Generalized Eigenproblem for Stochastic Process Covariances

Hans Knutsson `knutte@isy.liu.se`
Magnus Borga `magnus@isy.liu.se`
Tomas Landelius `tc@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 finding the solution of the generalized eigenproblem where the matrices involved contain expectation values from stochastic processes. The algorithm is iterative and sequential to its structure and uses on-line stochastic approximation to reach an equilibrium point. A quotient between two quadratic forms is suggested as an energy function for this problem and is shown to have zero gradient only at the points solving the eigenproblem. Furthermore it is shown that the algorithm for the generalized eigenproblem can be used to solve three important problems as special cases. For a stochastic process the algorithm can be used to find the directions for maximal variance, covariance, and canonical correlation as well as their magnitudes.

1 Introduction

When dealing with many scientific and engineering problems some version of the *generalized* or *two-matrix* eigenproblem needs to be solved along the way [1, 3, 5]:

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

When this equation turns up in mechanics the eigenvalue λ often refer to vibrations in some way. 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 of the matrices will be hermitian with $\mathbf{B}$ positive definite.

The next section will describe the generalized eigenproblem in some more detail and and an iterative algoritm to solve it is presented. In the subsequent sections it will be shown that three important problems will emerge as special cases of this problem. The algorithm for the generalized eigenproblem can be used to solve three important problems as special cases. For a stochastic process the algorithm can be used to find the directions and magnitudes for maximal variance, covariance, and correlation which corresponds do pricipal component analysis (PCA), singular value decomposition (SVD) and canonical correation analysis repectively .

The ability to perform dimensionality reduction is crucial to systems exposed to high dimensional data. One way of approaching this problem is to project the data on some of the directions of maximal data variation, the largest principal components. There are also a number of applications in signal processing where the largest eigenvalue and the corresponding eigenvalue of input data correlation or covariance matrices play an important role.

When relations between two sets of data, e.g. process input and output, are to be investigated it becomes interesting to find two directions, one in input and one in output space along which the data covariation is maximized. These directions turn out to be the ones accompanying the largest singular value of the between sets covariance matrix.

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 projections of the input and output signals is interesting since this measure of input-output relation is independent of the signal magnitudes.

2 The generalized eigenproblem

The problem described above can alternatively be seen as a special case of the more general problem of maximizing a ratio of *quadratic forms*, i.e. maximizing

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

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 lets look at the gradient of r:

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

where α is a positive factor. 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 in eq. 1. The solutions r_i and $\hat{\mathbf{w}}_i$ are the eigenvalues and eigenvectors respectively of the matrix $\mathbf{B}^{-1}\mathbf{A}$. 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 8.4 on page 13). This means that the $\mathbf{w}_i$'s are linear independent (see proof 8.5 on page 14). Since an n-dimensional space gives n eigenvectors which are linear independent $\{\mathbf{w}_1, \ldots, \mathbf{w}_n\}$ is a *base* and any $\mathbf{w}$ can be expressed as a linear combination of the eigenvectors. Now, it can be proved (see proof 8.6 on page 14) 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 8.7 on page 14). It can be shown (see proof 8.8 on page 15) 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 exist a direction in which r increases. In other words, all extremum points of the function r are saddles except for the global minimum and maximum points.

2.1 The algorithm

Since the only stable critical point is the global maximum it should be possible to find the largest eigenvalue and its corresponding vector by performing a stochastic gradient search on the energy function r. This can be done with an iterative algorithm:

$$\mathbf{w}(t+1) = \mathbf{w}(t) + \Delta\mathbf{w}(t), \tag{9}$$

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

$$E\{\Delta\mathbf{w}\} = \beta\frac{\partial r}{\partial \mathbf{w}} = \alpha(\mathbf{A}\hat{\mathbf{w}} - r\mathbf{B}\hat{\mathbf{w}}) = \alpha(\mathbf{A}\hat{\mathbf{w}} - \mathbf{B}\mathbf{w}) \tag{10}$$

where α and β are positive numbers. Here we use the length of the vector to represent the corresponding eigenvalue, i.e. $\|\mathbf{w}\| = r$.

It is, of course, possible to enhance this update rule 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 a procedures is, for the batch or off-line case, known as Gauss-Newton methods [2]. In this paper we will however 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 three sections it will be shown that three important problems mentioned in the introduction will emerge as special cases of the above problem. It will also be shown how to apply the algorithm to the data in a sequential manner in order to find not only the largest components but the entire eigensystem of the problem.

3 Direction of maximal data variation

For a set of random numbers with zero mean, $\{x_k\}$ the variance is defined as $E\{xx\}$. Now lets turn to a set of random vectors, again with zero mean. In this case we talk about the covariance matrix, here defined according to:

$$\mathbf{C}_{xx} = E\{\mathbf{x}\mathbf{x}^T\} \tag{11}$$

With the direction of maximal data variation we now mean the direction $\hat{\mathbf{w}}$ with the property that the linear combination $x = \hat{\mathbf{w}}^T\mathbf{x}$ posses maximal variance. Finding this direction is hence equal to finding the maximum of

$$\rho = E\{xx\} = E\{\hat{\mathbf{w}}^T\mathbf{x}\hat{\mathbf{w}}^T\mathbf{x}\} = \hat{\mathbf{w}}^T E\{\mathbf{x}\mathbf{x}^T\}\hat{\mathbf{w}} = \frac{\mathbf{w}^T\mathbf{C}_{xx}\mathbf{w}}{\mathbf{w}^T\mathbf{w}}. \tag{12}$$

Obviously the solution of this problem is a special case of that presented in eq. 2 with

$$\mathbf{A} = \mathbf{C}_{xx} \quad \text{and} \quad \mathbf{B} = \mathbf{I}. \tag{13}$$

Hence we can find the direction of maximal data variation by a stochastic gradient search according to eq. 10 with appropriate $\mathbf{A}$ and $\mathbf{B}$.

$$E\{\Delta\mathbf{w}\} = \beta\frac{\partial\rho}{\partial\mathbf{w}} = \alpha\,[\mathbf{C}_{xx}\hat{\mathbf{w}} - \rho\hat{\mathbf{w}}] = \alpha\,E\{\mathbf{x}\mathbf{x}^T\hat{\mathbf{w}} - \rho\hat{\mathbf{w}}\}. \tag{14}$$

Now let the length of the vector represent the estimated variance, i.e. $\|\mathbf{w}\| = \rho$. This leads to a novel unsupervised Hebbian learning algorithm that finds both the direction of maximal data variation and how much the data varies along that direction. The update rule for this algorithm is given by

$$\Delta\mathbf{w} = \alpha\,(\mathbf{x}\mathbf{x}^T\hat{\mathbf{w}} - \mathbf{w}), \tag{15}$$

where 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 and time dependent.

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

$$\mathbf{C}_{xx} = E\{\mathbf{x}\mathbf{x}^T\} = \sum \lambda_i \ \hat{\mathbf{e}}_i \hat{\mathbf{e}}_i^T \tag{16}$$

where λ_i and $\hat{\mathbf{e}}_i$ are the eigenvalues and orthogonal eigenvectors respectively. This is whown as principal component analysis (PCA).

The problem of maximizing the quotient, ρ, can then be seen as finding the largest eigenvalue, λ_1, and its corresponding eigenvector since:

$$\lambda_1 = \hat{\mathbf{e}}_1^T \mathbf{C}_{xx} \hat{\mathbf{e}}_1 = \max \frac{\mathbf{w}^T \mathbf{C}_{xx} \mathbf{w}}{\mathbf{w}^T \mathbf{w}} = \max \rho. \tag{17}$$

When the largest singular value and its corresponding vector is found it is a simple matter to proceed and find the second largest pair, and so on. In order to do this a new signal vector, containing all signal information orthogonal to the ones already estimated, is constructed and used as a new input to the algorithm. The details of this procedure is described in section 6.1.

It is also worth noting that it is possible to find the direction and magnitude of maximal data variation to the inverse of the covariance matrix. In this case we simply identify the matrices as $\mathbf{A} = \mathbf{I}$ and $\mathbf{B} = \mathbf{C}_{xx}$. The learning rule associated with this behavior then becomes:

$$\Delta \mathbf{w} = \alpha \ (\hat{\mathbf{w}} - \mathbf{x}\mathbf{x}^T \mathbf{w}). \tag{18}$$

4 Directions of maximal 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 again consider the multivariate case we can define the between sets covariance matrix according to:

$$\mathbf{C}_{xy} = E\{\mathbf{x}\mathbf{y}^T\} = (E\{\mathbf{y}\mathbf{x}^T\})^T = \mathbf{C}_{yx}^T \tag{19}$$

This time we look at the *two* directions of maximal data covariation, by which we now 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 maximal covariance. This means that we want to maximize the following function:

$$\rho = E\{xy\} = E\{\hat{\mathbf{w}}_x^T \mathbf{x} \hat{\mathbf{w}}_y^T \mathbf{y}\} = \hat{\mathbf{w}}_x^T E\{\mathbf{x}\mathbf{y}^T\} \hat{\mathbf{w}}_y = \frac{\mathbf{w}_x^T \mathbf{C}_{xy} \mathbf{w}_y}{\sqrt{\mathbf{w}_x^T \mathbf{w}_x \mathbf{w}_y^T \mathbf{w}_y^T}}. \tag{20}$$

Taking the derivatives of this function with respect to the vectors $\mathbf{w}_x$ and $\mathbf{w}_y$ gives

$$\begin{cases} \frac{\partial \rho}{\partial \mathbf{w}_x} & = \frac{1}{\|\mathbf{w}_x\|}(\mathbf{C}_{xy} \hat{\mathbf{w}}_y - \rho \hat{\mathbf{w}}_x) \\ \frac{\partial \rho}{\partial \mathbf{w}_y} & = \frac{1}{\|\mathbf{w}_y\|}(\mathbf{C}_{yx} \hat{\mathbf{w}}_x - \rho \hat{\mathbf{w}}_y). \end{cases} \tag{21}$$

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

$$\begin{cases} \mathbf{C}_{xy} \mathbf{C}_{yx} \hat{\mathbf{w}}_x & = \rho^2 \hat{\mathbf{w}}_x \\ \mathbf{C}_{yx} \mathbf{C}_{xy} \hat{\mathbf{w}}_y & = \rho^2 \hat{\mathbf{w}}_y. \end{cases} \tag{22}$$

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

$$\mathbf{A} = \begin{pmatrix} \mathbf{0} & \mathbf{C}_{xy} \\ \mathbf{C}_{yx} & \mathbf{0} \end{pmatrix} \quad , \quad \mathbf{B} = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \quad \text{, and} \quad \mathbf{w} = \begin{pmatrix} \mu_x \hat{\mathbf{w}}_x \\ \mu_y \hat{\mathbf{w}}_y \end{pmatrix}. \tag{23}$$

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

$$\begin{cases} \mathbf{C}_{xy} \hat{\mathbf{w}}_y & = r \frac{\mu_x}{\mu_y} \hat{\mathbf{w}}_x \\ \mathbf{C}_{yx} \hat{\mathbf{w}}_x & = r \frac{\mu_y}{\mu_x} \hat{\mathbf{w}}_y \end{cases} \tag{24}$$

and then solving for $\mathbf{w}_x$ and $\mathbf{w}_y$ which gives equation 22 with $r^2 = \rho^2$. Hence, the problem of finding the direction and magnitude of the largest data covariation can be seen as the special case of eq. 2 with the appropriate choice of matrices. In this case the length of the vector $\mathbf{w}$ would equal the magnitude of the covariance, $\|\mathbf{w}\| = \rho$ and the singular vectors are identified from $\mathbf{w}^T = (\mu_x \hat{\mathbf{w}}_x^T,\ \mu_y \hat{\mathbf{w}}_y^T)$.

This means that the algorithm for the solution of the generalized eigenproblem can be used also for the problem described above. Since we want to update $\mathbf{w}$, on average, in direction of the gradient and have specified the matrices $\mathbf{A}$ and $\mathbf{B}$ as above this leads us to:

$$E\{\Delta\mathbf{w}\} = \beta\frac{\partial r}{\partial \mathbf{w}} = \alpha\left[\begin{pmatrix}\mathbf{0} & \mathbf{C}_{xy}\\ \mathbf{C}_{yx} & \mathbf{0}\end{pmatrix}\hat{\mathbf{w}} - r\begin{pmatrix}\mathbf{I} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}\end{pmatrix}\hat{\mathbf{w}}\right]. \tag{25}$$

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

$$\Delta\mathbf{w} = \alpha\left[\begin{pmatrix}\mathbf{0} & \mathbf{x}\mathbf{y}^T\\ \mathbf{y}\mathbf{x}^T & \mathbf{0}\end{pmatrix}\hat{\mathbf{w}} - \mathbf{w}\right] \tag{26}$$

where the length of the vector represents the covariance, i.e. $\|\mathbf{w}\| = r = \rho$.

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:

$$\mathbf{C}_{xy} = E_{xy}\{\mathbf{x}\mathbf{y}^T\} = \sum \lambda_i\ \hat{\mathbf{e}}_{xi}\hat{\mathbf{e}}_{yi}^T \tag{27}$$

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. 20 is equal to finding the largest singular value:

$$\lambda_1 = \hat{\mathbf{e}}_{x1}^T\mathbf{C}_{xy}\hat{\mathbf{e}}_{y1} = \max\frac{\mathbf{w}_x^T\mathbf{C}_{xy}\mathbf{w}_y}{\sqrt{\mathbf{w}_x^T\mathbf{w}_x\mathbf{w}_y^T\mathbf{w}_y}} = \max\rho. \tag{28}$$

Again, a complete SVD can be found by construction of orthogonal signal vectors which are used as new inputs to the algorithm. This procedure will be discussed in section 6.2.

5 The directions of maximal canonical correlation

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

$$\mathbf{C} = \begin{pmatrix}\mathbf{C}_{xx} & \mathbf{C}_{xy}\\ \mathbf{C}_{yx} & \mathbf{C}_{yy}\end{pmatrix} = E\left\{\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix}\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix}^T\right\} \tag{29}$$

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 $E\{xy\}/\sqrt{E\{xx\}E\{yy\}}$ which means that the function we want to maximize can be written:

$$\rho = \frac{E\{xy\}}{\sqrt{E\{xx\}E\{yy\}}} = \frac{E\{\hat{\mathbf{w}}_x^T\mathbf{x}\mathbf{y}^T\hat{\mathbf{w}}_y\}}{\sqrt{E\{\hat{\mathbf{w}}_x^T\mathbf{x}\mathbf{x}^T\hat{\mathbf{w}}_x\}E\{\hat{\mathbf{w}}_y^T\mathbf{y}\mathbf{y}^T\hat{\mathbf{w}}_y\}}} = \frac{\hat{\mathbf{w}}_x^T\mathbf{C}_{xy}\hat{\mathbf{w}}_y}{\sqrt{\hat{\mathbf{w}}_x^T\mathbf{C}_{xx}\hat{\mathbf{w}}_x\hat{\mathbf{w}}_y^T\mathbf{C}_{yy}\hat{\mathbf{w}}_y}}. \tag{30}$$

We want to find the vectors $\hat{\mathbf{w}}_x$ and $\hat{\mathbf{w}}_y$ that maximizes ρ. The partial derivatives of ρ with respect to $\hat{\mathbf{w}}_x$ and $\hat{\mathbf{w}}_y$ are (see proof 8.1 on page 12)

$$\begin{cases}\frac{\partial\rho}{\partial\mathbf{w}_x} &= a\left(\mathbf{C}_{xy}\hat{\mathbf{w}}_y - \frac{\hat{\mathbf{w}}_x^T\mathbf{C}_{xy}\hat{\mathbf{w}}_y}{\hat{\mathbf{w}}_x^T\mathbf{C}_{xx}\hat{\mathbf{w}}_x}\mathbf{C}_{xx}\hat{\mathbf{w}}_x\right)\\ \frac{\partial\rho}{\partial\mathbf{w}_y} &= a\left(\mathbf{C}_{yx}\hat{\mathbf{w}}_x - \frac{\hat{\mathbf{w}}_y^T\mathbf{C}_{yx}\hat{\mathbf{w}}_x}{\hat{\mathbf{w}}_y^T\mathbf{C}_{yy}\hat{\mathbf{w}}_y}\mathbf{C}_{yy}\hat{\mathbf{w}}_y\right)\end{cases} \tag{31}$$

[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 normalised with the geometric mean of the variables' variances.

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

$$\begin{cases} \mathbf{C}_{xy}\hat{\mathbf{w}}_y = \rho\lambda_x \mathbf{C}_{xx}\hat{\mathbf{w}}_x \\ \mathbf{C}_{yx}\hat{\mathbf{w}}_x = \rho\lambda_y \mathbf{C}_{yy}\hat{\mathbf{w}}_y \end{cases} \tag{32}$$

where

$$\lambda_x = \lambda_y^{-1} = \sqrt{\frac{\hat{\mathbf{w}}_y^T \mathbf{C}_{yy}\hat{\mathbf{w}}_y}{\hat{\mathbf{w}}_x^T \mathbf{C}_{xx}\hat{\mathbf{w}}_x}} \tag{33}$$

λ_x is the ratio between the standard deviation of x and the standard deviation of y and vice versa. The λ's can be interpreted as a scaling factor between the linear combinations. Rewriting equation system 32 gives (see proof 8.2 on page 13)

$$\begin{cases} \mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}\mathbf{C}_{yy}^{-1}\mathbf{C}_{yx}\hat{\mathbf{w}}_x = \rho^2\hat{\mathbf{w}}_x \\ \mathbf{C}_{yy}^{-1}\mathbf{C}_{yx}\mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}\hat{\mathbf{w}}_y = \rho^2\hat{\mathbf{w}}_y. \end{cases} \tag{34}$$

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 and the corresponding eigenvalues ρ^2 are the squared *canonical correlations* [4]. It should, however, be noted that the relationship between $\hat{\mathbf{w}}_x$ and $\hat{\mathbf{w}}_y$ given by equation system 32 still applies even though it can't be seen from equation 34. That means, e.g. that a change of sign of $\hat{\mathbf{w}}_x$ also causes a change of sign of $\hat{\mathbf{w}}_x$, and vice versa. It also means that there are only N solutions $\{\hat{\mathbf{w}}_{xn}, \hat{\mathbf{w}}_{yn}, \rho_n\}, 1 \leq n \leq N$ where N is the minimum of the dimensionalities of $\mathbf{x}$ and $\mathbf{y}$. 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

$$\mathbf{A} = \begin{pmatrix} \mathbf{0} & \mathbf{C}_{xy} \\ \mathbf{C}_{yx} & \mathbf{0} \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} \mathbf{C}_{xx} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{yy} \end{pmatrix}, \quad \text{and} \quad \mathbf{w} = \begin{pmatrix} \mathbf{w}_x \\ \mathbf{w}_y \end{pmatrix} = \begin{pmatrix} \mu_x\hat{\mathbf{w}}_x \\ \mu_y\hat{\mathbf{w}}_y \end{pmatrix} \tag{35}$$

we can write equation 4 as

$$\begin{cases} \mathbf{C}_{xy}\hat{\mathbf{w}}_y = r\frac{\mu_x}{\mu_y}\mathbf{C}_{xx}\hat{\mathbf{w}}_x \\ \mathbf{C}_{yx}\hat{\mathbf{w}}_x = r\frac{\mu_y}{\mu_x}\mathbf{C}_{yy}\hat{\mathbf{w}}_y \end{cases} \tag{36}$$

which we recognize as equation 32 if we let $\lambda_x = \frac{\mu_x}{\mu_y}$ and $\lambda_y = \frac{\mu_y}{\mu_x}$. We can see that the gradient of r have the same directions as the gradient of ρ and is zero for the same $\hat{\mathbf{w}}$'s. If we solve for $\mathbf{w}_x$ and $\mathbf{w}_y$ in eq. 36 we will end up in eq. 34. This shows that we obtain the equations for the canonical correlations as the result of a gradient search on the energy function r.

Again the algorithm for solving the generalized eigenproblem can be used for the stochastic gradient search. In order to update $\mathbf{w}$, on average, in direction of the gradient of r, with the matrices $\mathbf{A}$ and $\mathbf{B}$ as in eq. 35 above, we obtain the update direction as:

$$E\{\Delta\mathbf{w}\} = \beta\frac{\partial r}{\mathbf{w}} = \alpha\left[\begin{pmatrix} \mathbf{0} & \mathbf{C}_{xy} \\ \mathbf{C}_{yx} & \mathbf{0} \end{pmatrix}\hat{\mathbf{w}} - r\begin{pmatrix} \mathbf{C}_{xx} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{yy} \end{pmatrix}\hat{\mathbf{w}}\right]. \tag{37}$$

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

$$\Delta\mathbf{w} = \alpha\left[\begin{pmatrix} \mathbf{0} & \mathbf{x}\mathbf{y}^T \\ \mathbf{y}\mathbf{x}^T & \mathbf{0} \end{pmatrix}\hat{\mathbf{w}} - \begin{pmatrix} \mathbf{x}\mathbf{x}^T & \mathbf{0} \\ \mathbf{0} & \mathbf{y}\mathbf{y}^T \end{pmatrix}\mathbf{w}\right]. \tag{38}$$

Since we will have $\|\mathbf{w}\| = r = \rho$ when the algorithm converges the length of the vector represents the correlation between the variates. From eq. 35 and eq. 36 we see that the quotient of the lengths of the vectors $\mathbf{w}_x$ and $\mathbf{w}_y$ equals the λ's which are the scaling factors (see equation 33). These quotients can be used to form least square estimators of y given x and vice versa. This means that the best (in a least square error sense) estimates $\tilde{x}$ and $\tilde{y}$ of the variates x and y are

$$\tilde{y} = \frac{\|\mathbf{w}\|}{\|\mathbf{w}_y\|}\,\mathbf{x}^T\mathbf{w}_x \quad \text{and} \quad \tilde{x} = \frac{\|\mathbf{w}\|}{\|\mathbf{w}_x\|}\,\mathbf{y}^T\mathbf{w}_y.$$

With these notations we can write equation 38 as

$$\Delta \mathbf{w} = \begin{pmatrix} \Delta \mathbf{w}_x \\ \Delta \mathbf{w}_y \end{pmatrix} = \frac{\alpha}{\|\mathbf{w}\|} \begin{pmatrix} \mu_y \mathbf{x} \, (y - \tilde{y}) \\ \mu_x \mathbf{y} \, (x - \tilde{x}) \end{pmatrix} \tag{39}$$

which shows that the algorithm have a stable point where $\tilde{y}$ on average equals y and $\tilde{x}$ on average equals x.

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 34. For invariance with respect to other linear transformations see proof 8.3 on page 13.

Also in this case it is possible to use the same algorithm in order to find the subsequent canonical correlations. This is the subject of the next section.

6 Finding successive eigenvalues and eigenvectors

Since the learning rule defined in eq. 10 tries to maximize the Rayleigh quotient in eq. 2 it will find the largest eigenvalue $\|\mathbf{w}_1\| = \lambda_1$ and the corresponding eigenvector $\hat{\mathbf{w}}_1 = \pm \hat{\mathbf{e}}_1$. 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. We will first present the general solution to this problem and then use this solution to modify our learning rule to suite the three special cases treated in the previous section.

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

$$\mathbf{GE} = \mathbf{ED} \quad \Rightarrow \quad \mathbf{G} = \mathbf{EDE}^{-1} = \sum \lambda_i \hat{\mathbf{e}}_i \mathbf{f}_i^T, \tag{40}$$

where the eigenvalues and vectors are gathered in the matrices $\mathbf{D}$ and $\mathbf{E}$ respectively:

$$\mathbf{D} = \begin{pmatrix} \lambda_1 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \lambda_n \end{pmatrix}, \quad \mathbf{E} = \begin{pmatrix} | & & | \\ \hat{\mathbf{e}}_1 & \cdots & \hat{\mathbf{e}}_n \\ | & & | \end{pmatrix}, \quad \mathbf{E}^{-1} = \begin{pmatrix} - & \mathbf{f}_1^T & - \\ & \vdots & \\ - & \mathbf{f}_n^T & - \end{pmatrix}. \tag{41}$$

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

$$\mathbf{f}_i^T \hat{\mathbf{e}}_j = \delta_{ij}. \tag{42}$$

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

$$\hat{\mathbf{e}}_i^T \mathbf{A} \hat{\mathbf{e}}_j = 0 \quad \text{and} \quad \hat{\mathbf{e}}_i^T \mathbf{B} \hat{\mathbf{e}}_j = 0 \quad \text{for} \quad i \neq j. \tag{43}$$

Hence we can use this result to find the dual vectors $\mathbf{f}_i$ possessing the property in e.q. 42, e.g. by choosing them according to:

$$\mathbf{f}_i = \frac{\mathbf{B}\hat{\mathbf{e}}_i}{\hat{\mathbf{e}}_i^T \mathbf{B} \hat{\mathbf{e}}_i}. \tag{44}$$

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

$$\mathbf{G} - \lambda_1 \hat{\mathbf{e}}_1 \mathbf{f}_1^T \tag{45}$$

will have the same eigenvectors and eigenvalues as $\mathbf{G}$ except for the eigenvalue corresponding to $\pm\hat{\mathbf{e}}_1$, i.e. λ_1 for $\mathbf{G}$ which now becomes 0 (see proof 8.9 on page 15). This means that the eigenvectors corresponding to the largest eigenvalue of the modified matrix are the same as those corresponding to the second largest eigenvalue of $\mathbf{G}$.

Since the algorithm will first find the vector $\mathbf{w}_1 = \pm\lambda_1\hat{\mathbf{e}}_1$ 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 build up $\mathbf{G}$ 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 versions of these two matrices, $\mathbf{A}^{'}$ and $\mathbf{B}^{'}$, with the following property:

$$\mathbf{B}^{'-1}\mathbf{A}^{'} = \mathbf{B}^{-1}\mathbf{A} - \lambda_1\hat{\mathbf{e}}_1\mathbf{f}_1^T. \tag{46}$$

A simple solution is obtained if we only modify one of the matrices and keep the other matrix fix, $\mathbf{B}^{'} = \mathbf{B}$:

$$\mathbf{A}^{'} = \mathbf{A} - \lambda_1\mathbf{B}\hat{\mathbf{e}}_1\mathbf{f}_1^T. \tag{47}$$

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

$$\mathbf{u}_1(t+1) = \mathbf{u}_1(t) + \Delta\mathbf{u}_1(t) \tag{48}$$

where

$$E\{\Delta\mathbf{u}_1\} = \alpha\left[\mathbf{B}\mathbf{w}_1 - \mathbf{u}_1\right]. \tag{49}$$

Once this estimate has converged, the outer product in e.q. 47 can be expressed as:

$$\lambda_1\mathbf{B}\hat{\mathbf{e}}_1\mathbf{f}_1^T = \frac{\mathbf{u}_1\mathbf{u}_1^T}{\hat{\mathbf{e}}_1^T\mathbf{u}_1}. \tag{50}$$

A modified version of the learning algorithm in e.q. 10 which finds the second eigenvalue and vector to the generalized eigenproblem has the following form:

$$E\{\Delta\mathbf{w}\} = \alpha\ \left[\mathbf{A}^{'}\hat{\mathbf{w}} - \mathbf{B}\mathbf{w}\right] = \alpha\ \left[(\mathbf{A} - \frac{\mathbf{u}_1\mathbf{u}_1^T}{\hat{\mathbf{w}}_1^T\mathbf{u}_1})\hat{\mathbf{w}} - \mathbf{B}\mathbf{w}\right]. \tag{51}$$

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

This scheme can of course be repeated to find the third canonical correlation by subtracting the second solution in the same way and so on.

6.1 Variance

In the special case where the interesting entity is the variance along a direction in the signal space we have the following structure for the matrices $\mathbf{A}$ and $\mathbf{B}$:

$$\mathbf{A} = \mathbf{C}_{xx} \quad \text{and} \quad \mathbf{B} = \mathbf{I}. \tag{52}$$

Hence we get 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 e.q. 47 then becomes:

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

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

$$E\{\Delta\mathbf{w}\} = \alpha\ \left[\mathbf{A}^{'}\hat{\mathbf{w}} - \mathbf{B}\mathbf{w}\right] = \alpha\ \left[(\mathbf{C}_{xx} - \mathbf{w}_1\hat{\mathbf{w}}_1^T)\hat{\mathbf{w}} - \mathbf{w}\right], \tag{54}$$

which is achieved if we at each time step update the vector $\mathbf{w}$ by

$$\Delta\mathbf{w} = \alpha\ \left[(\mathbf{x}\mathbf{x}^T - \mathbf{w}_1\hat{\mathbf{w}}_1^T)\hat{\mathbf{w}} - \mathbf{w}\right]. \tag{55}$$

6.2 Covariance

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 maximal data covariance. We have

$$\mathbf{A} = \begin{pmatrix} \mathbf{0} & \mathbf{C}_{xy} \\ \mathbf{C}_{yx} & \mathbf{0} \end{pmatrix} \quad \text{and} \quad \mathbf{B} = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}. \tag{56}$$

which again means that the compound matrix $\mathbf{G} = \mathbf{B}^{-1}\mathbf{A} = \mathbf{A}$ will be symmetric and have orthogonal eigenvectors, which means that they will be identical to their dual vectors. The outer product for modification of the matrix $\mathbf{A}$ in e.q. 47 becomes identical to the one presented in the previous section:

$$\lambda_1 \mathbf{B}\hat{\mathbf{e}}_1\mathbf{f}_1^T = \lambda_1 \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \hat{\mathbf{e}}_1\hat{\mathbf{e}}_1^T = \mathbf{w}_1\hat{\mathbf{w}}_1^T. \tag{57}$$

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

$$E\{\Delta\mathbf{w}\} = \alpha \left[\mathbf{A}^{'}\hat{\mathbf{w}} - \mathbf{B}\mathbf{w}\right] = \alpha \left[(\begin{pmatrix} \mathbf{0} & \mathbf{C}_{xy} \\ \mathbf{C}_{yx} & \mathbf{0} \end{pmatrix} - \mathbf{w}_1\hat{\mathbf{w}}_1^T)\hat{\mathbf{w}} - \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}\mathbf{w}\right], \tag{58}$$

which is achieved if we at each time step update the vector $\mathbf{w}$ by

$$\Delta\mathbf{w} = \alpha \left[(\begin{pmatrix} \mathbf{0} & \mathbf{x}\mathbf{y}^T \\ \mathbf{y}\mathbf{x}^T & \mathbf{0} \end{pmatrix} - \mathbf{w}_1\hat{\mathbf{w}}_1^T)\hat{\mathbf{w}} - \mathbf{w}\right]. \tag{59}$$

6.3 Canonical correlation

In the two previous cases it was easy to zero out an eigenvalue because the matrix $\mathbf{G}$ was symmetric. This is not the case when we now turn to canonical correlation. Here we have

$$\mathbf{A} = \begin{pmatrix} \mathbf{0} & \mathbf{C}_{xy} \\ \mathbf{C}_{yx} & \mathbf{0} \end{pmatrix} \quad \text{and} \quad \mathbf{B} = \begin{pmatrix} \mathbf{C}_{xx} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{yy} \end{pmatrix}, \tag{60}$$

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

$$\mathbf{G} = \mathbf{B}^{-1}\mathbf{A} = \begin{pmatrix} \mathbf{C}_{xx}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{yy}^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{0} & \mathbf{C}_{xy} \\ \mathbf{C}_{yx} & \mathbf{0} \end{pmatrix} = \begin{pmatrix} \mathbf{0} & \mathbf{C}_{xx}^{-1}\mathbf{C}_{xy} \\ \mathbf{C}_{yy}^{-1}\mathbf{C}_{yx} & \mathbf{0} \end{pmatrix}. \tag{61}$$

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 e.q. 49 for the general case:

$$E\{\Delta\mathbf{u}_1\} = \alpha \left[\mathbf{B}\mathbf{w}_1 - \mathbf{u}_1\right] = \alpha \left[\begin{pmatrix} \mathbf{C}_{xx} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{yy} \end{pmatrix}\mathbf{w}_1 - \mathbf{u}_1\right]. \tag{62}$$

To, on average, update according to the equation above we do the following:

$$\Delta\mathbf{u}_1 = \alpha \left[\begin{pmatrix} \mathbf{x}\mathbf{x}^T & \mathbf{0} \\ \mathbf{0} & \mathbf{y}\mathbf{y}^T \end{pmatrix}\mathbf{w}_1 - \mathbf{u}_1\right]. \tag{63}$$

With this estimate, the outer product in e.q. 47 can be used to modify the matrix $\mathbf{A}$:

$$\mathbf{A}^{'} = \mathbf{A} - \lambda_1\mathbf{B}\hat{\mathbf{e}}_1\mathbf{f}_1^T = \mathbf{A} - \frac{\mathbf{u}_1\mathbf{u}_1^T}{\hat{\mathbf{w}}_1^T\mathbf{u}_1}. \tag{64}$$

A modified version of the learning algorithm in e.q. 10 which finds the second largest canonical correlations and its corresponding directions can be written on the following form:

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

Again to do this on average we perform the update at each time step according to:

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

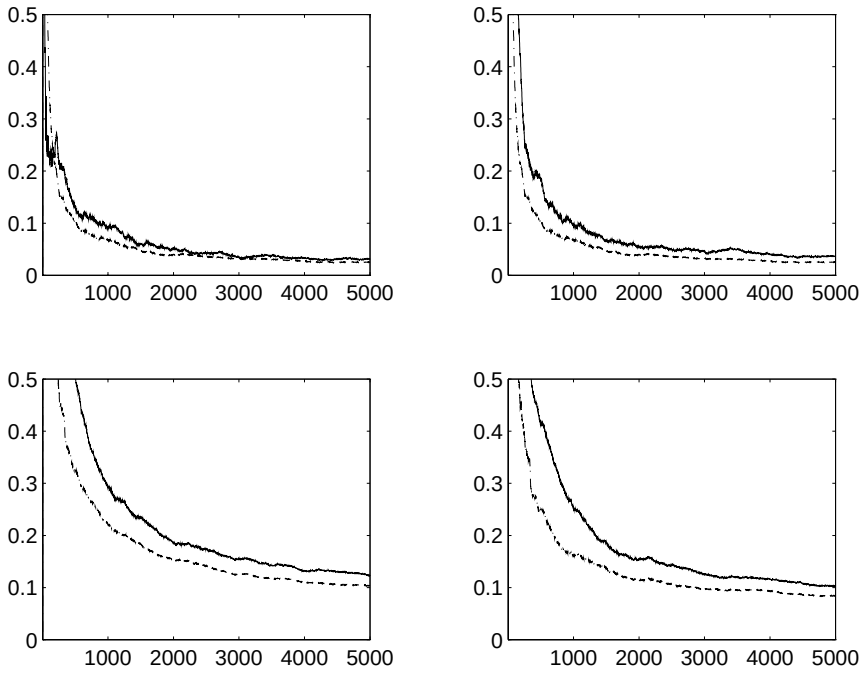

Figure 1: *Averaged errors in* $\mathbf{w}_x$ *(left) and* $\mathbf{w}_y$ *(right). Top row shows magnitude errors and bottom row angular errors. Estimates are shown with a solid line for the proposed algorithm and with a dashed line for the optimal approach.*

7 Experiments

This section presents some experiment with the proposed algorithm for solving two of the described specializations of the generalized eigenvalue problem, singular value decomposition and canonical correlation. In order to make a statement about the performance of the proposed algorithm the "optimal", in the sense of maximum likelihood, deterministic solution was calculated based on the data accumulated so far.

No extra time was spent on finding an optimal set of parameters for the algorithm. Instead the experiments and comparisons were carried out only to display the behaviour of the algorithm and show that it is robust and works.

7.1 Experiments on singular value decomposition

For use in this comparison, two random vector sequences were produced. In the shown example the dimensionality of an instance in the $\{\mathbf{x}_k\}$ and $\{\mathbf{y}_k\}$ sequences was 10 and 5 respectively. The largest singular value was equal to 10. The singular values of the covariance matrix describing the vector distribution declined as $\exp(-0.5i)$, where i is the index to the i:th largest singular value.

Magnitude and angular errors were calculated as $|1-\|\mathbf{w}\|/\|\mathbf{w}_c\||$ and $\arccos(\hat{\mathbf{w}}^T\hat{\mathbf{w}}_c)$ respectively, where $\mathbf{w}_c$ is the correct singular vector. The error estimates were averaged over 50 runs, each consisting of 5000 instances from the vector distributions. These averaged measurements are shown on the top and bottom row of figure 1. The right and left column of the figure corresponds to errors in $\{\mathbf{w}_x\}$ and $\{\mathbf{w}_y\}$ respectively. For the proposed algorithm the gain sequence $\gamma_k = (1+\beta)/(k+\beta)$, with $\beta = 1.25$, was used.

Even though the difference in computational is $\mathcal{O}(d^3)$ to $\mathcal{O}(d)$ in favour to the proposed algorithm, its behavior is still quite similar to that of optimal brute force SVD.

7.2 Experiments on canonical correlations

The results from this experiment show that the presented algorithm, which has complexity $\mathcal{O}(N)$, has a performance comparable to 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.

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 67.

$$\begin{cases} E_x \Leftarrow E_x + b\ (\ \|\mathbf{x}\mathbf{x}^T\mathbf{w}_x\| - E_x) \\ E_y \Leftarrow E_y + b\ (\ \|\mathbf{y}\mathbf{y}^T\mathbf{w}_y\| - E_y) \\ \alpha_x = a\lambda_x E_x^{-1} \\ \alpha_y = a\lambda_y E_y^{-1} \end{cases} \tag{67}$$

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 68.

$$\begin{cases} \Delta_x \Leftarrow \Delta_x + d\ (\Delta\mathbf{w}_x - \Delta_x) \\ \Delta_y \Leftarrow \Delta_y + d\ (\Delta\mathbf{w}_y - \Delta_y) \\ \mathbf{w}_{\mathbf{a}x} \Leftarrow \mathbf{w}_{\mathbf{a}x} + c\ \|\Delta_x\|\ \|\mathbf{w}_x\|^{-1}\ (\mathbf{w}_x - \mathbf{w}_{\mathbf{a}x}) \\ \mathbf{w}_{\mathbf{a}y} \Leftarrow \mathbf{w}_{\mathbf{a}y} + c\ \|\Delta_y\|\ \|\mathbf{w}_y\|^{-1}\ (\mathbf{w}_y - \mathbf{w}_{\mathbf{a}y}) \end{cases} \tag{68}$$

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 2 are the averages over the 10 runs. The errors in directions for the vectors $\mathbf{w}_{\mathbf{a}x1}, \mathbf{w}_{\mathbf{a}x2}, \mathbf{w}_{\mathbf{a}y1}$ and $\mathbf{w}_{\mathbf{a}y2}$ 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.

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

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.

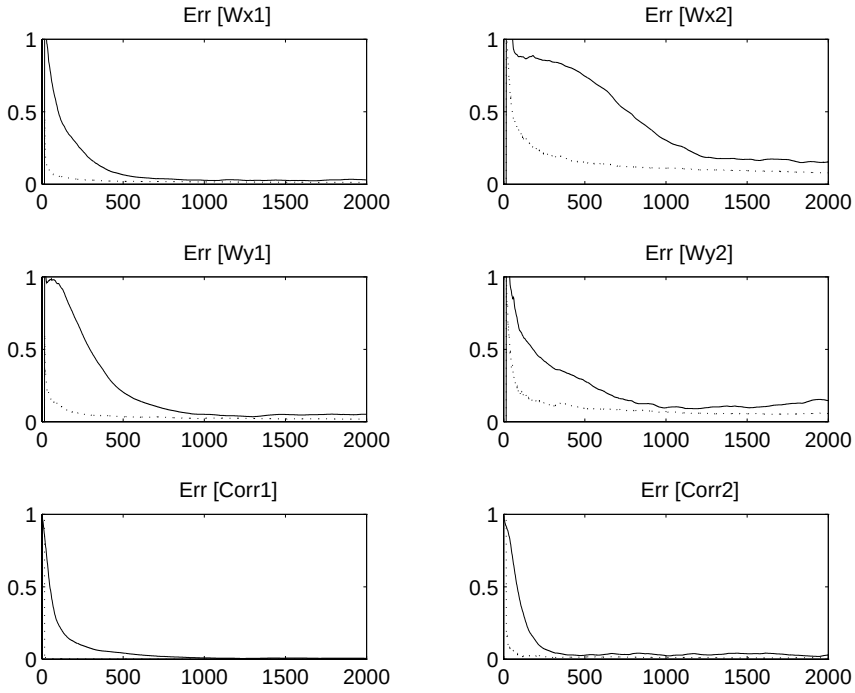

Figure 2: *Error magnitudes averaged over 10 runs of the algorithm. The solid lines displays the differences between the algorithm and the exact values. The dotted lines shows the differences between the optimal solutions obtained by solving the eigenvector equations and the exact values, (see text for further explanation). The top row shows the error angles in radians for* $\hat{\mathbf{w}}_{\mathbf{a}x}$*. The middle row shows the same errors for* $\hat{\mathbf{w}}_{\mathbf{a}y}$*. 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.*

8 Proofs

8.1 The partial derivatives of ρ (eq. 31)

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

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

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

□

8.2 Combining eigenvalue equations (eqs. 34)

Since $\mathbf{C}_{xx}$ and $\mathbf{C}_{yy}$ are nonsingular, equation system 32 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. 34. In the same way by inserting $\mathbf{w}_x$ from the first line into the second line is proved.

□

8.3 Invariance with respect to linear transformations (page 7)

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 34 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

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

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,

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

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

□

8.4 Orthogonality in the metrics A and B (eq. 5

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 symmetric 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 that $\hat{\mathbf{w}}_i^T \mathbf{B} \hat{\mathbf{w}}_j > 0$ 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 $\hat{\mathbf{w}}_i^T \mathbf{A} \hat{\mathbf{w}}_i = r_i \hat{\mathbf{w}}_i^T \mathbf{B} \hat{\mathbf{w}}_i$.

□

8.5 Linear independence

Suppose $\mathbf{w}_i$ are *not* linear 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

$$\hat{\mathbf{w}}_i^T \mathbf{B} \hat{\mathbf{w}}_k = \gamma_i \hat{\mathbf{w}}_i \mathbf{B} \hat{\mathbf{w}}_i \neq 0$$

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

□

8.6 The range of r (eq. 6)

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

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

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 and its minimum r_n when $\gamma_n \neq 0$ and $\gamma_i = 0 \; \forall \; i < n$ if r_n is the smallest eigenvalue.

□

8.7 The second derivative of r (eq. 7)

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

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

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

$$\frac{\partial r}{\partial \mathbf{w}} = \mathbf{A}\mathbf{w} - r\mathbf{B}\mathbf{w} = \mathbf{0}$$

and hence

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

□

8.8 Positive eigenvalues of the hessian (eq. 8)

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.

□

8.9 The successive eigenvalues (eq. 42)

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

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

where

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

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

$$\begin{aligned}
\mathbf{B}\mathbf{u} &= \left(\mathbf{A} - \lambda_1 \mathbf{u}_1 \mathbf{v}_1^T\right)(a\hat{\mathbf{u}}_1 + \mathbf{u}_o) \\
&= a\left(\mathbf{A}\hat{\mathbf{u}}_1 - \lambda_1 \hat{\mathbf{u}}_1\right) + \left(\mathbf{A}\hat{\mathbf{u}}_o - \mathbf{0}\right) \\
&= \mathbf{A}\mathbf{u}_o.
\end{aligned}$$

This shows that $\mathbf{A}$ and $\mathbf{B}$ have the same eigenvectors and eigenvalues except for the largest one. Obviously the eigenvector corresponding to the largest eigenvalue of $\mathbf{B}$ is $\hat{\mathbf{u}}_2$.

□

References

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

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

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

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

[5] 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.