ON THE REGULARIZATION OF CANONICAL CORRELATION ANALYSIS

Tijl De Bie

Katholieke Universiteit Leuven
ESAT-SCD
Kasteelpark Arenberg 10
3001 Leuven
tijl.debie@esat.kuleuven.ac.be

Bart De Moor

Katholieke Universiteit Leuven
ESAT-SCD
Kasteelpark Arenberg 10
3001 Leuven
bart.demoor@esat.kuleuven.ac.be

ABSTRACT

By elucidating a parallel between canonical correlation analysis (CCA) and least squares regression (LSR), we show how regularization of CCA can be performed and interpreted in the same spirit as the regularization applied in ridge regression (RR).

Furthermore, the results presented may have an impact on the practical use of regularized CCA (RCCA). More specifically, a relevant cross validation cost function for training the regularization parameter, naturally follows from the derivations.

1. INTRODUCTION

In this paper, we present a unifying approach to CCA [2] and RR [1] from the viewpoint of estimation, and from the viewpoint of regularization in order to deal with noise.

Regularization can be seen as a way to deal with numerical problems, due to finite sample sizes leading to inaccurate estimates of the process parameters used for the estimation problem.

Another point of view, and probably more relevant, focuses on the fact that learning is subject to overfitting, if the number of degrees of freedom is too large. A learned hypothesis can only be generalizing towards new examples, if the hypothesis space is small enough, so that each hypothesis can be falsified easily enough. This is often achieved by imposing a Bayesian prior on the solution, thus reducing the effective number of degrees of freedom.

A third way to approach it, is by considering the observed data as data corrupted by noise, and performing the estimation problem in a robust way, so that the influence of noise is as small as possible in a specific way. In a first section, we will show how this viewpoint gives rise to a natural interpretation of the regularization of LSR towards RR.

Given this interpretation of RR, elucidating the parallel between LSR and CCA will lead to a new interpretation of RCCA as described in [5], [6] and [7]. This is done in a following section.

A fundamental property of the approach we adopt, is the fact that we assume an underlying generative model for the data. However, this does not restrict the applicability of CCA, but rather gives an alternative interpretation, complementary to the usual interpretation in geometrical terms like correlation properties.

General notation. *Matrices* will be denoted by capital letters. Row and column *vectors* are represented by lower case letters. Indexed lower case letters represent columns of the capital case equivalent. Capital greek letters are *diagonal matrices*. Indexed lower case greek letters represent the *diagonal elements* of their capital case equivalent. With I, we denote an identity matrix.

2. FROM LEAST SQUARES REGRESSION TO RIDGE REGRESSION

Based on a very simple linear model, with two noise sources, we will review how it is possible to derive the least squares estimator as a maximum likelihood estimator (and thus the maximizer of the log likelihood). Subsequently, we will show how RR [1] naturally follows as a maximizer of the *expected* log likelihood. The expectation is carried out over all possible values of a noise source. An interpretation of the results will be provided.

Notation. $X, D \in \mathbf{R}^{d\times k}$ contain k d-dimensional samples x_i, d_i. The row vectors $y, n \in \mathbf{R}^{1\times k}$ contain k samples of the scalars y_i, n_i. The vector $w \in \mathbf{R}^{d\times 1}$ is a d-dimensional weight vector. The sample covariance matrices C_X and C_{Xy} are defined as $C_X = \frac{XX^T}{k}$ and $C_{Xy} = \frac{Xy^T}{k}$.

2.1. Least squares as a maximum likelihood estimator

We will briefly restate the following well known result on linear regression:

Theorem 1 *The maximum likelihood estimator of w given X and y and the model*

$$y = w^T X + n \tag{1}$$

where n is gaussian noise with zero mean and variance $E\{\frac{nn^T}{k}\} = \sigma^2$, is given by the least squares estimator

$$w_{LS} = C_X^{-1} C_{Xy} \tag{2}$$

Proof. The probability to observe the data, given w, is

$$\begin{aligned} P(X, y|w, \sigma^2) &\propto P(n|\sigma^2) \\ &\propto \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y - w^T X)(y - w^T X)^T}{2\sigma^2}\right] \end{aligned}$$

www.esat.kuleuven.ac.be/sista-cosic-docarch/

In order to maximize this with respect to w, we can minimize minus two times the log likelihood as well. Differentiating this with respect to w and equating to zero leads the optimality conditions on w (often called normal equations). Since XX^T is positive definite, the optimum w_{LS} corresponds to the maximum of the likelihood:

$$2XX^T w_{LS} - 2Xy^T = 0$$

and thus

$$w_{LS} = (XX^T)^{-1}(Xy^T) = C_X^{-1} C_{Xy}$$

This completes the proof.

2.2. Ridge regression as the maximizer of the expected log likelihood

In the model underlying the LSR solution, we included noise on y. But of course, in practice, the measurements X are not noise free either. If we do include it, the model becomes

$$y = w^T(X - D) + n \tag{3}$$

where D is the noise on the measurements X. Suppose we know however the covariance matrix of D, assuming furthermore that D is gaussianly distributed:

$$\begin{aligned} \frac{E\{DD^T\}}{k} &= \lambda^2 I \\ P(D) &= \frac{1}{\sqrt{2\pi\lambda^2}} \exp\left[-\mathrm{tr}\left(\frac{DD^T}{2\lambda^2}\right)\right] \end{aligned}$$

Minus two times the log likelihood, is then equal to (up to a constant term):

$$\begin{aligned} -2l(X, y|w, \sigma^2, D) &= -2\log(P(X, y|w, \sigma^2, D)) \\ &= \frac{\left(y - w^T(X - D)\right)\left(y - w^T(X - D)\right)^T}{2\sigma^2} \end{aligned}$$

The maximizer of the likelihood is equal to the minimizer of this quantity. Note that $-2l(X, y|w, \sigma^2, D)$ is equal to the square loss corresponding to the weight vector w.

Since we don't know the noise D, this minimization of the squared loss can not be carried out in an exact way. However, we are able to average the log likelihood with respect to D. This leads to

$$\begin{aligned} & L(X, y|w, \sigma^2) \qquad (4) \\ = \; & E_D\{-2l(X, y|w, \sigma^2, D)\} \\ = \; & \int_D \frac{\left(y - w^T(X - D)\right)\left(y - w^T(X - D)\right)^T}{2\sigma^2} P(D) dD \\ \propto \; & yy^T + w^T XX^T w + \lambda^2 w^T w - 2yX^T w \\ \propto \; & C_y - 2w^T C_{Xy} + w^T(C_X + \lambda^2 I)w \end{aligned}$$

Differentiating this equation with respect to w and equating to zero, leads to the optimal value for w_{RR}

$$w_{RR} = (C_X + \lambda^2 I)^{-1} C_{Xy} \tag{5}$$

Note again that $C_X + \lambda^2 I$ is positive definite, so, the optimality conditions correspond to a minimum of $L(X, y|w, \sigma^2)$. We have thus proven the following:

Theorem 2 *The estimator of w that maximizes the expected log likelihood of the data, given model (3), is given by the ridge regression estimator $w_{RR} = (C_X + \lambda^2 I)^{-1} C_{XY}$.*

2.3. Interpretation

As a first remark, we note that it is well known that by application of Jensen's inequality [3], the expected log likelihood is proved to be never larger than the log likelihood of the expectation:

$$\begin{aligned} L(X, y|w, \sigma^2) &\leq \log\left(E_D\{P(X, y|w, \sigma^2, D)\}\right) \\ &= \log\left(P(X, y|w, \sigma^2)\right) \end{aligned}$$

Optimizing this last quantity would lead to the maximum likelihood estimator of w in the presence of noise. Thus in fact, we are *optimizing a lower bound on the log likelihood.*

For reference, in appendix we add a derivation of the maximum likelihood estimator for w given the proposed model. The calculations are much more cumbersome. However, it is clear that it is not guaranteed that the maximum likelihood estimator will yield the best performance, where performance is measured in terms of expected squared error (the quantity optimized by the ridge regression estimator).

Therefore, as a second remark, note that the sample expectation over the (X, y) samples and the expectation over D (as given in equation (4)) of the squared error is the best estimate for the expected squared error over the distributions over X, y and D. (This is due to a standard property of the sample mean of a random variable, namely that it is the best estimate of the population mean. Furthermore, the law of large numbers guarantees convergence. Note however we assume that X, y and D are iid distributed and that there is no prior on the distributions of X and y.) Therefore, the RR estimator is the *minimum least squares error estimator* for this type of linear systems with noise.

As a third remark, we would like to point out that it is a standard result that maximum likelihood estimators are unbiased. Since the RR estimator is not a maximum likelihood estimator, we can conclude that it is *a biased estimator.*

2.4. Practical aspects

In general, λ will not be known. The standard approach to estimating this *hyperparameter* is applying cross validation. The cross validation score is then the squared error of the test set, evaluated for a w obtained using a training set.

So far our treatment of RR. Note that these ideas are no more than a reformulation of ideas existing in literature. We provided them for reference, and in order to motivate the development of thoughts in the next section on CCA.

3. FROM CANONICAL CORRELATION ANALYSIS TO ITS REGULARIZED VERSION

In contrast to regression, CCA should be catalogued under the denominator of multivariate statistics. While in regression, one searches for a direction in the X space that predicts the y space as good as possible, in doing CCA, one searches for a direction in the X space and one in the Y space (that is multidimensional as well) so that order the correlation between the projections on these directions is maximal. One can still go further, and look for directions

uncorrelated with the previous ones, that, under this additional constraint, maximize the correlation between the X and the Y space. And so on.

The classical approach to CCA is using these geometrical arguments [2]. However, in this paper, we will approach CCA as an estimator of a linear system underlying the data X and Y.

For conciseness and ease of notation, we will assume that the dimensions of X and Y are both equal to d. However, most of the results are easy to extend towards different dimensions for X and Y.

Notation. In contrast to the previous section, we will not work with a one-dimensional y, but with $Y \in \mathbf{R}^{d\times k}$. Also, $C \in \mathbf{R}^{d\times k}$. The sample covariance matrices C_Y and $C_{XY} = C_{YX}^T$ are defined as $C_Y = \frac{YY^T}{k}$ and $C_{XY} = \frac{XY^T}{k}$. $M_X, M_Y \in \mathbf{R}^{d\times d}$ are square mixing matrices. Furthermore, $W_X = M_X^{-T}$ and $W_Y = M_Y^{-T}$. These matrices contain d-dimensional weight vectors $w_{X,i}$ and $w_{Y,i}$ in their columns. Analogously, $v_{X,i}$ and $v_{Y,i}$ are the d-dimensional columns of $V_X, V_Y \in \mathbf{R}^{d\times k}$. The diagonal elements of the diagonal matrices $\Xi, \Sigma, \Lambda_X, \Lambda_Y \in \mathbf{R}^{d\times d}$ are $\xi_i, \sigma_i, \lambda_{X,i}$ and $\lambda_{Y,i}$.

Most of the results can be generalized towards different dimensions $d_X \times k$ and $d_Y \times k$ for X and Y, however, due to space limitations we will not do this in this paper.

3.1. Standard geometrical approach to canonical correlation analysis

CCA solves the following optimization problem:

$$\begin{aligned} \xi_i = \quad & \max_{v_{X,i}, v_{Y,j}, \forall i,j} \quad \frac{v_{X,i}^T X Y^T v_{Y,i}}{\sqrt{v_{X,i}^T X X^T v_{X,i} v_{Y,i}^T Y Y^T v_{Y,i}}} \qquad (6) \\ & \text{s.t. } \forall j < i \qquad v_{X,i}^T X X^T v_{X,j} = 0 \\ & \qquad\qquad\quad v_{Y,j}^T Y Y^T v_{Y,i} = 0 \end{aligned}$$

where $v_{X,i}, v_{Y,i} \in \mathbf{R}^{d\times 1}$ and $i = 1, \ldots, d$.

One can show this problem reduces to solving the following generalized eigenvalue problem:

$$\begin{aligned} & \begin{pmatrix} 0 & C_{XY} \\ C_{YX} & 0 \end{pmatrix} \cdot \begin{pmatrix} v_{X,i} \\ v_{Y,i} \end{pmatrix} \qquad (7) \\ & = \xi_i \begin{pmatrix} C_X & 0 \\ 0 & C_Y \end{pmatrix} \cdot \begin{pmatrix} v_{X,i} \\ v_{Y,i} \end{pmatrix} \end{aligned}$$

The generalized eigenvectors $v_{X,i}$ with $v_{Y,i}$ are the corresponding canonical components in both spaces, and ξ_i is the canonical correlation corresponding to $v_{X,i}$ and $v_{Y,i}$.

3.2. Canonical correlation analysis as a maximum likelihood estimator

Now, assume the following model underlying the data

$$\begin{aligned} X &= M_X(C + N_X) \qquad (8) \\ Y &= M_Y(C + N_Y) \end{aligned}$$

and thus

$$\begin{aligned} W_X^T X &= C + N_X \qquad (9) \\ W_Y^T Y &= C + N_Y \end{aligned}$$

where we assume that N_X and N_Y are gaussianly distributed with covariance matrices

$$\frac{E\{N_X N_X^T\}}{k} = \quad \Sigma^2 \quad = \frac{E\{N_Y N_Y^T\}}{k}$$

where for each $i < j$, $\sigma_i > \sigma_j > 0$. Furthermore, without loss of generality, we assume that the covariance matrix of C is equal to the identity.

We are now ready to state the following theorem:

Theorem 3 *The generalized eigenvectors V_X and V_Y given by the generalized eigenvalue problem solved by CCA (if properly normalized) and $(\Xi^{-1} - 1)$ where ξ_i are the canonical correlations, correspond to a stationary point of the likelihood function with variables W_X, W_Y and Σ^2 respectively, parameters in the model (8), given the data X and Y.*

That this stationary point corresponds to a maximum, ie that the CCA solution is a maximum likelihood estimator of the parameters of (8), will be left as a conjecture in this paper.

Proof. The probability of X and Y given W_X, W_Y, Σ^2 and C is given by

$$\begin{aligned} & P(X, Y | W_X, W_Y, \Sigma^2, C) \\ = \quad & P(X | W_X, \Sigma^2, C) P(Y | W_Y, \Sigma^2, C) \end{aligned}$$

where

$$\begin{aligned} & P(X | W_X, \Sigma^2, C) \qquad (10) \\ = \quad & \frac{\exp\Big(-\frac{1}{2}\mathrm{tr}\big[(W_X^T X - C)\Sigma^{-2}(W_X^T X - C)^T\big]\Big)}{\sqrt{2\pi \det(W_X \Sigma^{-2} W_X^T)^{-1}}} \\ \text{and} \quad & P(Y | W_Y, \Sigma^2, C) \qquad (11) \\ = \quad & \frac{\exp\Big(-\frac{1}{2}\mathrm{tr}\big[(W_Y^T Y - C)\Sigma^{-2}(W_Y^T Y - C)^T\big]\Big)}{\sqrt{2\pi \det(W_Y \Sigma^{-2} W_Y^T)^{-1}}} \end{aligned}$$

After some tedious but straightforward calculations, one can thus show that the evidence $P(X, Y | W_X, W_Y, \Sigma^2)$ is equal to

$$\begin{aligned} & P(X, Y | W_X, W_Y, \Sigma^2) \qquad (12) \\ = \quad & \int_C P(X, Y | W_X, W_Y, \Sigma^2, C) P(C) dC \\ \propto \quad & \sqrt{\det(I + 2\Sigma^{-2})} \cdot \\ & \sqrt{\det((W_X \Sigma^{-2} W_X^T)(W_Y \Sigma^{-2} W_Y^T))} \cdot \\ & \exp\Bigg(-\frac{1}{2}\mathrm{tr}\Big[(W_X^T X)^T \Sigma^{-2} (W_X^T X) \\ & + (W_Y^T Y)^T \Sigma^{-2} (W_Y^T Y) \\ & - \Sigma^{-2}(W_X^T X + W_Y^T Y)^T \cdot (I + 2\Sigma^{-2})^{-1} \cdot \\ & (W_X^T X + W_Y^T Y)\Sigma^{-2}\Big]\Bigg) \end{aligned}$$

In order to maximize the likelihood, we can as well minimize minus 2 times the log likelihood:

$$-2l(X, Y | W_X, W_Y, \Sigma^2) \qquad (13)$$

$$
\begin{aligned}
= \quad & -\log\det(I+2\Sigma^{-2}) \\
& -\log\det((W_X\Sigma^{-2}W_X^T)(W_Y\Sigma^{-2}W_Y^T)) \\
& +\mathrm{tr}\Big[(W_X^TX)^T\Sigma^{-2}(W_X^TX) \\
& +(W_Y^TY)^T\Sigma^{-2}(W_Y^TY) \\
& -(W_X^TX+W_Y^TY)^T\cdot\Sigma^{-2}(I+2\Sigma^{-2})^{-1}\Sigma^{-2}\cdot \\
& (W_X^TX+W_Y^TY)\Big] + \text{a constant}
\end{aligned}
$$

Differentiating this with respect to the matrix W_X and equating to zero, gives

$$
\begin{aligned}
& -2W_X^{-T} \\
& +2C_XW_X[\Sigma^{-2}-\Sigma^{-2}(I+2\Sigma^{-2})^{-1}\Sigma^{-2}] \\
& -2C_{XY}W_Y[\Sigma^{-2}(I+2\Sigma^{-2})^{-1}\Sigma^{-2}] \\
= \quad & 0
\end{aligned}
$$

After multiplication on the right with $\Sigma^4(I+2\Sigma^{-2})(I+\Sigma^2)^{-1}$ and on the left with W_X^T, this leads to

$$
\begin{aligned}
& W_X^TC_XW_X - W_XC_{XY}W_Y(I+\Sigma^2)^{-1} \\
= \quad & (I+\Sigma^2)-(I+\Sigma^2)^{-1} \qquad (14)
\end{aligned}
$$

Similarly, we can derive an analogous equation by equating the derivative with respect to W_Y to zero.

$$
\begin{aligned}
& W_Y^TC_YW_Y - W_YC_{YX}W_X(I+\Sigma^2)^{-1} \\
= \quad & (I+\Sigma^2)-(I+\Sigma^2)^{-1} \qquad (15)
\end{aligned}
$$

Furthermore, differentiating this with respect to Σ^{-2} (taking the diagonality of Σ into account), leads to

$$
\begin{aligned}
- \quad & 2(I+2\Sigma^{-2})^{-1}-2\Sigma^2 \qquad (16) \\
+ \quad & \mathrm{diag}[W_X^TC_XW_X+W_Y^TC_YW_Y \\
- \quad & 2(W_X^TC_XW_X+W_Y^TC_YW_Y+W_X^TC_{XY}W_Y \\
& +W_Y^TC_{YX}W_X)\Sigma^{-2}(I+\Sigma^{-2})(I+2\Sigma^{-2})^{-2}] = 0
\end{aligned}
$$

Now, one can see that these equations (14), (15) and (16) hold for the CCA solution, if the columns of V_X and V_Y are properly normalized, so that $V_X^TC_XV_X = I+\Sigma^2$, $V_Y^TC_YV_Y = I+\Sigma^2$ and with $\Xi = (I+\Sigma^2)^{-1}$. We can see this since, if we multiply the CCA generalized eigenvalue equation by $\begin{pmatrix} V_X^T & V_Y^T \end{pmatrix}$ on the left hand side, we obtain:

$$
\begin{aligned}
V_X^TC_{XY}V_Y &= V_X^TC_XV_X\Xi \\
V_Y^TC_{YX}V_X &= V_Y^TC_YV_Y\Xi
\end{aligned}
$$

Filling everything out completes the proof.

3.3. Regularized CCA as the maximizer of the expected log likelihood

In general, however, we will not only encounter noise on the latent variables C, but there will be measurement noise on X and on Y as well. Therefore, we adopt the following model

$$
\begin{aligned}
W_X^T(X-D_X) &= C-N_X \qquad (17) \\
W_Y^T(Y-D_Y) &= C-N_Y
\end{aligned}
$$

We will assume the noise terms D_X and D_Y are gaussianly distributed, with covariance matrices equal to Λ_X^2 and Λ_Y^2.

We are now ready for the main theorem of the paper:

Theorem 4 *Properly normalized V_X and V_Y, and $(\Xi^{-1}-I)$, given by the regularized canonical correlation (RCCA) estimate as defined by the following generalized eigenvalue problem*

$$
\begin{aligned}
& \begin{pmatrix} 0 & C_{XY} \\ C_{YX} & 0 \end{pmatrix} \cdot \begin{pmatrix} V_{X,i} \\ V_{Y,i} \end{pmatrix} \qquad (18) \\
= \xi_i & \begin{pmatrix} C_X+\Lambda_X^2 & 0 \\ 0 & C_Y+\Lambda_Y^2 \end{pmatrix} \cdot \begin{pmatrix} V_{X,i} \\ V_{Y,i} \end{pmatrix}
\end{aligned}
$$

correspond to a stationary point of the expected log likelihood with variables W_X, W_Y and Σ^2 respectively, parameters in the generative model (17). The expectation is carried out over the distributions of D_X and D_Y.

That this stationary point corresponds to a maximum, will be left as a conjecture in this paper.

Proof. The outline of the proof is clear, given the maximum likelihood derivation of ordinary CCA, and the RR derivation. We will thus only state some intermediate results of the proof, for conciseness.

The probability of X and Y given W_X, W_Y, Σ^2, C and D_X and D_Y is given by

$$
\begin{aligned}
& P(X,Y|W_X,W_Y,\Sigma^2,C,D_X,D_Y) \\
= \quad & P(X|W_X,\Sigma^2,C,D_X)P(Y|W_Y,\Sigma^2,C,D_Y) \\
= \quad & P(X-D_X|W_X,\Sigma^2,C,D_X)P(Y-D_Y|W_Y,\Sigma^2,C,D_Y)
\end{aligned}
$$

Each of these factors can be expressed in the same way as equations (10) and (11), where X and Y have to be replaced by $X-D_X$ and $Y-D_Y$.

We can again take the expectation of (19) over C, leading to the analogon to equation (12), with the same replacements. Taking the logarithm multiplied by minus two, and subsequently averaging with respect to D_X and D_Y, leads to minus two times the average log likelihood

$$
\begin{aligned}
& -2L(X,Y|W_X,W_Y,\Sigma^2) \\
= \quad & E_{D_X,D_Y}\{l(X,Y|W_X,W_Y,\Sigma^2,D_X,D_Y)\}
\end{aligned}
$$

to be equal to a sum of two terms, the first one of which is equal equation (13), and the second one of which is

$$
\begin{aligned}
\mathrm{tr} \quad & \Big[\Sigma^{-1}W_X^T\Lambda_X^2W_X\Sigma^{-1} \\
+ \quad & \Sigma^{-1}W_Y^T\Lambda_Y^2W_Y\Sigma^{-1} \\
- \quad & (I+2\Sigma^{-2})^{-1}\Sigma^{-2}W_X^T\Lambda_X^2W_X\Sigma^{-2}(I+2\Sigma^{-2})^{-1} \\
- \quad & (I+2\Sigma^{-2})^{-1}\Sigma^{-2}W_Y^T\Lambda_Y^2W_Y\Sigma^{-2}(I+2\Sigma^{-2})^{-1}\Big]
\end{aligned}
$$

By simply writing out the equations, this can be shown to be equal to (13), with every $C_X = \frac{XX^T}{k}$ and $C_Y = \frac{YY^T}{k}$ replaced by $C_X = \frac{XX^T}{k}+\Lambda_X^2$ and $C_Y = \frac{YY^T}{k}+\Lambda_Y^2$. The cross product terms in $C_{XY} = \frac{XY^T}{k} = C_{YX}^T$ remain unchanged.

Now, we can differentiate this cost function with respect to W_X and W_Y again, leading to optimality conditions that are fulfilled by the regularized CCA solution, given that W_X and W_Y are normalized so that $W_X^T(C_X+\Lambda_X^2)W_X = I+\Sigma^2$, $W_Y^T(C_Y+\Lambda_Y^2)W_Y = I+\Sigma^2$ and with $\Xi = (I+\Sigma^2)^{-1}$.

This completes the outline of the proof.

Note that it is straightforward to extend the theorem to general covariance matrices for D_X and D_Y.

3.4. Interpretation

We can interpret this in a similar way as RR. The expected log likelihood is a lower bound of the log of the expectation. Therefore, the maximum of the expected log likelihood is a lower bound on the likelihood, D_X and D_Y taken into account.

Again, this does not represent an unbiased estimator, since it is not the maximum likelihood estimator. Neither is it the least squares estimator, as it was in the RR case. However, instead of least squares, another measure is appropriate, namely the log likelihood (in the RR case, least squares was equivalent with log likelihood).

3.5. Practical aspects

The result is mainly theoretical. However, there is an important practical aspect.

In general, the noise covariance matrices Λ_X and Λ_Y are not known. Thus, they have to be estimated using cross validation, or by some other method. If we use cross validation, we need a cost function to be optimized. For this, based on the analogy with RR, it has now been made acceptable to use the log likelihood given the data (equation (13)). Thus, for each of the regularization parameters, we solve the generalized eigenvalue problem making use of the training set. The regularization parameter for which the log likelihood of the test set is maximal, can be taken as the estimate for the noise covariance.

4. CONCLUSIONS

LSR and RR are well established regression methods. By analogy to a derivation of LSR, we propose an interpretation of CCA in terms of a maximum likelihood estimator. Extending the result along the same lines as the extension of LSR towards RR, we derived a regularized version CCA, that is around for quite some time already, however, was not entirely understood.

Apart from theoretical results mostly concerning the interpretation of regularization of CCA, we made clear how to train the regularization parameter using cross validation. This was not known before, and at hoc techniques were applied up till now.

An important fact that is not pointed out yet, is the similarity between the model underlying CCA, and the independent component analysis (ICA) model (for an introduction to ICA, see [8]). Where the identification for ICA is possible thanks to supposed independencies among the components of C (and this is basically exploited using higher order information), CCA only uses second order information to identify essentially the same model. An important consequence of this is that CCA and the regularized version are easily kernelizable and can thus be made nonlinear ([7], [13], [12], [11]; for an introduction to kernel methods, see [9]).

5. ACKNOWLEDGEMENTS

Tijl De Bie is a Research Assistant with the Fund for Scientific Research - Flanders (F.W.O.-Vlaanderen). Dr. Bart De Moor is a full professor at the Katholieke Universiteit Leuven, Belgium. Our research is supported by:

Research Council KUL: GOA-Mefisto 666, IDO (IOTA Oncology, Genetic networks);

Flemish Government: *FWO*: projects G.0115.01 (microarrays / oncology), G.0407.02 (support vector machines), G.0240.99 (multilinear algebra), G.0413.03 (inference in bioi), G.0388.03 (microarrays for clinical use), research communities (ICCoS, ANMMM); *IWT*: STWW-Genprom (gene promotor prediction), GBOU-McKnow (Knowledge management algorithms), GBOU-ANA (biosensors);

Belgian Federal Government: DWTC (IUAP IV-02 (1996-2001) and IUAP V-22 (2002-2006);

EU: CAGE; ERNSI;

Furthermore, sincere thanks go to Lieven De Lathauwer, Nello Cristianini and Michael Jordan for very helpful discussions and suggestions. TDB would like to thank Laurent El Ghaoui and Gert Lanckriet for their hospitality during a very enriching stay in the SALSA group, EECS Department, U.C.Berkeley, where part of this work has been completed.

6. REFERENCES

[1] A.E. Hoerl and R.W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems", Technometrics, 12(3):55-67, 1970.

[2] H. Hotelling, "Relations between two sets of variates", Biometrica, 28, 321-377, 1935.

[3] T.M. Cover, J.A. Thomas, "Elements of information theory", Wiley-interscience, 1991.

[4] A.N. Tikhonov and V.Y. Arsenin, "Solutions of Ill-Posed Problems", Winston, Washington, 1977.

[5] H.D. Vinod, "Canonical Ridge and Econometrics of Joint Production", J. Econometrics, 4:147-166, 1976.

[6] F.A. Nielsen, L.K. Hansen, S.C. Strother, "Canonical ridge analysis with ridge parameter optimization", NeuroImage 7: S758, 1998.

[7] F. Bach, M. Jordan, "Kernel Independent component Analysis", Journal of Machine Learning Research, 3, 1-48, 2002.

[8] A. Hyvärinen, J. Karhunen, E. Oja, "Independent Component Analysis", Wiley-interscience, 2001.

[9] N. Cristianini, J. Shawe-Taylor, "An Introduction to Support Vector Machines", Cambridge University Press, 2000.

[10] B. De Moor, J. David,"Total least squares and the algebraic Riccati equation", Systems and Control Letters, vol.18, no.5, May 1992, pp. 329-337.

[11] T. Van Gestel, "From linear to kernel based methods in classification, modelling and prediction", PhD thesis, Faculty of Engineering, K.U.Leuven (Leuven, Belgium), May, 2002, 286 p. 01-24

[12] T. Van Gestel, J. Suykens, J. De Brabanter, B. De Moor, J. Vandewalle, "Kernel canonical correlation analysis and least squares support vector machines", in Proc. of the International Conference on Artificial Neural Networks (ICANN 2001), Vienna, Austria, Aug. 2001, pp. 381-386.

[13] J.A.K. Suykens, T. Van Gestel, J. Vandewalle, B. De Moor, "A support vector machine formulation to PCA analysis and its kernel version", Internal Report 02-68, ESAT-SISTA, K.U.Leuven (Leuven, Belgium), 2002. Accepted for publication in IEEE Transactions on Neural Networks.

7. APPENDIX - A MAXIMUM LIKELIHOOD ESTIMATOR FOR THE RIDGE REGRESSION MODEL

RR is the maximizer of the expected log likelihood, if the identity matrix multiplied by a constant, added to C_X is equal to the noise covariance.

However, it is possible to write an analytical solution for the maximum likelihood estimator for w. Interesting to note is that the solution takes the same form, however, the identity matrix added to C_X then is multiplied with a different constant, which can be negative or positive.

Theorem 5 *The maximum likelihood estimator w_{ML} for w, given model 3, is equal to*

$$w_{ML} = (C_X + \gamma I)^{-1} C_{Xy} \tag{19}$$

where

$$\gamma \quad = \quad -\lambda^2 \cdot \frac{1}{k}\left(\frac{(y - w_{ML}^T X)(y - w_{ML}^T X)^T}{\sigma^2 + \lambda^2 w_{ML}^T w_{ML}} - 1\right) \tag{20}$$

Important to note is that this equation is implicit. An iterative procedure to compute w_{ML} is however suggested by the theorem. In this paper, we will not go into the convergence properties of this iteration. Furthermore, we will only prove that the given solution represents a stationary point of the likelihood, and leave as a conjecture that this stationary point corresponds to a maximum.

Proof. We want to maximize the probability of the data X and y given the parameters w, γ, σ, with respect to w.

$$\begin{aligned}
& P(X, y|w, \gamma, \sigma) \\
= & \int_D P(X, y|w, \gamma, \sigma, D) P(D|w, \gamma, \sigma) dD \\
= & \int_D \exp\big(-\frac{1}{2\sigma^2}[y - w^T(X - D)][y - w^T(X - D)]^T\big) \\
& \cdot \frac{1}{\sqrt{(2\pi\sigma^2)^n}} \cdot \exp\big(-\frac{1}{2\lambda^2}\mathrm{tr}[DD^T]\big) \cdot \frac{1}{\sqrt{(2\pi\lambda^2)^n}} dD \\
\propto & \frac{1}{\sqrt{\det(I\sigma^2 + ww^T\lambda^2)}} \cdot \\
& \cdot \exp\left[-\frac{1}{2}\mathrm{tr}\left(\frac{(y - w^T X)^T (y - w^T X)}{\sigma^2}\right.\right. \\
& \left.\left. -\frac{(w(y - w^T X))^T}{\sigma^2}\Big(\frac{I}{\lambda^2} + \frac{ww^T}{\sigma^2}\Big)^{-1}\frac{(w(y - w^T X))}{\sigma^2}\right)\right]
\end{aligned}$$

Taking the logarithm and multiplying with -2, leads to

$$\begin{aligned}
& \log\det(I\sigma^2 + ww^T\lambda^2) + \frac{(y - w^T X)(y - w^T X)^T}{\sigma^2} \\
- & \quad \mathrm{tr}\Big[\frac{(w(y - w^T X))^T}{\sigma^2}(\frac{I}{\lambda^2} + \frac{ww^T}{\sigma^2})^{-1}\frac{(w(y - w^T X))}{\sigma^2}\Big]
\end{aligned}$$

We will derive this with respect to w, and equate it to 0, in order to find the optimum. Thereby, we will use the following equality (due to the matrix inversion lemma):

$$w^T\left(\frac{I}{\lambda^2} + \frac{ww^T}{\sigma^2}\right)^{-1} w \quad = \quad \frac{\lambda^2\sigma^2 w^T w}{\lambda^2 w^T w + \sigma^2}$$

the derivative with respect to w of which is

$$\frac{2w\lambda^2\sigma^4}{(\sigma^2 + \lambda^2 w^T w)^2}$$

Furthermore, we need the derivative of the logdet term with respect to w, it turns out to be equal to

$$2\left(\frac{I}{\lambda^2} + \frac{ww^T}{\sigma^2}\right)^{-1} w \quad = \quad \frac{2w\lambda^2}{\sigma^2 + \lambda^2 w^T w} \tag{21}$$

Combining all these results leads to the proof of the theorem.

Although this theorem provides only an implicit formula for w_{ML}, it allows some intuitive interpretation: $\gamma \neq 0$ leads to a correction on C_X that has the following properties:

- for $\sigma \to \infty$, all the noise will be attributed to n, since this will have a negligible effect on the likelihood as compared to noise attributed to D. The result is that $\gamma \to \frac{\lambda^2}{k}$, meaning that for large samples size k, it is assumed that the estimate of C_X is probably correct (since the noise $D = 0$).
- for $\sigma \to 0$ on the other hand, all the noise will be attributed to the high variance noise D, since this will lead to the highest likelihood. The result is that $\gamma \to \frac{(y - w_{ML}^T X)(y - w_{ML}^T X)^T}{k w_{ML}^T w_{ML}} - \frac{\lambda^2}{k}$. The second term is unimportant in the sense that it quickly disappears as the sample size increases. The first term has the squared error in the numerator. This will be assumed to be due to D only, since this doesn't lower the likelihood. Thus, γ converges to a sample estimate of $-\lambda^2$, namely to $-\frac{w_{ML}^T \frac{DD^T}{k} w_{ML}}{w_{ML}^T w_{ML}}$. By subtracting this estimated sample noise covariance from C_X, one removes the noise influence in an optimal way.
- for intermediate values of σ, the noise can be seen as attributed proportionally to both noise terms. The parameter γ is then an estimate of minus the variance λ^2 of the noise in this sample that is due to D.

Note the similarities with total least squares (TLS) [10]: asymptotically for large sample sizes k, TLS yields the same solution, since the sample estimate of λ^2 will converge to its true value. In TLS, one always takes $\gamma = -\lambda^2$. In other words: whereas TLS assumes that λ^2 is the sample covariance of D, the ML estimator given in this paper provides an optimal estimate for this sample covariance, given the data and the population covariances.

A simple analogous derivation of a maximum likelihood estimator of the model (17) is clearly not possible. However, note that we do not necessarily want a maximum likelihood estimate. Most of the time, we are not interested in estimating the weight vector w, but rather in a w that most likely gives good estimates for y given yet unobserved samples X. This is precisely what RR does.