# Canonical Correlation
# *a Tutorial*

Magnus Borga

January 12, 2001

## Contents

## 1 About this tutorial

This is a printable version of a tutorial in HTML format. The tutorial may be modified at any time as will this version. The latest version of this tutorial is available at `http://people.imt.liu.se/˜magnus/cca/`.

# 2 Introduction

Canonical correlation analysis (CCA) is a way of measuring the linear relationship between two multidimensional variables. It finds two bases, one for each variable, that are optimal with respect to correlations and, at the same time, it finds the corresponding correlations. In other words, it finds the two bases in which the correlation matrix between the variables is diagonal and the correlations on the diagonal are maximized. The dimensionality of these new bases is equal to or less than the smallest dimensionality of the two variables.

An important property of canonical correlations is that they are invariant with respect to affine transformations of the variables. This is the most important difference between CCA and ordinary correlation analysis which highly depend on the basis in which the variables are described.

CCA was developed by H. Hotelling [10]. Although being a standard tool in statistical analysis, where canonical correlation has been used for example in economics, medical studies, meteorology and even in classification of malt whisky, it is surprisingly unknown in the fields of learning and signal processing. Some exceptions are [2, 13, 5, 4, 14],

For further details and applications in signal processing, see my PhD thesis [3] and other publications.

# 3 Definition

Canonical correlation analysis can be defined as the problem of finding two sets of basis vectors, one for $\mathbf{x}$ and the other for $\mathbf{y}$, such that the correlations between the *projections* of the variables onto these basis vectors are mutually maximized.

Let us look at the case where only one pair of basis vectors are sought, namely the ones corresponding to the largest canonical correlation: Consider the linear combinations $x = \mathbf{x}^T \hat{\mathbf{w}}_x$ and $y = \mathbf{y}^T \hat{\mathbf{w}}_y$ of the two variables respectively. This means that the function to be maximized is

$$\begin{aligned}\rho &= \frac{E[xy]}{\sqrt{E[x^2]E[y^2]}} = \frac{E[\hat{\mathbf{w}}_x^T \mathbf{x}\mathbf{y}^T \hat{\mathbf{w}}_y]}{\sqrt{E[\hat{\mathbf{w}}_x^T \mathbf{x}\mathbf{x}^T \hat{\mathbf{w}}_x]E[\hat{\mathbf{w}}_y^T \mathbf{y}\mathbf{y}^T \hat{\mathbf{w}}_y]}} \\ &= \frac{\mathbf{w}_x^T \mathbf{C}_{xy} \mathbf{w}_y}{\sqrt{\mathbf{w}_x^T \mathbf{C}_{xx} \mathbf{w}_x \mathbf{w}_y^T \mathbf{C}_{yy} \mathbf{w}_y}}.\end{aligned} \tag{1}$$

The maximum of $\rho$ with respect to $\mathbf{w}_x$ and $\mathbf{w}_y$ is the maximum canonical correlation. The subsequent canonical correlations are uncorrelated for different solutions, i.e.

$$\begin{cases} E[x_i x_j] &= E[\mathbf{w}_{xi}^T \mathbf{x}\mathbf{x}^T \mathbf{w}_{xj}] = \mathbf{w}_{xi}^T \mathbf{C}_{xx} \mathbf{w}_{xj} = 0 \\ E[y_i y_j] &= E[\mathbf{w}_{yi}^T \mathbf{y}\mathbf{y}^T \mathbf{w}_{yj}] = \mathbf{w}_{yi}^T \mathbf{C}_{yy} \mathbf{w}_{yj} = 0 \\ E[x_i y_j] &= E[\mathbf{w}_{xi}^T \mathbf{x}\mathbf{y}^T \mathbf{w}_{yj}] = \mathbf{w}_{xi}^T \mathbf{C}_{xy} \mathbf{w}_{yj} = 0 \end{cases} \quad \text{for} \quad i \neq j. \tag{2}$$

The projections onto $\mathbf{w}_x$ and $\mathbf{w}_y$, i.e. $x$ and $y$, are called *canonical variates*.

# 4 Calculating canonical correlations

Consider two random variables $\mathbf{x}$ and $\mathbf{y}$ with zero mean. The total covariance matrix

$$\mathbf{C} = \begin{bmatrix} \mathbf{C}_{xx} & \mathbf{C}_{xy} \\ \mathbf{C}_{yx} & \mathbf{C}_{yy} \end{bmatrix} = E\left[ \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix}^T \right] \tag{3}$$

is a block matrix where $\mathbf{C}_{xx}$ and $\mathbf{C}_{xx}$ are the within-sets covariance matrices of $\mathbf{x}$ and $\mathbf{y}$ respectively and $\mathbf{C}_{xy} = \mathbf{C}_{yx}^T$ is the between-sets covariance matrix.

The canonical correlations between $\mathbf{x}$ and $\mathbf{y}$ can be found by solving the eigenvalue equations

$$\begin{cases} \mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}\mathbf{C}_{yy}^{-1}\mathbf{C}_{yx}\hat{\mathbf{w}}_x = \rho^2\hat{\mathbf{w}}_x \\ \mathbf{C}_{yy}^{-1}\mathbf{C}_{yx}\mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}\hat{\mathbf{w}}_y = \rho^2\hat{\mathbf{w}}_y \end{cases} \tag{4}$$

where the eigenvalues $\rho^2$ are the squared *canonical correlations* and the eigenvectors $\hat{\mathbf{w}}_x$ and $\hat{\mathbf{w}}_y$ are the normalized canonical correlation *basis vectors*. The number of non-zero solutions to these equations are limited to the smallest dimensionality of $\mathbf{x}$ and $\mathbf{y}$. E.g. if the dimensionality of $\mathbf{x}$ and $\mathbf{y}$ is 8 and 5 respectively, the maximum number of canonical correlations is 5.

Only one of the eigenvalue equations needs to be solved since the solutions are related by

$$\begin{cases} \mathbf{C}_{xy}\hat{\mathbf{w}}_y = \rho\lambda_x\mathbf{C}_{xx}\hat{\mathbf{w}}_x \\ \mathbf{C}_{yx}\hat{\mathbf{w}}_x = \rho\lambda_y\mathbf{C}_{yy}\hat{\mathbf{w}}_y, \end{cases} \tag{5}$$

where

$$\lambda_x = \lambda_y^{-1} = \sqrt{\frac{\hat{\mathbf{w}}_y^T\mathbf{C}_{yy}\hat{\mathbf{w}}_y}{\hat{\mathbf{w}}_x^T\mathbf{C}_{xx}\hat{\mathbf{w}}_x}}. \tag{6}$$

# 5 Relating topics

## 5.1 The difference between CCA and ordinary correlation analysis

Ordinary correlation analysis is dependent on the coordinate system in which the variables are described. This means that even if there is a very strong linear relationship between two multidimensional signals, this relationship may not be visible in a ordinary correlation analysis if one coordinate system is used, while in another coordinate system this linear relationship would give a very high correlation.

CCA finds the coordinate system that is optimal for correlation analysis, and the eigenvectors of equation 4 defines this coordinate system.

**Example:** Consider two normally distributed two-dimensional variables $\mathbf{x}$ and $\mathbf{y}$ with unit variance. Let $y_1 + y_2 = x_1 + x_2$. It is easy to confirm that the correlation matrix between $\mathbf{x}$ and $\mathbf{y}$ is

$$\mathbf{R}_{xy} = \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix}. \tag{7}$$

This indicates a relatively weak correlation of 0.5 despite the fact that there is a perfect linear relationship (in one dimension) between $\mathbf{x}$ and $\mathbf{y}$.

A CCA on this data shows that the largest (and only) canonical correlation is one and it also gives the direction $[11]^T$ in which this perfect linear relationship lies. If the variables are described in the bases given by the canonical correlation basis vectors (i.e. the eigenvectors of equation 4), the correlation matrix between the variables is

$$\mathbf{R}_{xy} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}. \tag{8}$$

### 5.2 Relation to mutual information

There is a relation between correlation and mutual information. Since information is additive for statistically independent variables and the canonical variates are uncorrelated, the mutual information between $\mathbf{x}$ and $\mathbf{y}$ is the sum of mutual information between the variates $x_i$ and $y_i$ if there are no higher order statistic dependencies than correlation (second-order statistics). For Gaussian variables this means

$$I(\mathbf{x};\mathbf{y}) = \frac{1}{2}\log\left(\frac{1}{\prod_i (1-\rho_i^2)}\right) = \frac{1}{2}\sum_i \log\left(\frac{1}{(1-\rho_i^2)}\right). \tag{9}$$

Kay [13] has shown that this relation plus a constant holds for all elliptically symmetrical distributions of the form

$$cf((\mathbf{z}-\bar{\mathbf{z}})^T\mathbf{C}^{-1}(\mathbf{z}-\bar{\mathbf{z}})). \tag{10}$$

### 5.3 Relation to other linear subspace methods

Instead of the two eigenvalue equations in 4 we can formulate the problem in one single eigenvalue equation:

$$\mathbf{B}^{-1}\mathbf{A}\hat{\mathbf{w}} = \rho\hat{\mathbf{w}} \tag{11}$$

where

$$\mathbf{A} = \begin{bmatrix} \mathbf{0} & \mathbf{C}_{xy} \\ \mathbf{C}_{yx} & \mathbf{0} \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} \mathbf{C}_{xx} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{yy} \end{bmatrix} \quad \text{and} \quad \hat{\mathbf{w}} = \begin{pmatrix} \mu_x\hat{\mathbf{w}}_x \\ \mu_y\hat{\mathbf{w}}_y \end{pmatrix}. \tag{12}$$

Solving the eigenproblem in equation 11 with slightly different matrices will give solutions to *principal component analysis* (PCA), *partial least squares (PLS) and* multivariate linear regression (MLR). The matrices are listed in table 1.

|  | $\mathbf{A}$ | $\mathbf{B}$ |
|---|---|---|
| PCA | $\mathbf{C}_{xx}$ | $\mathbf{I}$ |
| PLS | $\begin{pmatrix} \mathbf{0} & \mathbf{C}_{xy} \\ \mathbf{C}_{yx} & \mathbf{0} \end{pmatrix}$ | $\begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}$ |
| CCA | $\begin{pmatrix} \mathbf{0} & \mathbf{C}_{xy} \\ \mathbf{C}_{yx} & \mathbf{0} \end{pmatrix}$ | $\begin{pmatrix} \mathbf{C}_{xx} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{yy} \end{pmatrix}$ |
| MLR | $\begin{pmatrix} \mathbf{0} & \mathbf{C}_{xy} \\ \mathbf{C}_{yx} & \mathbf{0} \end{pmatrix}$ | $\begin{pmatrix} \mathbf{C}_{xx} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}$ |

Table 1: The matrices $\mathbf{A}$ and $\mathbf{B}$ for PCA, PLS, CCA and MLR.

### 5.4 Relation to SNR

Correlation is strongly related to signal to noise ratio (SNR), which is a more commonly used measure in signal processing. Consider a signal $x$ and two noise signals $\eta_1$ and $\eta_2$ all having zero mean[1] and all being uncorrelated with each other. Let $S = E[x^2]$ and $N_i = E[\eta_i^2]$ be the energy of the signal and the noise signals respectively. Then the correlation between $a(x + \eta_1)$ and $b(x + \eta_2)$ is

$$\begin{aligned} \rho &= \frac{E\left[a(x+\eta_1)b(x+\eta_2)\right]}{\sqrt{E\left[a^2(x+\eta_1)^2\right]E\left[b^2(x+\eta_2)^2\right]}} \\ &= \frac{E\left[x^2\right]}{\sqrt{\left(E\left[x^2\right]+E\left[\eta_1^2\right]\right)\left(E\left[x^2\right]+E\left[\eta_2^2\right]\right)}} \\ &= \frac{S}{\sqrt{(S+N_1)(S+N_2)}}. \end{aligned} \tag{13}$$

Note that the amplification factors $a$ and $b$ do not affect the correlation or the SNR.

#### 5.4.1 Equal noise energies

In the special case where the noise energies are equal, i.e.
$N_1 = N_2 = N$, equation 13 can be written as

$$\rho = \frac{S}{S+N}. \tag{14}$$

This means that the SNR can be written as

$$\frac{S}{N} = \frac{\rho}{1-\rho}. \tag{15}$$

[1] The assumption of zero mean is for convenience. A non-zero mean does not affect the SNR or the correlation.

Here, it should be noted that the noise affects the signal *twice*, so this relation between SNR and correlation is perhaps not so intuitive. This relation is illustrated in figure 1 (top).

### 5.4.2 Correlation between a signal and the corrupted signal

Another special case is when $N_1 = 0$ and $N_2 = N$. Then, the correlation between a signal and a noise-corrupted version of that signal is

$$\rho = \frac{S}{\sqrt{S(S+N)}}. \tag{16}$$

In this case, the relation between SNR and correlation is

$$\frac{S}{N} = \frac{\rho^2}{1-\rho^2}. \tag{17}$$

This relation between correlation and SNR is illustrated in figure 1 (bottom).

# A Explanations

## A.1 A note on correlation and covariance matrices

In neural network literature, the matrix $\mathbf{C}_{xx}$ in equation 3 is often called a correlation matrix. This can be a bit confusing, since $\mathbf{C}_{xx}$ does not contain the correlations between the variables in a statistical sense, but rather the expected values of the products between them. The correlation between $x_i$ and $x_j$ is defined as

$$\rho_{ij} = \frac{E[(x_i - \bar{x}_i)(x_j - \bar{x}_j)]}{\sqrt{E[(x_i - \bar{x}_i)^2]E[(x_j - \bar{x}_j)^2]}}, \tag{18}$$

see for example[1], i.e. the covariance between $x_i$ and $x_j$ normalized by the geometric mean of the variances of $x_i$ and $x_j$ ($\bar{x} = E[x]$). Hence, the correlation is bounded, $-1 \leq \rho_{ij} \leq 1$. In this tutorial, correlation matrices are denoted $\mathbf{R}$.

The diagonal terms of $\mathbf{C}_{xx}$ are the second order *origin* moments, $E[x_i^2]$, of $x_i$. The diagonal terms in a *covariance matrix* are the variances or the second order *central* moments, $E[(x_i - \bar{x_i})^2]$, of $x_i$.

The maximum likelihood estimator of $\rho$ is obtained by replacing the expectation operator in equation 18 by a sum over the samples. This estimator is sometimes called the *Pearson correlation coefficient* after K. Pearson[16].

## A.2 Affine transformations

An affine transformation is simply a translation of the origin followed by a linear transformation. In mathematical terms an affine transformation of $\mathbb{R}^n$ is a map $F : \mathbb{R}^n \to \mathbb{R}^n$ of the form

$$F(\mathbf{p}) = \mathbf{A}\mathbf{p} + \mathbf{q} \quad \forall \mathbf{p} \in \mathbb{R}^n$$

where $\mathbf{A}$ is a linear transformation of $\mathbb{R}^n$ and $\mathbf{q}$ is a translation vector in $\mathbb{R}^n$.

## A.3 A piece of information theory

Consider a discrete random variable $\mathbf{x}$:

$$\mathbf{x} \in \{\mathbf{x}_i\},\ i \in \{1, 2, \ldots, N\}. \tag{19}$$

(There is, in practice, no limitation in $\mathbf{x}$ being discrete since all measurements have finite precision.) Let $P(\mathbf{x}_k)$ be the probability of $\mathbf{x} = \mathbf{x}_k$ for a randomly chosen $\mathbf{x}$. The *information* content in the vector (or symbol) $\mathbf{x}_k$ is defined as

$$I(\mathbf{x}_k) = \log\left(\frac{1}{P(\mathbf{x}_k)}\right) = -\log P(\mathbf{x}_k). \tag{20}$$

If the basis 2 is used for the logarithm, the information is measured in *bits*. The definition of information has some appealing properties. First, the information is 0 if $P(\mathbf{x}_k) = 1$; if the receiver of a message knows that the message will be $\mathbf{x}_k$, he does not get any information when he receives the message. Secondly, the information is always positive. It is not possible to lose information by receiving a message. Finally, the information is additive, i.e. the information in two independent symbols is the sum of the information in each symbol:

$$\begin{aligned} I(\mathbf{x}_i, \mathbf{x}_j) &= -\log\left(P(\mathbf{x}_i, \mathbf{x}_j)\right) = -\log\left(P(\mathbf{x}_i)P(\mathbf{x}_j)\right) \\ &= -\log P(\mathbf{x}_i) - \log P(\mathbf{x}_j) = I(\mathbf{x}_i) + I(\mathbf{x}_j) \end{aligned} \tag{21}$$

if $\mathbf{x}_i$ and $\mathbf{x}_j$ are statistically independent.

The information measure considers each *instance* of the stochastic variable $\mathbf{x}$ but it does not say anything about the stochastic variable itself. This can be accomplished by calculating the average information of the stochastic variable:

$$H(\mathbf{x}) = \sum_{i=1}^{N} P(\mathbf{x}_i)I(\mathbf{x}_i) = -\sum_{i=1}^{N} P(\mathbf{x}_i)\log(P(\mathbf{x}_i)). \tag{22}$$

$H(\mathbf{x})$ is called the *entropy* of $\mathbf{x}$ and is a measure of *uncertainty* about $\mathbf{x}$.

Now, we introduce a second discrete random variable $\mathbf{y}$, which, for example, can be an output signal from a system with $\mathbf{x}$ as input. The *conditional entropy* [18] of $\mathbf{x}$ given $\mathbf{y}$ is

$$H(\mathbf{x}|\mathbf{y}) = H(\mathbf{x}, \mathbf{y}) - H(\mathbf{y}). \tag{23}$$

The conditional entropy is a measure of the average information in $\mathbf{x}$ given that $\mathbf{y}$ is known. In other words, it is the remaining uncertainty of $\mathbf{x}$ after observing $\mathbf{y}$. The *average mutual information*[2] $I(\mathbf{x}; \mathbf{y})$ between $\mathbf{x}$ and $\mathbf{y}$ is defined as the average information about $\mathbf{x}$ gained when observing $\mathbf{y}$:

$$I(\mathbf{x}; \mathbf{y}) = H(\mathbf{x}) - H(\mathbf{x}|\mathbf{y}). \tag{24}$$

[2] Shannon48 originally used the term *rate of transmission*. The term *mutual information* was introduced later.

The mutual information can be interpreted as the difference between the uncertainty of $\mathbf{x}$ and the remaining uncertainty of $\mathbf{x}$ after observing $\mathbf{y}$. In other words, it is the reduction in uncertainty of $\mathbf{x}$ gained by observing $\mathbf{y}$. Inserting equation 23 into equation 24 gives

$$I(\mathbf{x};\mathbf{y}) = H(\mathbf{x}) + H(\mathbf{y}) - H(\mathbf{x},\mathbf{y}) = I(\mathbf{y};\mathbf{x}) \tag{25}$$

which shows that the mutual information is symmetric.

Now let $\mathbf{x}$ be a continuous random variable. Then the *differential entropy* $h(\mathbf{x})$ is defined as [18]

$$h(\mathbf{x}) = -\int_{\mathbb{R}^N} p(\mathbf{x}) \log p(\mathbf{x})\; d\mathbf{x}, \tag{26}$$

where $p(\mathbf{x})$ is the probability density function of $\mathbf{x}$. The integral is over all dimensions in $\mathbf{x}$. The average information in a continuous variable would of course be infinite since there are an infinite number of possible outcomes. This can be seen if the discrete entropy definition (eq. 22) is calculated in limes when $x$ approaches a continuous variable:

$$H(x) = -\lim_{\delta x \to 0} \sum_{i=-\infty}^{\infty} p(x_i)\delta x \log\left(p(x_i)\delta x\right) = h(x) - \lim_{\delta x \to 0} \log \delta x, \tag{27}$$

where the last term approaches infinity when $\delta x$ approaches zero [8]. But since mutual information considers the difference in entropy, the infinite term will vanish and continuous variables can be used to simplify the calculations. The mutual information between the continuous random variables $\mathbf{x}$ and $\mathbf{y}$ is then

$$I(\mathbf{x};\mathbf{y}) = h(\mathbf{x}) + h(\mathbf{y}) - h(\mathbf{x},\mathbf{y}) = \int_{\mathbb{R}^N}\int_{\mathbb{R}^M} p(\mathbf{x},\mathbf{y}) \log\left(\frac{p(\mathbf{x},\mathbf{y})}{p(\mathbf{x})p(\mathbf{y})}\right) d\mathbf{x}d\mathbf{y}, \tag{28}$$

where $N$ and $M$ are the dimensionalities of $\mathbf{x}$ and $\mathbf{y}$ respectively.

Consider the special case of Gaussian distributed variables. The differential entropy of an $N$-dimensional Gaussian variable $\mathbf{z}$ is

$$h(\mathbf{z}) = \frac{1}{2}\log\left((2\pi e)^N |\mathbf{C}|\right) \tag{29}$$

where $\mathbf{C}$ is the covariance matrix of $\mathbf{z}$ [3]. This means that the mutual information between two $N$-dimensional Gaussian variables is

$$I(\mathbf{x};\mathbf{y}) = \frac{1}{2}\log\left(\frac{|\mathbf{C}_{xx}|\;|\mathbf{C}_{yy}|}{|\mathbf{C}|}\right), \tag{30}$$

where

$$\mathbf{C} = \begin{bmatrix} \mathbf{C}_{xx} & \mathbf{C}_{xy} \\ \mathbf{C}_{yx} & \mathbf{C}_{yy} \end{bmatrix}.$$

$\mathbf{C}_{xx}$ and $\mathbf{C}_{yy}$ are the within-set covariance matrices and $\mathbf{C}_{xy} = \mathbf{C}_{yx}^T$ is the between-sets covariance matrix. For more details on information theory, see for example [7].

## A.4 Principal component analysis

Principal component analysis (PCA) is an old tool in multivariate data analysis. It was used already in 1901 [17]. The principal components are the eigenvectors of the covariance matrix. The projection of data onto the principal components is sometimes called the Hotelling transform after H. Hotelling[9] or Karhunen-Loéve transform (KLT) after K. Karhunen and [12] and M. Loéve [15]. This transformation is as an orthogonal transformation that diagonalizes the covariance matrix.

## A.5 Partial least squares

partial least squares (PLS) was developed in econometrics in the 1960s by Herman Wold. It is most commonly used for regression in the field of chemometrics [19]. PLS i basically the singular-value decomposition (SVD) of a between-sets covariance matrix.

For an overview, see for example [6] and [11]. In PLS regression, the principal vectors corresponding to the largest principal values are used as a new, lower dimensional, basis for the signal. A regression of $\mathbf{y}$ onto $\mathbf{x}$ is then performed in this new basis. As in the case of PCA, the scaling of the variables affects the solutions of the PLS.

## A.6 Multivariate linear regression

Multivariate linear regression (MLR) is the problem of finding a set of basis vectors $\hat{\mathbf{w}}_{xi}$ and corresponding regressors $\beta_i$ in order to minimize the mean square error of the vector $\mathbf{y}$:

$$\epsilon^2 = E\left[ \|y_i - \sum_{i=1}^{M} \beta_i \hat{\mathbf{w}}_{xi}^T \mathbf{x}\|^2 \right] \tag{31}$$

where $M = \dim(\mathbf{y})$. The basis vectors are described by the matrix $\mathbf{C}xx^{-1}\mathbf{C}xy$ which is also known as the *Wiener filter*. A low-rank approximation to this problem can be defined by minimizing

$$\epsilon^2 = E\left[ \|\mathbf{y} - \sum_{i=1}^{N} \beta_i \hat{\mathbf{w}}_{xi}^T \mathbf{x} \hat{\mathbf{w}}_{yi}\|^2 \right] \tag{32}$$

where $N < M$ and the orthogonal basis $\hat{\mathbf{w}}_{yi}$s span the subspace of $\mathbf{y}$ which gives the smallest mean square error given the rank $N$. The bases $\{\hat{\mathbf{w}}_{wi}\}$ and $\{\hat{\mathbf{w}}_{yi}\}$ are given by the solutions to

$$\begin{cases} \mathbf{C}_{xy}\hat{\mathbf{w}}_y = \beta \mathbf{C}_{xx}\hat{\mathbf{w}}_x \\ \mathbf{C}_{yx}\hat{\mathbf{w}}_x = \frac{\rho^2}{\beta}\hat{\mathbf{w}}_y, \end{cases} \tag{33}$$

which can be recognized from equation 11 with $\mathbf{A}$ and $\mathbf{B}$ from the lower row in table 1.

### A.7 Signal to noise ratio

The signal to noise ratio (SNR) is the quotient between the signal energy and the noise energy. It is usually expressed as dB (decibel) which is a logarithmic scale:

$$\text{SNR} = 10 \log \frac{S}{N}$$

where $S$ is the signal energy and $N$ is the noise energy.

## References

[1] T. W. Anderson. *An Introduction to Multivariate Statistical Analysis*. John Wiley & Sons, second edition, 1984.

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

[3] M. Borga. *Learning Multidimensional Signal Processing*. PhD thesis, Linköping University, Sweden, SE-581 83 Linköping, Sweden, 1998. Dissertation No 531, ISBN 91-7219-202-X.

[4] S. Das and P. K. Sen. Restricted canonical correlations. *Linear Algebra and its Applications*, 210:29–47, 1994.

[5] P. W. Fieguth, W. W. Irving, and A. S. Willsky. Multiresolution model development for overlapping trees via canonical correlation analysis. In *International Conference on Image Processing*, pages 45–48, Washington DC., 1995. IEEE.

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

[7] R. M. Gray. *Entropy and Information Theory*. Springer-Verlag, New York, 1990.

[8] S. Haykin. *Neural Networks: A Comprehensive Foundation*. Macmillan College Publishing Company, 1994.

[9] H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24:417–441, 498–520, 1933.

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

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

[12] K. Karhunen. Uber lineare methoden in der Wahrsccheilichkeitsrechnung. *Annales Academiae Scientiarum Fennicae, Seried A1: Mathematica-Physica*, 37:3–79, 1947.

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

[14] P. Li, J. Sun, and B. Yu. Direction finding using interpolated arrays in unknown noise fields. *Signal Processing*, 58:319–325, 1997.

[15] M. Loéve. *Probability Theory*. Van Nostrand, New York, 1963.

[16] K. Pearson. Mathematical contributions to the theory of evolution–III. Regression, heridity and panmixia. *Philosophical Transaction of the Royal Society of London, Series A*, 187:253–318, 1896.

[17] K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2:559–572, 1901.

[18] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 1948. Also in N. J. A. Sloane and A. D. Wyner (ed.) *Claude Elwood Shannon Collected Papers*, IEEE Press 1993.

[19] 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.

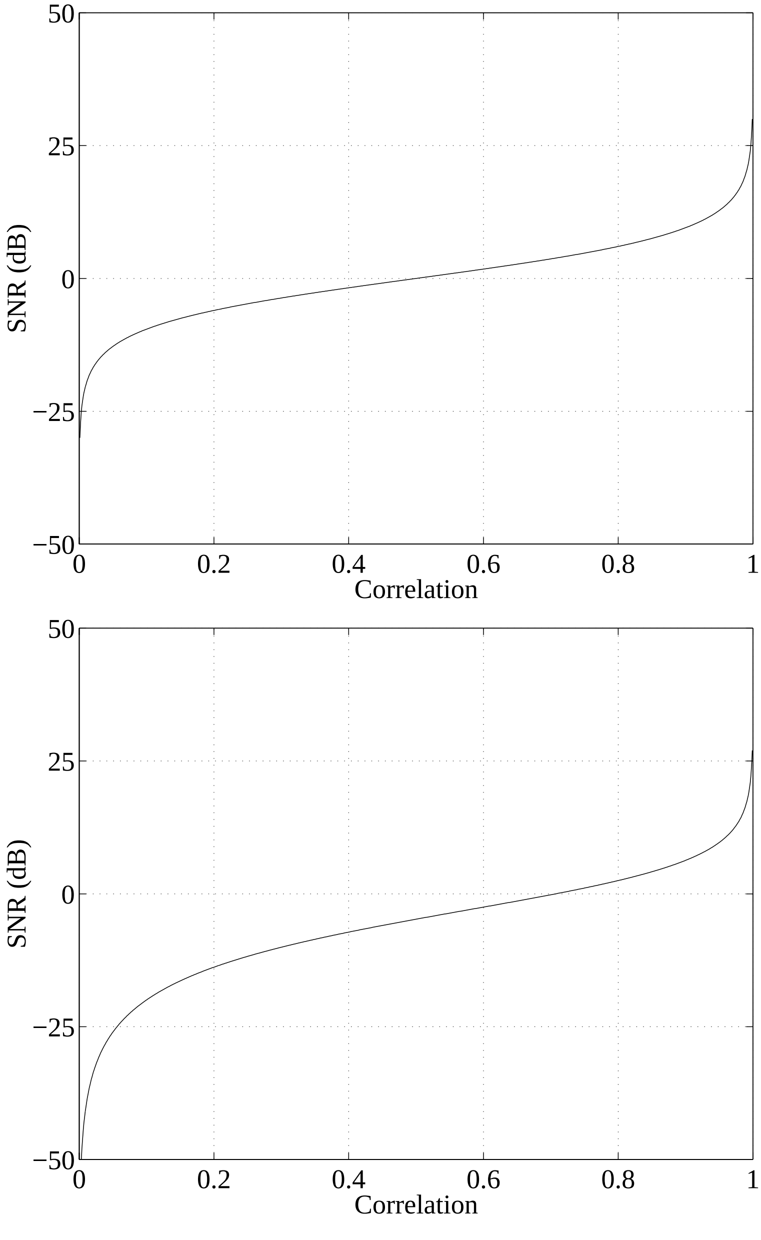

Figure 1: **Top:** The relation between correlation and SNR for two signals each corrupted by uncorrelated noise. Both noise signals have the same energy. **Bottom:** The relation between correlation and SNR. The correlation is measured between a signal and a noise-corrupted version of that signal.