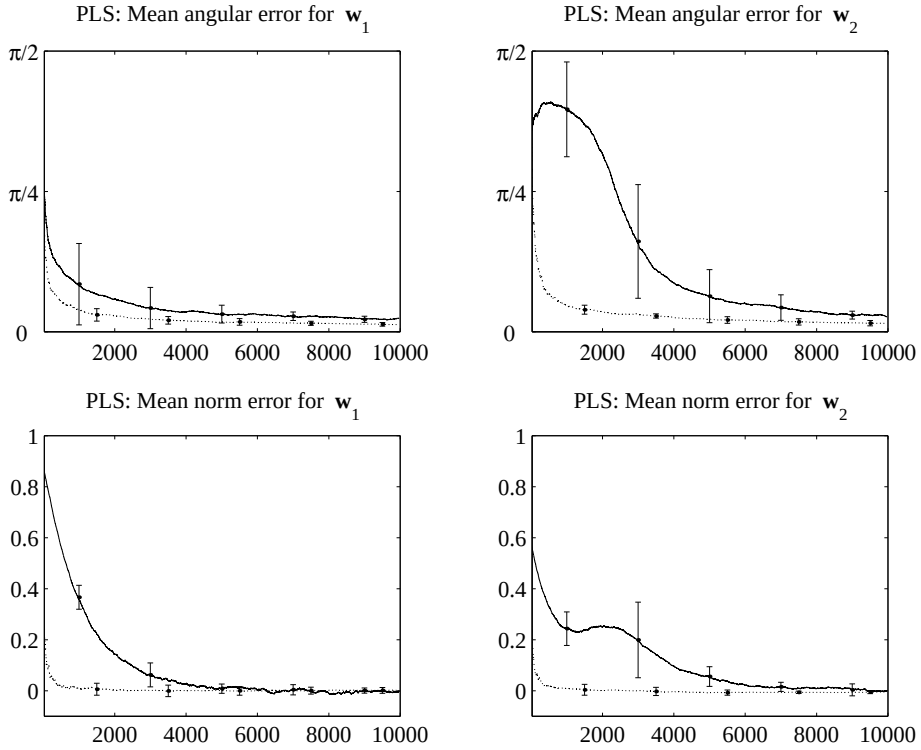

Figure 6: *Results for the PLS case*

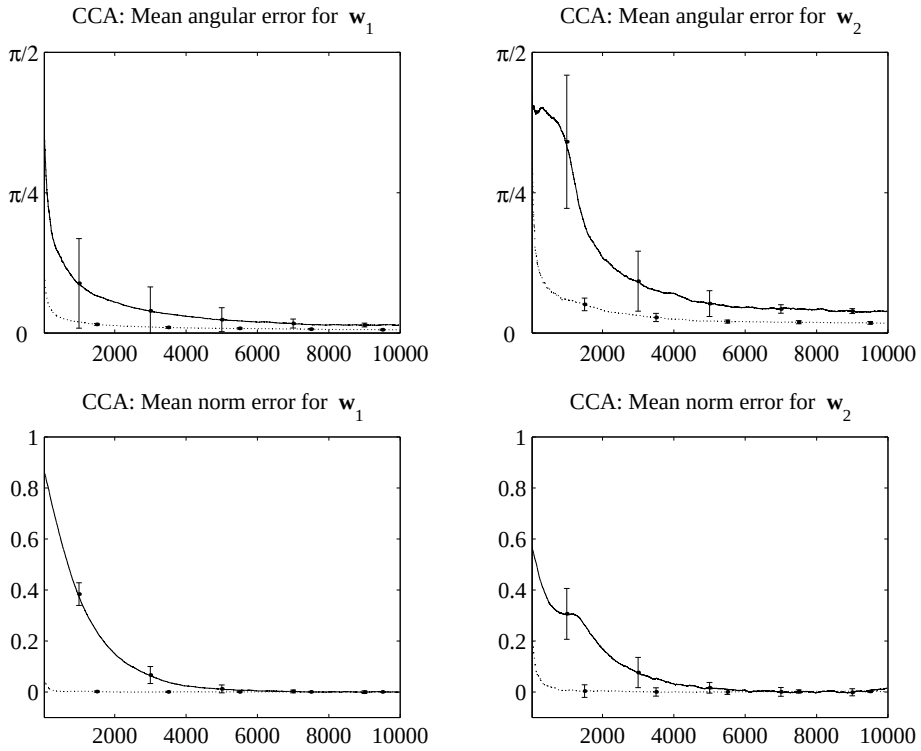

Figure 7: *Results for the CCA case*

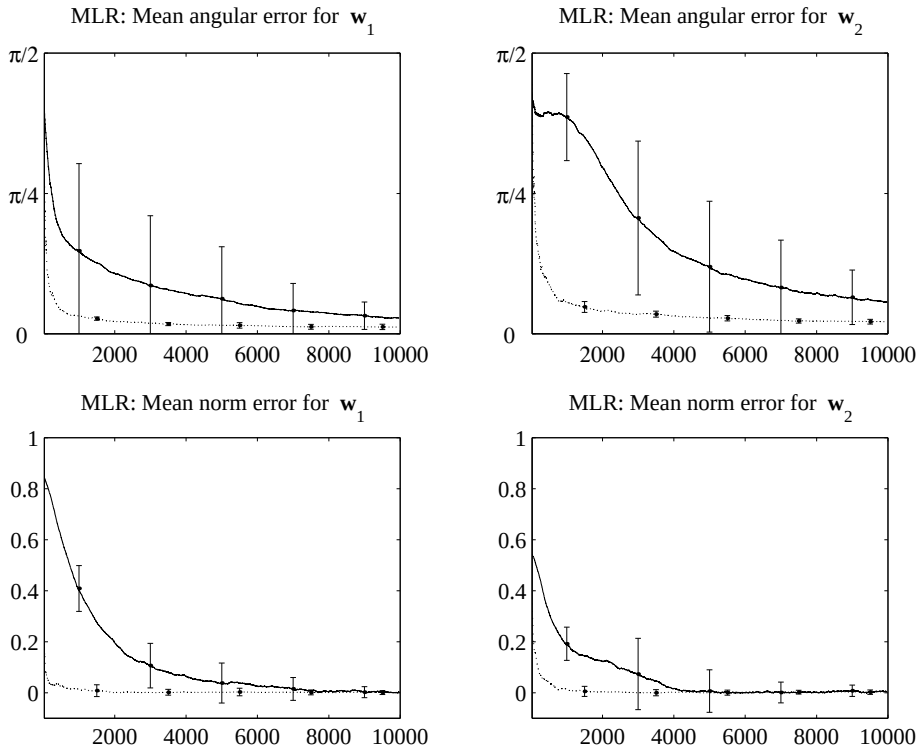

Figure 8: *Results for the MLR case*

where γ depends on the consistency of the changes in $\mathbf{w}$. When there is a consistent change in $\mathbf{w}$, γ is small and the averaging window is short and $\mathbf{w_a}$ follows $\mathbf{w}$ quickly. When the changes in $\mathbf{w}$ are less consistent, the window gets longer and $\mathbf{w_a}$ is the average of an increasing number of instances of $\mathbf{w}$. This means, for example, that if $\mathbf{w}$ is moving symmetrically around the correct solution with a constant variance, the error of $\mathbf{w_a}$ will still tend towards zero (see figure 9).

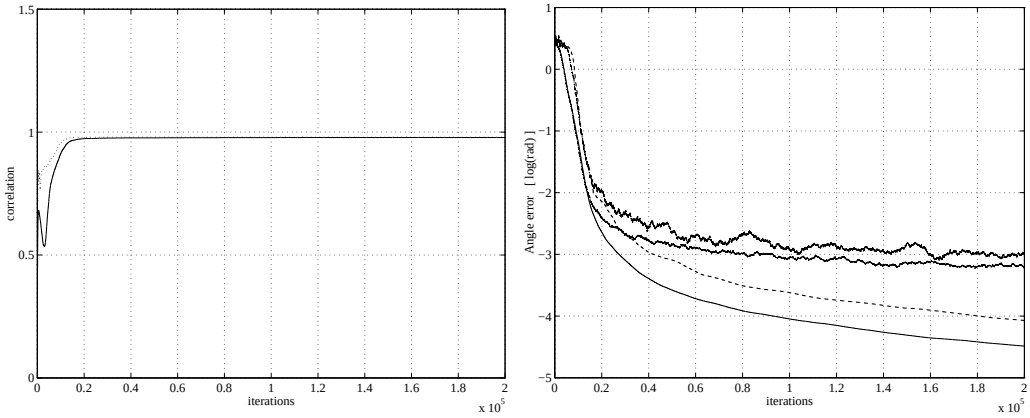

Figure 9: **Left:** *Figure showing the estimated first canonical correlation as a function of number of actual events (solid line) and the true correlation in the current directions found by the algorithm (dotted line). The dimensionality of one set of variables is* 800 *and of the second set* 200*. Right:Figure showing the log of the angular error as a function of number of actual events.*

The experiment was carried out using a randomly chosen distribution of a 800-dimensional $\mathbf{x}$ variable and a 200-dimensional $\mathbf{y}$ variable. Two $\mathbf{x}$ and two $\mathbf{y}$ dimensions were correlated. The other 798 dimensions of $\mathbf{x}$ and 198 dimensions of $\mathbf{y}$ were uncorrelated. The variances in the 1000 dimensions were of the same order of magnitude.

The left plot in figure 9 shows the estimated first canonical correlation as a function of number of actual events (solid line) and the true correlation in the current directions found by the algorithm (dotted line).

The right plot in figure 9 shows the effect of the adaptive averaging. The two upper noisy curves show the angular errors of the 'raw' estimates in the $\mathbf{x}$ and $\mathbf{y}$ spaces and the two lower curves shows the angular errors for $\mathbf{x}$ (dashed) and $\mathbf{y}$ (solid). The angle errors of the smoothed estimates are much more stable and decrease more rapidly than the 'raw' estimates. The errors after 2×10^5 samples is below one degree. (It should be noted that this is an extreme precision as, with a resolution of 1 degree, a low estimate of the number of different orientations in a 1000-dimensional space is 10^{2000}.) The angular errors 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.

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

5 Summary and conclusions

We have presented an iterative algorithm for analysis of stochastic processes in terms of PCA, PLS, CCA, and MLR. The directions of maximal variance, covariance, correlation, and least square error are found by a novel algorithm performing a stochastic gradient search on suitable Rayleigh quotients. The algorithm operates on-line which allows non-stationary data to be analyzed. When searching for an m-rank approximation, the computational complexity is $\mathcal{O}(md)$ for each iteration. Finding a full rank solution have a computational complexity of order $\mathcal{O}(d^3)$ using traditional techniques.

The equilibrium points of the algorithm were shown to correspond to solutions of the generalized eigenproblem. Hence, PCA, PLS, CCA and MLR were presented as special cases of this more general problem. In PCA, PLS and CCA, the eigenvalues corresponds to variance, covariance and correlation respectively of the projection of the data onto the eigenvectors. In MLR, the eigenvalues,

together with a function of the corresponding eigenvector, provide the regression coefficients. The eigenvalues are given by the lengths of the basis vectors found by the proposed algorithm. A low rank approximation is obtained when only the solutions with the largest eigenvalues and their corresponding vectors are used.

Reduced rank MLR can, for example, be used to increase the stability of the predictors when there are more parameters than observations, when the relation is known to be of low rank or, maybe most importantly, when a full rank solution is unobtainable due to computational costs. The regression coefficients can of course also be used for regression in the first three cases. In the case of PCA, the idea is to separately reduce the dimensionality of the X and Y spaces and do a regression of the first principal components of Y on the first principal components of X. This method is known as principal components regression. The obvious disadvantage here is that there is no reason that the principal components of X are related to the principal components of Y. To avoid this problem, PLS regression is some times used. Clearly, this choice of basis is better than PCA for regression purposes since directions of high covariance are selected, which means that a linear relation is easier to find. However, neither of these solutions results in minimum least squares error. This is only obtained using the directions corresponding to the MLR problem.

PCA differs from the other three methods in that it concerns only one set of variables while the other three concerns relations between two sets of variables. The difference between PLS, CCA and MLR can be seen by comparing the matrices in the corresponding eigenproblems. In CCA, the between sets covariance matrices are normalized with respect to the within set covariances in both the $\mathbf{x}$ and the $\mathbf{y}$ spaces. In MLR, the normalization is done only with respect to the $\mathbf{x}$ space covariance while the $\mathbf{y}$ space, where the square error is defined, is left unchanged. In PLS, no normalization is done. Hence, these three cases can be seen as the same problem, covariance maximization, where the variables have been subjected to different, data dependent, scaling.

In some PLS applications, the variances of the variables are scaled to unity [25, 8, 13]. This may indicate that the aim is really to maximize *correlation* and that CCA would be the proper method to use.

Recently, the neural network community has taken an increased interest in information theoretcal approaches[11]. In particular, the concepts *independant components* and *mutual information* has been the basis for a number of successful applications, e.g. blind separation and blind deconvolution [2]. It is appropriate to point out that there is a strong relation between these concepts and canonical correlation [1, 15]. The relevance of the present paper in this context is apparent.

6 Proofs

6.1 Orthogonality in the metrics A and B (eq. 5)

$$\hat{\mathbf{w}}_i^T \mathbf{B} \hat{\mathbf{w}}_j = \begin{cases} 0 & \text{for} \quad i \neq j \\ \beta_i > 0 & \text{for} \quad i = j \end{cases} \quad \text{and} \quad \hat{\mathbf{w}}_i^T \mathbf{A} \hat{\mathbf{w}}_j = \begin{cases} 0 & \text{for} \quad i \neq j \\ r_i \beta_i & \text{for} \quad i = j \end{cases} \tag{5}$$

Proof: For solution i we have

$$\mathbf{A} \hat{\mathbf{w}}_i = r_i \mathbf{B} \hat{\mathbf{w}}_i$$

The scalar product with another eigenvector gives

$$\hat{\mathbf{w}}_j^T \mathbf{A} \hat{\mathbf{w}}_i = r_i \hat{\mathbf{w}}_j^T \mathbf{B} \hat{\mathbf{w}}_i$$

and of course also

$$\hat{\mathbf{w}}_i^T \mathbf{A} \hat{\mathbf{w}}_j = r_j \hat{\mathbf{w}}_i^T \mathbf{B} \hat{\mathbf{w}}_j$$

Since $\mathbf{A}$ and $\mathbf{B}$ are Hermitian we can change positions of $\hat{\mathbf{w}}_i$ and $\hat{\mathbf{w}}_j$ which gives

$$r_j \hat{\mathbf{w}}_i^T \mathbf{B} \hat{\mathbf{w}}_j = r_i \hat{\mathbf{w}}_i^T \mathbf{B} \hat{\mathbf{w}}_j$$

and hence

$$(r_i - r_j) \hat{\mathbf{w}}_i^T \mathbf{B} \hat{\mathbf{w}}_j = 0.$$

For this expression to be true when $i \neq j$, we have that $\hat{\mathbf{w}}_i^T \mathbf{B} \hat{\mathbf{w}}_j = 0$ if $r_i \neq r_j$. For $i = j$ we now have that $\hat{\mathbf{w}}_i^T \mathbf{B} \hat{\mathbf{w}}_i = \beta_i > 0$ since $\mathbf{B}$ is positive definite. In the same way we have

$$\left(\frac{1}{r_i} - \frac{1}{r_j} \right) \hat{\mathbf{w}}_i^T \mathbf{A} \hat{\mathbf{w}}_j = 0$$

which means that $\hat{\mathbf{w}}_i^T \mathbf{A} \hat{\mathbf{w}}_j = 0$ for $i \neq j$. For $i = j$ we know that $\hat{\mathbf{w}}_i^T \mathbf{A} \hat{\mathbf{w}}_i = r_i \hat{\mathbf{w}}_i^T \mathbf{B} \hat{\mathbf{w}}_i = r_i \beta_i$.

□

6.2 Linear independence

$\{\mathbf{w}_i\}$ are linearly independent.

Proof: Suppose $\{\mathbf{w}_i\}$ are *not* linearly independent. This would mean that we could write an eigenvector $\mathbf{w}_k$ as

$$\hat{\mathbf{w}}_k = \sum_{j \neq k} \gamma_j \hat{\mathbf{w}}_j.$$

This means that for $j \neq k$,

$$\mathbf{w}_j^T \mathbf{B} \mathbf{w}_k = \gamma_j \mathbf{w}_j \mathbf{B} \mathbf{w}_j \neq 0$$

which violates equation 5. Hence, $\{\mathbf{w}_i\}$ are linear independent.

□

6.3 The range of r (eq. 6)

$$r_n \leq r \leq r_1 \tag{6}$$

Proof: If we express a vector $\mathbf{w}$ in the base of the eigenvectors $\hat{\mathbf{w}}_i$, i.e.

$$\mathbf{w} = \sum_i \gamma_i \hat{\mathbf{w}}_i$$

we can write

$$r = \frac{\sum \gamma_i \hat{\mathbf{w}}_i^T \mathbf{A} \sum \gamma_i \hat{\mathbf{w}}_i}{\sum \gamma_i \hat{\mathbf{w}}_i^T \mathbf{B} \sum \gamma_i \hat{\mathbf{w}}_i} = \frac{\sum \gamma_i^2 \alpha_i}{\sum \gamma_i^2 \beta_i},$$

where $\alpha_i = \hat{\mathbf{w}}_i^T \mathbf{A} \hat{\mathbf{w}}_i$. Now, since $\alpha_i = \beta_i r_i$ (see equation 5), we get

$$r = \frac{\sum \gamma_i^2 \beta_i r_i}{\sum \gamma_i^2 \beta_i}.$$

Obviously this function has the maximum value r_1 when $\gamma_1 \neq 0$ and $\gamma_i = 0 \; \forall \; i > 1$ if r_1 is the largest eigenvalue. The minimum value, r_n, is obtained when $\gamma_n \neq 0$ and $\gamma_i = 0 \; \forall \; i < n$ if r_n is the smallest eigenvalue.

□

6.4 The second derivative of r (eq. 7)

$$\mathbf{H}_i = \left. \frac{\partial^2 r}{\partial \mathbf{w}^2} \right|_{\mathbf{w} = \hat{\mathbf{w}}_i} = \frac{2}{\hat{\mathbf{w}}_i^T \mathbf{B} \hat{\mathbf{w}}_i} (\mathbf{A} - r_i \mathbf{B}) \tag{7}$$

Proof: From the gradient in equation 3 we get the second derivative as

$$\frac{\partial^2 r}{\partial \mathbf{w}^2} = \frac{2}{(\mathbf{w}^T \mathbf{B} \mathbf{w})^2} \left[\left(\mathbf{A} - \frac{\partial r}{\partial \mathbf{w}} \mathbf{w}^T \mathbf{B} - r \mathbf{B} \right) \mathbf{w}^T \mathbf{B} \mathbf{w} - (\mathbf{A} \mathbf{w} - r \mathbf{B} \mathbf{w}) 2 \mathbf{w}^T \mathbf{B} \right].$$

If we insert one of the solutions $\hat{\mathbf{w}}_i$, we have

$$\left.\frac{\partial r}{\partial \mathbf{w}}\right|_{\mathbf{w}=\hat{\mathbf{w}}_i} = \frac{2}{\hat{\mathbf{w}}_i^T \mathbf{B} \hat{\mathbf{w}}_i} \left(\mathbf{A}\hat{\mathbf{w}}_i - r\mathbf{B}\hat{\mathbf{w}}_i\right) = \mathbf{0}$$

and hence

$$\left.\frac{\partial^2 r}{\partial \mathbf{w}^2}\right|_{\mathbf{w}=\hat{\mathbf{w}}_i} = \frac{2}{\hat{\mathbf{w}}_i^T \mathbf{B} \hat{\mathbf{w}}_i} \left(\mathbf{A} - r_i\mathbf{B}\right).$$

□

6.5 Positive eigenvalues of the Hessian (eq. 8)

$$\mathbf{w}^T \mathbf{H}_i \mathbf{w} > 0 \quad \forall\, i > 1 \tag{8}$$

Proof: If we express a vector $\mathbf{w}$ as a linear combination of the eigenvectors we get

$$\begin{aligned}
\frac{\beta_i}{2}\mathbf{w}^T\mathbf{H}_i\mathbf{w} &= \mathbf{w}^T(\mathbf{A} - r_i\mathbf{B})\mathbf{w} \\
&= \mathbf{w}^T\mathbf{B}(\mathbf{B}^{-1}\mathbf{A} - r_i\mathbf{I})\mathbf{w} \\
&= \sum \gamma_j \hat{\mathbf{w}}_j^T \mathbf{B}(\mathbf{B}^{-1}\mathbf{A} - r_i\mathbf{I}) \sum \gamma_j \hat{\mathbf{w}}_j \\
&= \sum \gamma_j \hat{\mathbf{w}}_j^T \mathbf{B} \left(\sum r_j\gamma_j\hat{\mathbf{w}}_j - \sum r_i\gamma_j\hat{\mathbf{w}}_j\right) \\
&= \sum \gamma_j \hat{\mathbf{w}}_j^T \mathbf{B} \sum (r_j - r_i)\gamma_j\hat{\mathbf{w}}_j \\
&= \sum \gamma_j^2 \beta_j (r_j - r_i)
\end{aligned}$$

where $\beta_i = \hat{\mathbf{w}}_i^T \mathbf{B} \hat{\mathbf{w}}_i > 0$. Now, $(r_j - r_i) > 0$ for $j < i$ so if $i > 1$ there is at least one choice of $\mathbf{w}$ that makes this sum positive.

□

6.6 The partial derivatives of the covariance (eq. 16)

$$\begin{cases} \frac{\partial \rho}{\partial \mathbf{w}_x} &= \frac{1}{\|\mathbf{w}_x\|}(\mathbf{C}_{xy}\hat{\mathbf{w}}_y - \rho\hat{\mathbf{w}}_x) \\ \frac{\partial \rho}{\partial \mathbf{w}_y} &= \frac{1}{\|\mathbf{w}_y\|}(\mathbf{C}_{yx}\hat{\mathbf{w}}_x - \rho\hat{\mathbf{w}}_y). \end{cases} \tag{16}$$

Proof: The partial derivative of ρ with respect to $\mathbf{w}_x$ is

$$\begin{aligned}
\frac{\partial \rho}{\partial \mathbf{w}_x} &= \frac{\mathbf{C}_{xy}\mathbf{w}_y\|\mathbf{w}_x\|\|\mathbf{w}_y\| - \mathbf{w}_x^T\mathbf{C}_{xy}\mathbf{w}_y\|\mathbf{w}_x\|^{-1}\mathbf{w}_x\|\mathbf{w}_y\|}{\|\mathbf{w}_x\|^2\|\mathbf{w}_y\|^2} \\
&= \frac{\mathbf{C}_{xy}\hat{\mathbf{w}}_y}{\|\mathbf{w}_x\|} - \frac{\rho\mathbf{w}_x}{\|\mathbf{w}_x\|^2} \\
&= \frac{1}{\|\mathbf{w}_x\|}\left(\mathbf{C}_{xy}\hat{\mathbf{w}}_y - \rho\hat{\mathbf{w}}_x\right)
\end{aligned}$$

The same calculations can be made for $\frac{\partial \rho}{\partial \mathbf{w}_y}$ by exchanging x and y .

□

6.7 The partial derivatives of the correlation (eq. 24)

$$\begin{cases} \frac{\partial \rho}{\partial \mathbf{w}_x} &= \frac{a}{\|\mathbf{w}_x\|}\left(\mathbf{C}_{xy}\hat{\mathbf{w}}_y - \frac{\hat{\mathbf{w}}_x^T\mathbf{C}_{xy}\hat{\mathbf{w}}_y}{\hat{\mathbf{w}}_x^T\mathbf{C}_{xx}\hat{\mathbf{w}}_x}\mathbf{C}_{xx}\hat{\mathbf{w}}_x\right) \\ \frac{\partial \rho}{\partial \mathbf{w}_y} &= \frac{a}{\|\mathbf{w}_y\|}\left(\mathbf{C}_{yx}\hat{\mathbf{w}}_x - \frac{\hat{\mathbf{w}}_y^T\mathbf{C}_{yx}\hat{\mathbf{w}}_x}{\hat{\mathbf{w}}_y^T\mathbf{C}_{yy}\hat{\mathbf{w}}_y}\mathbf{C}_{yy}\hat{\mathbf{w}}_y\right) \end{cases} \tag{24}$$

Proof: The partial derivative of ρ with respect to $\mathbf{w}_x$ is

$$\begin{aligned}
\frac{\partial \rho}{\partial \mathbf{w}_x} &= \frac{(\mathbf{w}_x^T \mathbf{C}_{xx} \mathbf{w}_x \mathbf{w}_y^T \mathbf{C}_{yy} \mathbf{w}_y)^{1/2} \mathbf{C}_{xy} \mathbf{w}_y}{\mathbf{w}_x^T \mathbf{C}_{xx} \mathbf{w}_x \mathbf{w}_y^T \mathbf{C}_{yy} \mathbf{w}_y} \\
&\quad - \frac{\mathbf{w}_x^T \mathbf{C}_{xy} \mathbf{w}_y (\mathbf{w}_x^T \mathbf{C}_{xx} \mathbf{w}_x \mathbf{w}_y^T \mathbf{C}_{yy} \mathbf{w}_y)^{-1/2} \mathbf{C}_{xx} \mathbf{w}_x \mathbf{w}_y^T \mathbf{C}_{yy} \mathbf{w}_y}{\mathbf{w}_x^T \mathbf{C}_{xx} \mathbf{w}_x \mathbf{w}_y^T \mathbf{C}_{yy} \mathbf{w}_y} \\
&= (\mathbf{w}_x^T \mathbf{C}_{xx} \mathbf{w}_x \mathbf{w}_y^T \mathbf{C}_{yy} \mathbf{w}_y)^{-1/2} \left(\mathbf{C}_{xy} \mathbf{w}_y - \frac{\mathbf{w}_x^T \mathbf{C}_{xy} \mathbf{w}_y}{\mathbf{w}_x^T \mathbf{C}_{xx} \mathbf{w}_x} \mathbf{C}_{xx} \mathbf{w}_x \right) \\
&= \|\mathbf{w}_x\|^{-1} (\underbrace{\hat{\mathbf{w}}_x^T \mathbf{C}_{xx} \hat{\mathbf{w}}_x \hat{\mathbf{w}}_y^T \mathbf{C}_{yy} \hat{\mathbf{w}}_y}_{\geq 0})^{-1/2} \left(\mathbf{C}_{xy} \hat{\mathbf{w}}_y - \frac{\hat{\mathbf{w}}_x^T \mathbf{C}_{xy} \hat{\mathbf{w}}_y}{\hat{\mathbf{w}}_x^T \mathbf{C}_{xx} \hat{\mathbf{w}}_x} \mathbf{C}_{xx} \hat{\mathbf{w}}_x \right) \\
&= \frac{a}{\|\mathbf{w}_x\|} \left(\mathbf{C}_{xy} \hat{\mathbf{w}}_y - \frac{\hat{\mathbf{w}}_x^T \mathbf{C}_{xy} \hat{\mathbf{w}}_y}{\hat{\mathbf{w}}_x^T \mathbf{C}_{xx} \hat{\mathbf{w}}_x} \mathbf{C}_{xx} \hat{\mathbf{w}}_x \right), \qquad a \geq 0.
\end{aligned}$$

The same calculations can be made for $\frac{\partial \rho}{\partial \mathbf{w}_y}$ by exchanging x and y .

□

6.8 The partial derivatives of the MLR-quotient (eq. 35)

$$\begin{cases} \frac{\partial \rho}{\partial \mathbf{w}_x} &= \frac{a}{\|\mathbf{w}_x\|} \left(\mathbf{C}_{xy} \hat{\mathbf{w}}_y - \beta \mathbf{C}_{xx} \hat{\mathbf{w}}_x \right) \\ \frac{\partial \rho}{\partial \mathbf{w}_y} &= \frac{a}{\|\mathbf{w}_x\|} \left(\mathbf{C}_{yx} \hat{\mathbf{w}}_x - \frac{\rho^2}{\beta} \hat{\mathbf{w}}_y \right). \end{cases} \tag{35}$$

Proof: The partial derivative of ρ with respect to $\mathbf{w}_x$ is

$$\begin{aligned}
\frac{\partial \rho}{\partial \mathbf{w}_x} &= \frac{(\mathbf{w}_x^T \mathbf{C}_{xx} \mathbf{w}_x \mathbf{w}_y^T \mathbf{w}_y)^{1/2} \mathbf{C}_{xy} \mathbf{w}_y}{\mathbf{w}_x^T \mathbf{C}_{xx} \mathbf{w}_x \mathbf{w}_y^T \mathbf{w}_y} \\
&\quad - \frac{\mathbf{w}_x^T \mathbf{C}_{xy} \mathbf{w}_y (\mathbf{w}_x^T \mathbf{C}_{xx} \mathbf{w}_x \mathbf{w}_y^T \mathbf{w}_y)^{-1/2} \mathbf{C}_{xx} \mathbf{w}_x \mathbf{w}_y^T \mathbf{w}_y}{\mathbf{w}_x^T \mathbf{C}_{xx} \mathbf{w}_x \mathbf{w}_y^T \mathbf{w}_y} \\
&= (\mathbf{w}_x^T \mathbf{C}_{xx} \mathbf{w}_x \mathbf{w}_y^T \mathbf{w}_y)^{-1/2} \left(\mathbf{C}_{xy} \mathbf{w}_y - \frac{\mathbf{w}_x^T \mathbf{C}_{xy} \mathbf{w}_y}{\mathbf{w}_x^T \mathbf{C}_{xx} \mathbf{w}_x} \mathbf{C}_{xx} \mathbf{w}_x \right) \\
&= \|\mathbf{w}_x\|^{-1} (\underbrace{\hat{\mathbf{w}}_x^T \mathbf{C}_{xx} \hat{\mathbf{w}}_x \hat{\mathbf{w}}_y^T \hat{\mathbf{w}}_y}_{\geq 0})^{-1/2} \left(\mathbf{C}_{xy} \hat{\mathbf{w}}_y - \frac{\hat{\mathbf{w}}_x^T \mathbf{C}_{xy} \hat{\mathbf{w}}_y}{\hat{\mathbf{w}}_x^T \mathbf{C}_{xx} \hat{\mathbf{w}}_x} \mathbf{C}_{xx} \hat{\mathbf{w}}_x \right) \\
&= \frac{a}{\|\mathbf{w}_x\|} \left(\mathbf{C}_{xy} \hat{\mathbf{w}}_y - \beta \mathbf{C}_{xx} \hat{\mathbf{w}}_x \right), \qquad a \geq 0.
\end{aligned}$$

The partial derivative of ρ with respect to $\mathbf{w}_y$ is

$$\begin{aligned}\frac{\partial \rho}{\partial \mathbf{w}_y} &= \frac{(\mathbf{w}_x^T \mathbf{C}_{xx} \mathbf{w}_x \mathbf{w}_y^T \mathbf{w}_y)^{1/2} \mathbf{C}_{yx} \mathbf{w}_x}{\mathbf{w}_x^T \mathbf{C}_{xx} \mathbf{w}_x \mathbf{w}_y^T \mathbf{w}_y} \\ &\quad - \frac{\mathbf{w}_x^T \mathbf{C}_{xy} \mathbf{w}_y (\mathbf{w}_x^T \mathbf{C}_{xx} \mathbf{w}_x \mathbf{w}_y^T \mathbf{w}_y)^{-1/2} \mathbf{w}_x^T \mathbf{C}_{xx} \mathbf{w}_x \mathbf{w}_y}{\mathbf{w}_x^T \mathbf{C}_{xx} \mathbf{w}_x \mathbf{w}_y^T \mathbf{w}_y} \\ &= (\mathbf{w}_x^T \mathbf{C}_{xx} \mathbf{w}_x \mathbf{w}_y^T \mathbf{w}_y)^{-1/2} \left(\mathbf{C}_{yx} \mathbf{w}_x - \frac{\mathbf{w}_x^T \mathbf{C}_{xy} \mathbf{w}_y \mathbf{w}_x^T \mathbf{C}_{xx} \mathbf{w}_x}{\mathbf{w}_x^T \mathbf{C}_{xx} \mathbf{w}_x \mathbf{w}_y^T \mathbf{w}_y} \mathbf{w}_y \right) \\ &= \|\mathbf{w}_y\|^{-1} (\underbrace{\hat{\mathbf{w}}_x^T \mathbf{C}_{xx} \hat{\mathbf{w}}_x}) ^{-1/2} \left(\mathbf{C}_{yx} \hat{\mathbf{w}}_x - \hat{\mathbf{w}}_x^T \mathbf{C}_{xy} \hat{\mathbf{w}}_y \hat{\mathbf{w}}_y \right) \\ &= \frac{a}{\|\mathbf{w}_x\|} \left(\mathbf{C}_{yx} \hat{\mathbf{w}}_x - \frac{\rho^2}{\beta} \hat{\mathbf{w}}_y \right), \qquad a \geq 0.\end{aligned}$$

□

6.9 Combining eigenvalue equations (eqs. 27)

$$\begin{cases} \mathbf{C}_{xx}^{-1} \mathbf{C}_{xy} \mathbf{C}_{yy}^{-1} \mathbf{C}_{yx} \hat{\mathbf{w}}_x = \rho^2 \hat{\mathbf{w}}_x \\ \mathbf{C}_{yy}^{-1} \mathbf{C}_{yx} \mathbf{C}_{xx}^{-1} \mathbf{C}_{xy} \hat{\mathbf{w}}_y = \rho^2 \hat{\mathbf{w}}_y. \end{cases} \tag{27}$$

Proof: Since $\mathbf{C}_{xx}$ and $\mathbf{C}_{yy}$ are nonsingular, equation system 25 can be written as

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

Inserting $\hat{\mathbf{w}}_y$ from the second line into the first line gives

$$\mathbf{C}_{xx}^{-1} \mathbf{C}_{xy} \mathbf{C}_{yy}^{-1} \mathbf{C}_{yx} \hat{\mathbf{w}}_x = \rho^2 \lambda_x \lambda_y \hat{\mathbf{w}}_x = \rho^2 \hat{\mathbf{w}}_x,$$

since $\lambda_x = \lambda_y^{-1}$. This proves the first line in eq. 27. In the same way, solving for $\hat{\mathbf{w}}_x$ proves the second line in eq. 27.

□

6.10 Invariance with respect to linear transformations

Canonical correlations are invariant with respect to linear transformations.

Proof: Let

$$\mathbf{x} = \mathbf{A}_x \mathbf{x}' \quad \text{and} \quad \mathbf{y} = \mathbf{A}_y \mathbf{y}'.$$

where $\mathbf{A}_x$ and $\mathbf{A}_y$ are non-singular matrices. If we denote

$$\mathbf{C}'_{xx} = E\{\mathbf{x}' \mathbf{x}'^T\},$$

then the covariance matrix for $\mathbf{x}$ can be written as

$$\mathbf{C}_{xx} = E\{\mathbf{x}\mathbf{x}^T\} = E\{\mathbf{A}_x \mathbf{x}' \mathbf{x}'^T \mathbf{A}_x^T\} = \mathbf{A}_x \mathbf{C}'_{xx} \mathbf{A}_x^T.$$

In the same way we have

$$\mathbf{C}_{xy} = \mathbf{A}_x \mathbf{C}'_{xy} \mathbf{A}_y^T \quad \text{and} \quad \mathbf{C}_{yy} = \mathbf{A}_y \mathbf{C}'_{yy} \mathbf{A}_y^T.$$

Now, the equation system 27 can be written as

$$\begin{cases} (\mathbf{A}_x^T)^{-1} \mathbf{C}'^{-1}_{xx} (\mathbf{A}_x)^{-1} \mathbf{A}_x \mathbf{C}'_{xy} \mathbf{A}_y^T (\mathbf{A}_y^T)^{-1} \mathbf{C}'^{-1}_{yy} (\mathbf{A}_y)^{-1} \mathbf{A}_y \mathbf{C}'_{yx} \mathbf{A}_x^T \hat{\mathbf{w}}_x &= \rho^2 \hat{\mathbf{w}}_x \\ (\mathbf{A}_y^T)^{-1} \mathbf{C}'^{-1}_{yy} (\mathbf{A}_y)^{-1} \mathbf{A}_y \mathbf{C}'_{yx} \mathbf{A}_x^T (\mathbf{A}_x^T)^{-1} \mathbf{C}'^{-1}_{xx} (\mathbf{A}_x)^{-1} \mathbf{A}_x \mathbf{C}'_{xy} \mathbf{A}_y^T \hat{\mathbf{w}}_y &= \rho^2 \hat{\mathbf{w}}_y, \end{cases}$$

or

$$\begin{cases} {\mathbf{C}'}_{xx}^{-1}\mathbf{C}'_{xy}{\mathbf{C}'}_{yy}^{-1}\mathbf{C}'_{yx}\hat{\mathbf{w}}'_x &= \rho^2\hat{\mathbf{w}}'_x \\ {\mathbf{C}'}_{yy}^{-1}\mathbf{C}'_{yx}{\mathbf{C}'}_{xx}^{-1}\mathbf{C}'_{xy}\hat{\mathbf{w}}'_y &= \rho^2\hat{\mathbf{w}}'_y, \end{cases}$$

where $\hat{\mathbf{w}}'_x = \mathbf{A}_x^T\hat{\mathbf{w}}_x$ and $\hat{\mathbf{w}}'_y = \mathbf{A}_y^T\hat{\mathbf{w}}_y$. Obviously this transformation leaves the roots ρ unchanged. If we look at the canonical variates,

$$\begin{cases} x' &= {\mathbf{w}'}_x^T\mathbf{x}' = \mathbf{w}_x^T\mathbf{A}\mathbf{A}^{-1}\mathbf{x} = x \\ y' &= {\mathbf{w}'}_y^T\mathbf{y}' = \mathbf{w}_y^T\mathbf{A}\mathbf{A}^{-1}\mathbf{y} = y, \end{cases}$$

we see that these too are unaffected by the linear transformation.

□

6.11 The successive eigenvalues (eq. 48)

$$\mathbf{H} = \mathbf{G} - \lambda_1\hat{\mathbf{e}}_1\mathbf{f}_1^T \tag{48}$$

Proof: Consider a vector $\mathbf{u}$ which we express as the sum of one vector parallel to the eigenvector $\hat{\mathbf{e}}_1$, and another vector $\mathbf{u}_o$ that is a linear combination of the other eigenvectors and, hence, orthogonal to the dual vector $\mathbf{f}_1$.

$$\mathbf{u} = a\hat{\mathbf{e}}_1 + \mathbf{u}_o$$

where

$$\mathbf{f}_1^T\hat{\mathbf{e}}_1 = 1 \quad \text{and} \quad \mathbf{f}_1^T\hat{\mathbf{u}}_o = 0.$$

Multiplying $\mathbf{H}$ with $\mathbf{u}$ gives

$$\begin{aligned} \mathbf{H}\mathbf{u} &= \left(\mathbf{G} - \lambda_1\hat{\mathbf{e}}_1\mathbf{f}_1^T\right)\left(a\hat{\mathbf{e}}_1 + \mathbf{u}_o\right) \\ &= a\left(\mathbf{G}\hat{\mathbf{e}}_1 - \lambda_1\hat{\mathbf{e}}_1\right) + \left(\mathbf{G}\mathbf{u}_o - \mathbf{0}\right) \\ &= \mathbf{G}\mathbf{u}_o. \end{aligned}$$

This shows that $\mathbf{G}$ and $\mathbf{H}$ have the same eigenvectors and eigenvalues except for the largest eigenvalue and eigenvector of $\mathbf{G}$. Obviously the eigenvector corresponding to the largest eigenvalue of $\mathbf{H}$ is $\hat{\mathbf{e}}_2$.

□

References

[1] S. Becker. Mutual information maximization: models of cortical self-organization. *Network: Computation in Neural Systems*, 7:7–31, 1996.

[2] A. J. Bell and T. J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7:1129–59, 1995.

[3] R. D. Bock. *Multivariate Statistical Methods in Behavioral Research*. McGraw-Hill series in psychology. McGraw-Hill, 1975.

[4] M. Borga. Reinforcement Learning Using Local Adaptive Models, August 1995. Thesis No. 507, ISBN 91–7871–590–3.

[5] M. Borga, H. Knutsson, and T. Landelius. Learning Canonical Correlations. In *Proceedings of the 10th Scandinavian Conference on Image Analysis*, Lappeenranta, Finland, June 1997. SCIA.

[6] R. Bracewell. *The Fourier Transform and its Applications*. McGraw-Hill, 2nd edition, 1986.

[7] J. Dennis and R. Schnabel. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Prentice Hall, Englewood Cliffs, New Jersey, 1983.

[8] P. Geladi and B. R. Kowalski. Parial least-squares regression: a tutorial. *Analytica Chimica Acta*, 185:1–17, 1986.

[9] G. H. Golub and C. F. Van Loan. *Matrix Computations*. The Johns Hopkins University Press, second edition, 1989.

[10] G. H. Granlund and H. Knutsson. *Signal Processing for Computer Vision*. Kluwer Academic Publishers, 1995. ISBN 0-7923-9530-1.

[11] S. Haykin. Neural networks expand sp's horizons. *IEEE Signal Processing Magazine*, pages 24–49, March 1996.

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

[13] A. Höskuldsson. PLS regression methods. *Journal of Chemometrics*, 2:211–228, 1988.

[14] A. J. Izenman. Reduced-rank regression for the multivariate linear model. *Journal of Multivariate Analysis*, 5:248–264, 1975.

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

[16] H. Knutsson, M. Borga, and T. Landelius. Learning Canonical Correlations. Report LiTH-ISY-R-1761, Computer Vision Laboratory, S–581 83 Linköping, Sweden, June 1995.

[17] T. Landelius. Behavior Representation by Growing a Learning Tree, September 1993. Thesis No. 397, ISBN 91–7871–166–5.

[18] T. Landelius. *Reinforcement Learning and Distributed Local Model Synthesis*. PhD thesis, Linköping University, Sweden, S–581 83 Linköping, Sweden, 1997. Dissertation No 469, ISBN 91–7871–892–9.

[19] T. Landelius, H. Knutsson, and M. Borga. On-Line Singular Value Decomposition of Stochastic Process Covariances. Report LiTH-ISY-R-1762, Computer Vision Laboratory, S–581 83 Linköping, Sweden, June 1995.

[20] E. Oja and J. Karhunen. On stochastic approximation of the eigenvectors and eigenvalues of the expectation of a random matrix. *Journal of Mathematical Analysis and Applications*, 106:69–84, 1985.

[21] D. K. Stewart and W. A. Love. A general canonical correlation index. *Psychological Bulletin*, 70:160–163, 1968.

[22] G. W. Stewart. A bibliographical tour of the large, sparse generalized eigenvalue problem. In J. R. Bunch and D. J. Rose, editors, *Sparse Matrix Computations*, pages 113–130, 1976.

[23] A. L. van den Wollenberg. Redundancy analysis: An alternative for canonical correlation analysis. *Psychometrika*, 36:207–209, 1977.

[24] E. van der Burg. *Nonlinear Canonical Correlation and Some Related Techniques.* DSWO Press, 1988.

[25] S. Wold, A. Ruhe, H. Wold, and W. J. Dunn. The collinearity problem in linear regression. the partial least squares (pls) approach to generalized inverses. *SIAM J. Sci. Stat. Comput.*, 5(3):735–743, 1984.