A Weighted Logarithmic Merit Function for Canonical Correlation Analysis

Mohammed A. Hasan
Department of Electrical & Computer Engineering
University of Minnesota Duluth
E.mail:mhasan@d.umn.edu

Abstract: *A weighted logarithmic merit function that incorporates a diagonal matrix is utilized for deriving a gradient dynamical system that converges to the actual canonical correlation coordinates of arbitrary data matrices. The equilibrium points of the resulting gradient system are determined and their stability is thoroughly analyzed. Qualitative properties of the proposed systems are analyzed in detail including the limit of solutions as time approaches infinity. The performance of this system is also examined.*

Keywords: Canonical correlation analysis(CCA), SVD, gradient dynamical system, unconstrained optimization, matrix calculus

1 Introduction

Canonical correlation analysis (CCA) is a multivariate statistical technique that has been widely used in many modern information processing fields, such as text mining, statistical and medical signal/image processing, facial expression recognition, and communication theory. It is originally developed in [1]-[2]. The applications of CCA require effective estimation of the singular vectors of the coherence matrix of a pair of multivariate information sources. Although the singular value decomposition (SVD) of the coherence matrix can be used to implement the CCA, the conventional matrix algebraic approach is often unsuitable for a higher dimensional data as well as for the adaptive CCA due to the very high computational load.

The CCA technique is described in most standard textbooks on multivariate statistics, e.g., [3,4,5]. Serial and parallel algorithms for CCA have been proposed in [6]. Work on nonlinear canonical correlations analysis is dealt with in [7]-[8]. CCA has been generalized in several directions. For example, multiset CCA is proposed in [9]-[10] by maximizing some generalized measure of correlation.

The following notation will be used throughout. The symbol $\mathbb{R}$ denotes the set of real numbers. The transpose of a real matrix x is denoted by x^T, and the derivative of x with respect to time is written as x'. If B is a square matrix, then $tr(B)$ and $det(B)$ denote the trace and determinant of B, respectively. The identity matrix of appropriate dimension is expressed with the symbol I. The gradient and the Hessian matrix of a function f are denoted by ∇f and $H(f)$, respectively.

2 Background and Preliminaries

Given two multivariate data sets X and Y, let $R_{xx} \in \mathbb{R}^{m\times m}$, and $R_{yy} \in \mathbb{R}^{n\times n}$, be an auto-covariance of x and y, respectively. Let $R_{xy} \in \mathbb{R}^{m\times n}$, be a cross-covariance between x and y. The conventional CCA consists of performing the singular value decomposition of the coherence matrix $\hat{C}$ [12] defined as

$$\hat{C} = R_{xx}^{-\frac{1}{2}} R_{xy} R_{yy}^{-\frac{1}{2}}, \tag{1}$$

where $R_{xx}^{\frac{1}{2}}$ and $R_{yy}^{\frac{1}{2}}$ are any symmetric square roots of R_{xx} and R_{yy}, respectively. Assume that the SVD of $\hat{C}$ is

$$\hat{C} = u\Sigma v^T + u_2\Sigma_2 v_2^T, \tag{2}$$

where $\Sigma = \text{diag}(\sigma_1, \cdots, \sigma_p)$ and $\Sigma_2 = \text{diag}(\sigma_{p+1}, \cdots, \sigma_n)$ are diagonal matrices so that $\sigma_i > \sigma_j$ for $i = 1, \cdots, p$ and $j = p+1, \cdots, n$. The matrices $u, v \in \mathbb{R}^{n\times p}$ and $u_2, v_2 \in \mathbb{R}^{n\times n-p}$ are orthogonal, i.e., $u^T u = I, v^T v = I$ and $u_2^T u_2 = I, v_2^T v_2 = I$, $u^T u_2 = 0$, $v^T v_2 = 0$. It can be easily verified that the matrix

$$U = \frac{1}{\sqrt{2}} \begin{bmatrix} u & -u \\ v & v \end{bmatrix}, \tag{3a}$$

is orthogonal, i.e., $U^T U = I$, and that

$$U^T \hat{C} U = \begin{bmatrix} \Sigma & 0 \\ 0 & -\Sigma \end{bmatrix}, \tag{3b}$$

where

$$\bar{\hat{C}} = \begin{bmatrix} 0 & \hat{C} \\ \hat{C}^T & 0 \end{bmatrix}. \tag{4a}$$

Thus $\bar{\hat{C}}$ can be expressed as

$$\bar{\hat{C}} = U\bar{\Sigma}U^T + U_2\bar{\Sigma}_2 U_2^T, \tag{4b}$$

where

$$\begin{aligned} \bar{\Sigma} = \begin{bmatrix} 0 & \Sigma \\ \Sigma & 0 \end{bmatrix}, \bar{\Sigma}_2 = \begin{bmatrix} 0 & \Sigma_2 \\ \Sigma_2 & 0 \end{bmatrix}, \\ U_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} u_2 & -u_2 \\ v_2 & v_2 \end{bmatrix}. \end{aligned} \tag{4c}$$

Note that U_2 is orthogonal, i.e., $U_2^T U_2 = I$.

2.1 First and Second Order Differentials

Let $F : \mathbb{R}^{n\times p} \to \mathbb{R}$ be twice continously differentiable function, the first and second order differentials of F are defined by

$$dF(x) = \frac{dF(x + \epsilon dx)}{d\epsilon}|_{\epsilon=0}, \tag{5a}$$

and

$$d^2F(x) = \frac{d^2F(x + \epsilon dx)}{d\epsilon^2}|_{\epsilon=0}. \tag{5b}$$

To compute the gradient and the Hessian matrix for a merit function F, the first and second order differentials need to be derived first. In the next result, the first and second order differentials for linear, quadratic, quartic, and logarithmic functions are computed.

Proposition 1. *Let $E \in \mathbb{R}^{n\times n}$, $b \in \mathbb{R}^{n\times p}$, $D \in \mathbb{R}^{p\times p}$, where D is diagonal matrix, and consider the functions F_1, F_2, F_3, F_4 defined over $\mathbb{R}^{n\times p}$ by*

$$\begin{aligned} F_1(z) &= tr(b^T z), \\ F_2(z) &= tr(z^T E z D), \\ F_3(z) &= tr((z^T E z)^2), \\ F_4(z) &= tr\{D \log(z^T E z)\}. \end{aligned} \tag{6a}$$

Then the first and second order differentials of F_1, F_2, F_3, and F_4 are given by:

$$dF_1(z) = tr(b^T dz),\ d^2F_1 = 0, \tag{6b}$$

$$\begin{aligned} dF_2 &= tr\{dz^T E z D + z^T E dz D)\}, \\ d^2F_2(z) &= tr\{2dz^T E dz D\}, \end{aligned} \tag{7a}$$

$$\begin{aligned} dF_3 &= 4tr\{dz^T E z z^T E z\}, \\ d^2F_3(z) &= 4tr\{dz^T E dz z^T E z + dz^T E z dz^T E z \\ &\quad + dz^T E z z^T E dz)\}, \end{aligned} \tag{7b}$$

$$dF_4(z) = tr\{D(z^T E z)^{-1}(dz^T E z + z^T E dz)\}, \tag{7c}$$

and

$$\begin{aligned} d^2F_4 &= 2tr\{D(z^T E z)^{-1} dz^T E dz\} \\ &- tr\{D(z^T E^T z)^{-1} z^T E^T dz (z^T E^T z)^{-1} z^T E^T dz \\ &- 2tr\{D(z^T E z)^{-1} dz^T E z (z^T E z)^{-1} z^T E dz \\ &- tr\{D(z^T E z)^{-1} z^T E dz (z^T E z)^{-1} z^T E dz. \end{aligned} \tag{7d}$$

Therefore, the gradient and the Hessian matrix of $F_i, i = 1, 2, 3, 4$, are

$$\begin{aligned} \nabla F_1 &= 0, H(F_1) = 0 \\ \nabla F_2 &= (E + E^T) z D \\ H F_2 &= D \otimes (E + E^T) \\ \nabla F_3 &= 2Ez(z^T E z) + 2E^T z (z^T E^T z) \\ H(F_3) &= 4I \otimes E z z^T E + 4I \otimes E^T z z^T E^T + 4KEz \otimes z^T E^T \\ \nabla F_4 &= EzD(z^T E z)^{-1} + E^T z (z^T E^T z)^{-1} D, \end{aligned} \tag{8a}$$

$$\begin{aligned} H(F_4) &= \frac{1}{2}(z^T E^T z)^{-1} D \otimes E + D(z^T E z)^{-1} \otimes E^T \\ &- (z^T E^T z)^{-1} D \otimes E z (z^T E z)^{-1} z^T E \\ &- D(z^T E z)^{-1} \otimes E^T z (z^T E^T z)^{-1} z^T E^T \\ &- \frac{1}{2} K E z (z^T E z)^{-1} D \otimes (z^T E^T z)^{-1} z^T E^T \\ &- \frac{1}{2} K E z (z^T E z)^{-1} \otimes D (z^T E^T z)^{-1} z^T E^T \\ &- \frac{1}{2} K E^T z (z^T E^T z)^{-1} D \otimes (z^T E z)^{-1} z^T E \\ &- \frac{1}{2} K E^T z (z^T E^T z)^{-1} \otimes D (z^T E z)^{-1} z^T E \end{aligned} \tag{8b}$$

for some permutation matrix K.

Proof: The proof is a direct application of the definitions (5a), (5b), and Lemma 3 (see Appendix).

Proposition 2. *Let $A \in \mathbb{R}^{n\times m}$, $x \in \mathbb{R}^{n\times p}$, $y \in \mathbb{R}^{m\times p}$, $D \in \mathbb{R}^{p\times p}$, D is diagonal, and consider the function defined by*

$$F_5(x, y) = tr\{D \log(x^T A y)\}. \tag{9a}$$

Then the differentials of F_5 with respect to x and y are

$$\begin{aligned} d_x F_5(x, y) &= tr\{D(x^T A y)^{-1}(dx^T A y) \\ &\quad + D(y^T A^T x)^{-1}(y^T A^T dx)\}, \end{aligned} \tag{9b}$$

$$\begin{aligned} d_y F_5(x, y) &= tr\{D(x^T A y)^{-1}(x^T A dy) \\ &\quad + D(y^T A^T x)^{-1}(dy^T A^T dx)\}, \end{aligned} \tag{9c}$$

and hence

$$\begin{aligned} \nabla_x F_5 &= AyD(x^T A y)^{-1} + Ay(x^T A y)^{-1} D \\ \nabla_y F_5 &= A^T Dx(y^T A^T x)^{-1} + A^T x(y^T A^T x)^{-1} D. \end{aligned} \tag{9d}$$

Proof: The proof is a direct application of the definitions (5a), (5b), and the trace identity

$$tr(DW) = \frac{1}{2} tr(DW + DW^T),$$

where D and W are any two square matrices of same dimensions and D is diagonal.

3 A Weighted Logarithmic merit Function

In this section, a gradient dynamical system for computing canonical correlations is derived from a logarithmic merit function weighted by a diagonal matix. The merit function that will be considered are defined as

$$G(x, y) = tr\{D \log(x^T A y)\} - \frac{\alpha}{2} tr\{(x^T B x + y^T C y)\}, \tag{10}$$

where $A = R_{xy} \in \mathbb{R}^{n\times m}$, $B = R_{xx} \in \mathbb{R}^{n\times n}$, $C = R_{yy} \in \mathbb{R}^{m\times m}$, $D \in \mathbb{R}^{p\times p}$, $x \in \mathbb{R}^{n\times p}$, $y \in \mathbb{R}^{m\times p}$, and $\alpha > 0$. It will be assumed that B and C are positive definite matrices. The diagonal matrix is incorporated within the merit function G, since the principal directions x and y estimated using the merit function G with $D = I$ are rotated singular vectors of the coherence matrix $\hat{C}$. On the other hand, if D is a diagonal matrix whose eigenvalues are all positive and distinct, then the function G is optimized at exact principal x and y directions.

The merit function G can be shown to be upper bounded and $-G$ is radially unbounded. Thus gradient systems converge to the principal singular components of the coherence matrix. The case where $B = I$, $C = I$, and $D = I$ leads to computing singular subspaces.

The behavior of the function G can be illustrated in the following simple example.

Example 1: Let $F(x, y) = d \log(axy) - \frac{b}{2}x^2 - \frac{c}{2}y^2$, where $b, c, d > 0$ and $a \neq 0$. The function F is defined for all (x, y) such that $axy > 0$. The objective is to find the minima and maxima of F over $\mathbb{R}^2$. The gradient and the Hessian matrix of F can be verified to be

$$\nabla F = \begin{bmatrix} \frac{d}{x} - bx \\ \frac{d}{y} - cy \end{bmatrix}, \tag{11a}$$

and

$$\nabla^2 F = \begin{bmatrix} -\frac{d}{x^2} - b & 0 \\ 0 & \frac{-d}{y^2} - c \end{bmatrix}. \tag{11b}$$

The equilibrium points of F are solutions of the equations

$$\begin{aligned} \frac{d}{x} - bx &= 0, \\ \frac{d}{y} - cy &= 0. \end{aligned} \tag{12}$$

When $x \neq 0$ and $y \neq 0$, these equations imply that

$$x^2 = \frac{d}{b},$$

$$y^2 = \frac{d}{c}.$$

Thus the set of equilibrium points consists of the point $(\hat{x}, \hat{y})$ such that

$$\hat{x} = \pm\sqrt{\frac{d}{b}},$$
$$\hat{y} = \pm\sqrt{\frac{d}{c}}.$$

Since $\frac{d}{\hat{x}^2} = b$, and $\frac{d}{\hat{y}^2} = c$, the Hessian matrix simplifies to

$$\nabla^2 F = \begin{bmatrix} -2b & 0 \\ 0 & -2c \end{bmatrix}. \tag{13}$$

Obviously this matrix is negative definite at the equilibrium points $(\hat{x}, \hat{y})$. The maximum of $F = d\{\log|a| + \frac{1}{2}\log(\frac{d^2}{bc}) - 1\} = d\{\frac{1}{2}\log(\frac{|a|^2 d^2}{bc}) - 1\}$. If $d = 1$, then the maximum of $F = \log|a| - \frac{1}{2}\log(bc) - 1 = \log\frac{|a|}{\sqrt{bc}} - 1$.

In the next section, example 1 will be generalized to higher dimensions.

3.1 A Gradient Dynamical System

In this section we analyze the dynamical system resulted from the merit function (10). From Proposition 1, it follows that the gradient can be expressed in terms of the matrices A, D, x, y as follows:

$$\nabla G = \begin{bmatrix} AyD(x^TAy)^{-1} + Ay(x^TAy)^{-1}D - \alpha Bx \\ A^TDx(y^TA^Tx)^{-1} + A^Tx(y^TA^Tx)^{-1}D - \alpha Cy \end{bmatrix}. \tag{14a}$$

Thus a gradient dynamical system for maximizing G may be expressed by

$$\begin{aligned} x' &= AyD(x^TAy)^{-1} + Ay(x^TAy)^{-1}D - \alpha Bx \\ y' &= A^TDx(y^TA^Tx)^{-1} + A^Tx(y^TA^Tx)^{-1}D - \alpha Cy. \end{aligned} \tag{14b}$$

Let $(x(t), y(t))$ be a solution of (14b) for $t \geq 0$, then $x(t)^Tx(t)$, $y(t)^Ty(t)$, $x(t)^TAy(t)$, $x(t)^TBx(t)$, and $y(t)^TCy(t)$ can be shown to converge to diagonal matrices as $t \to \infty$.

For convenience, assume that $\alpha = 1$ and let $(x(t), y(t))$ be a solution of (14b) for $t \geq 0$. Set $\bar{A} = \lim_{t\to\infty} x(t)^TAy(t)$, $\bar{B} = \lim_{t\to\infty} x(t)^TBx(t)$, and $\bar{C} = \lim_{t\to\infty} x(t)^TCx(t)$. In what follows it will be assumed that $\bar{A}$ is invertible and that $\bar{A}^T\bar{A}$ has distinct eigenvalues. Equation (14b) implies that

$$D + \bar{A}D\bar{A}^{-1} = \bar{B}, \tag{15a}$$
$$D + \bar{A}^TD\bar{A}^{-T} = \bar{C}. \tag{15b}$$

Since $\bar{B} - D$ and $\bar{C} - D$ are symmetric, it follows that $\bar{A}D\bar{A}^{-1}$ and $\bar{A}^TD\bar{A}^{-T}$ are symmetric, i.e.,

$$\bar{A}D\bar{A}^{-1} = \bar{A}^{-T}D\bar{A}^T,$$
$$\bar{A}^TD\bar{A}^{-T} = \bar{A}^{-1}D\bar{A}.$$

After a few rearrangement, the last two equations yield

$$\bar{A}^T\bar{A}D = D\bar{A}^T\bar{A},$$

and

$$\bar{A}\bar{A}^TD = D\bar{A}\bar{A}^T.$$

Since D is diagonal and all its eigenvalues are distinct, then $\bar{A}^T\bar{A}$ and $\bar{A}\bar{A}^T$ are diagonal:

$$\bar{A}\bar{A}^T = D_1,$$
$$\bar{A}^T\bar{A} = D_2,$$

where D_1 and D_2 are diagonal matrices.

From these two equations, we have

$$\bar{A}D_1 = D_2\bar{A}.$$

Since $\bar{A}$ is assumed to be invertible, then $\bar{A}D_1\bar{A}^{-1} = D_2$ and hence $\bar{A}D_1\bar{A}^{-1} = D_2 = PD_1P^T$, where P is a permutation matrix. The last equation implies that

$$P^T\bar{A}D_1 = D_1P^T\bar{A}.$$

Assuming that all eigenvalues of $\bar{A}\bar{A}^T$ are distinct, it follows that

$$P^T\bar{A} = D_3,$$

or equivalently

$$\bar{A} = PD_3,$$

where D_3 is diagonal matrix. From the relation

$$\bar{A}^T\bar{A} = D_2 = D_3P^TPD_3 = D_3^2,$$

it follows that

$$D_3 = \sqrt{D_2},$$

and consequently

$$\bar{A} = P\sqrt{D_2}.$$

Additionally,

$$\bar{A}\bar{A}^T = D_1 = PD_3^2P^T.$$

Therefore,

$$D_2 = P^TD_1P.$$

Next we show that both $\bar{B}$ and $\bar{C}$ are diagonal. Clearly,

$$\bar{B} = D + P\sqrt{D_2}D\sqrt{D_2^{-1}}P^T = D + PDP^T.$$

Since PDP^T is diagonal, $\bar{B} = D + PDP^T$ is also diagonal. Similarly,

$$\bar{C} = D + \sqrt{D_2}P^TDP\sqrt{D_2^{-1}} = D + P^TDP,$$

is diagonal.

To check whether the equilibrium points $(\hat{x}, \hat{y})$ of the system (14) are maximizers for the function G, the Hessian matrix $H(G)$ evaluated at $(\hat{x}, \hat{y})$ is negative definite. This matrix is obtain by simplifying (8b) after a proper choice of E.

4 Simulation Results

In this section, we present an example that demonstrates the behavior and the applicability of the proposed algorithms. In this example we tested the proposed algorithm using the matrices $A \in \mathbb{R}^{9\times7}$, $B \in \mathbb{R}^{9\times9}$, and $C \in \mathbb{R}^{7\times7}$. These are generated using the following formulas:

$$A = \frac{1}{100}\sum_{k=1}^{100} x(k)y(k)^T,$$
$$B = \frac{1}{100}\sum_{k=1}^{100} x(k)x(k)^T,$$
$$C = \frac{1}{100}\sum_{k=1}^{100} y(k)y(k)^T,$$

where $x(k)$ and $y(k)$ are vectors of sizes 9×1 and 9×1, respectively. The vectors $x(k)$ and $y(k)$ are generated using the Matlab function *rand*. The matrices A, B and C which are used in this example are

```
A =0.3710 0.3653 0.3792 0.3878 0.4032 0.3628 0.3873
   0.3289 0.3334 0.3542 0.3526 0.3611 0.3191 0.3525
   0.3821 0.3567 0.3673 0.3753 0.3926 0.3518 0.3854
   0.3560 0.3442 0.3732 0.3843 0.3825 0.3500 0.3674
   0.3676 0.3660 0.3759 0.3902 0.4122 0.3586 0.3839
   0.3575 0.3591 0.3702 0.3709 0.3881 0.3453 0.3782
   0.3591 0.3606 0.3805 0.3806 0.3845 0.3524 0.3701
   0.3904 0.3829 0.4118 0.4014 0.4215 0.3685 0.4142
   0.3436 0.3371 0.3651 0.3552 0.3784 0.3460 0.3679
```

The nine columns of the matrix B are

```
B=
0.5054 0.3536 0.3821 0.3772 0.3909 0.3719 0.3818
0.3536 0.4488 0.3508 0.3415 0.3538 0.3357 0.3413
0.3821 0.3508 0.4986 0.3742 0.3823 0.3742 0.3889
0.3772 0.3415 0.3742 0.5005 0.3784 0.3658 0.3703
0.3909 0.3538 0.3823 0.3784 0.5289 0.3967 0.4017
0.3719 0.3357 0.3742 0.3658 0.3967 0.4798 0.3714
0.3818 0.3413 0.3889 0.3703 0.4017 0.3714 0.4929
0.4228 0.3739 0.4034 0.3980 0.4114 0.3927 0.4050
0.3758 0.3282 0.3615 0.3747 0.3808 0.3598 0.3527

    0.4228 0.3758
    0.3739 0.3282
    0.4034 0.3615
    0.3980 0.3747
    0.4114 0.3808
    0.3927 0.3598
    0.4050 0.3527
    0.5597 0.3985
    0.3985 0.4862

C =0.4663 0.3558 0.3652 0.3820 0.3767 0.3468 0.3720
   0.3558 0.4638 0.3607 0.3609 0.3665 0.3457 0.3622
   0.3652 0.3607 0.4997 0.3855 0.3920 0.3529 0.3834
   0.3820 0.3609 0.3855 0.5190 0.3992 0.3614 0.3818
   0.3767 0.3665 0.3920 0.3992 0.5269 0.3618 0.3930
   0.3468 0.3457 0.3529 0.3614 0.3618 0.4574 0.3441
   0.3720 0.3622 0.3834 0.3818 0.3930 0.3441 0.4911
```

The canonical correlations of the coherence matrix $\hat{C} = B^{\frac{-1}{2}} A C^{\frac{-1}{2}}$ are 0.9563, 0.2371, 0.1738, 0.1380, 0.1296, 0.0895, 0.0412. Algorithm (14b) is applied with input matrices A, B, C, as given and the diagonal matrix D is

```
D =23.0914        0        0
         0  15.6121        0
         0        0   6.9251
```

Euler method with stepsize $\alpha = 0.0615$ is used to solve the dynamical system (14b). After 7,000 iterations $x(k)$ and $y(k)$ converge so that

```
x'*A*y=-0.0000   -0.0000   -7.1165
       -0.0000  -29.8604    0.0000
       -5.2174   -0.0000    0.0000

x'*B*x= 30.0165    0.0000    0.0000
         0.0000   31.2242    0.0000
         0.0000    0.0000   30.0165

y'*C*y= 30.0165    0.0000    0.0000
         0.0000   31.2242    0.0000
         0.0000    0.0000   30.0165
```

The singular values of $(x^T Bx)^{-.5}(x^T Ay)(y^T Cy)^{-.5}$ are 0.9563, 0.2371, 0.1738. These are the largest three canonical correlations of $\hat{C} = B^{\frac{-1}{2}} A C^{\frac{-1}{2}}$. It can be verified that $\bar{B} = \bar{C} = D + PDP^T$, where $P = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}$.

5 Appendix

In this appendix, we list a number of results that are used in proving some of the propositions of this work.

5.1 Gradient and Hessian Matrices

The gradient and Hessian matrices can be obtained from first and second order differentials as the following lemma [11].

Lemma 3. *Let ϕ be a twice differentiable real-valued function of an $n \times p$ matrix. Then, the following relationships hold:*

$$d\phi(X) = tr(A^T dX) \Leftrightarrow \nabla\phi(X) = A \quad (A.1)$$

$$d^2\phi(X) = tr(B(dX)^T CdX) \Leftrightarrow H\phi(X) = \frac{1}{2}(B^T \otimes C + B \otimes C^T) \quad (A.2)$$

$$d^2\phi(X) = tr(B(dX)CdX) \Leftrightarrow H\phi(X) = \frac{1}{2}K_{rn}(B^T \otimes C + C^T \otimes B) \quad (A.3)$$

where d denotes the differential, and A, B, and C are matrices, each of which may be a function of X. The gradient of ϕ with respect to X and the Hessian matrix of ϕ at X are defined as

$$\nabla\phi(X) = \frac{\partial\phi(X)}{\partial X}$$

$$H\phi(X) = \frac{\partial}{(vecX)^T}\left(\frac{\partial\phi(X)}{\partial(vecX)^T}\right)^T \quad (A.4)$$

where vec is the vector operator and stands for the operation of stacking the columns of a matrix into one column, and $\otimes$ denotes the Kronecker product. The matrix K_{pn} denotes the $pn \times pn$ commutation matrix; $K_{pn}^T = K_{pn}^{-1} = K_{pn}$ and $K_{pm}(A \otimes C) = (C \otimes A)K_{qn}$, where $A \in \mathbb{R}^{m\times n}$ and $C \in \mathbb{R}^{r\times q}$.

Proposition 4. *Let $D, A \in \mathbb{R}^{n\times n}$ be positive definite matrices and assume that D is diagonal having distinct eigenvalues. If $AD = DA$, then A is diagonal.*

Proof: Assume that $A = [a_{ij}]$ and $D = \mathrm{diag}(\mu_1, \cdots, \mu_n)$, then for each i, j we have $a_{ij}\mu_j = \mu_i a_{ij}$ or $(\mu_j - \mu_i)a_{ij} = 0$. Thus $a_{ij} = 0$ for $i \neq j$, i.e., A is diagonal.

Proposition 5 [12]. *Let $B, D \in \mathbb{R}^{p\times p}$ and assume that D is diagonal and all eigenvalues of D are distinct. If $BD + DB$ is diagonal, then B is diagonal.*

References

[1] H. Hotelling, The most predictable criterion, J. Educ. Psychol. 26 (1935) 139-142.

[2] H. Hotelling, "Relations between two sets of variates," Biometrika, Vol. 28, pp. 321-377, 1936.

[3] T. W. Anderson, An Introduction to Multivariate Statistical Analysis, 2nd ed. New York: Wiley, 1984.

[4] K.V. Mardia, J.T. Kent, J.M. Bibby, Multivariate Analysis, Academic Press, New York, 1979.

[5] S.G. Shi, W. Taam, "Non-linear canonical correlation analysis with a simulated annealing solution," J. Appl. Statist. 19 (1) (1992) 155-165.

[6] M.A. Hasan, "A new approach for computing canonical correlations and coordinates," Proceedings of the 2004 International Symposium on Circuits and Systems, 2004, Volume: 3, 23-26 May 2004, Pages:309-312.

[7] E. van der Burg and J. de Leeuw, "Non-linear Canonical Correlation Analysis", Brit. J. Math. Statist. Psychol., Vol 36, pp. 54-80, 1983.

[8] J.R. Kettenring, "Canonical analysis of several sets of variables," Biometrika 58 (1971) 433-451.

[9] A.A. Nielsen, "Multiset canonical correlations analysis and multispectral, truly multitemporal remote sensing data Image Processing," IEEE Transactions on, Volume: 11 Issue: 3 , March 2002, Page(s): 293-305.

[10] Malte Kuss and Thore Graepel, "The Geometry Of Kernel Canonical Correlation Analysis," Technical Report No. 108, May 2003, Max Planck Institute for Biological Cybernetics.

[11] J. R. Magnus and H. Neudecker, Matrix Differential Calculus with Applications in Statistics and Econometrics, 2nd ed. New York: Wiley, 1991.

[12] M.A. Hasan, "Natural Gradient for Minor Component Extraction," Proceedings of the IEEE International Symposium on Circuits and Systems, ISCAS 2005. 23-26 May 2005.