Simultaneous Minor Component Extraction via Weighted Inverse Rayleigh Quotient

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

Abstract. *New criteria are proposed for extracting multiple minor components associated with the covariance matrix of an input process. The proposed minor component analysis (MCA) algorithms are based on optimizing a weighted inverse Rayleigh quotient so that the optimum weights at equilibrium points are exactly the desired eigenvectors of a covariance matrix instead of an arbitrary orthonormal basis of the minor subspace. Variations of the derived MCA learning rules are obtained by imposing orthogonal and quadratic constraints and change of variables. Some of the proposed algorithms can also perform PCA by merely changing the sign of the step-size. These algorithms may be seen as MCA counterparts of Oja's and Xu's systems for computing multiple principal component analysis. Simulation results to demonstrate algorithm performance are also presented.*

Keywords: minor component analysis, principal component analysis, Oja's learning rule, inverse Rayleigh quotient.

1. Introduction

The minor subspace of dimension p associated with a covariance matrix B is spanned by the eigenvectors corresponding to the smallest p eigenvalues of the sample covariance matrix. The minor components are the directions in which the data have the smallest variances. They usually represent the statistics of the additive noise. MCA has been applied to frequency estimation [1], [2], bearing estimation [3], and digital beamforming [4]. It has also been applied to total least squares algorithms for parameter estimation [5], [6]; The problem encountered in many real-time applications is the difficulty of successively estimating the minor eigenvectors that span the desired subspace. Several stochastic gradient algorithms have been reported as viable solutions [7], but most of them suffer from a slow convergence rate. A comprehensive analysis of single minor component analysis is given in [8].

It is known that the Oja flow is only capable of extracting the principal subspace but not the principal eigenvectors. Flows that achieve the principal component analysis were first proposed by Sanger [9], Oja et al. [10], and Xu [11]. Dynamical systems for computing minor components and diagonalization are given in [12] and [13], respectively. In this paper we develop gradient flows that are capable of extracting the minor subspace and the minor eigenvectors from the optimization of a weighted inverse Rayleigh quotient (WIRQ). WIRQ has several attractive properties that can be exploited for deriving MCA and PCA algorithms. These include boundedness, homogeneity and some orthogonality properties. Additionaly, WIRQ has only one minima and one maxima.

2. Problem Formulation

Suppose that the input vector sequence $x_k \in \mathbb{R}^n$ is a stationary stochastic process with zero mean and covariance matrix $B = E(xx^T)$ with the eigenvalues $0 < \lambda_1 < \cdots < \lambda_n$ and the corresponding orthonormal eigenvectors $z_1, \cdots, z_n$. Let p be an integer such that $1 \leq p \leq n$ and let the eigendecomposition of B be given as $B = Z_1\Lambda_1 Z_1^T + Z_2\Lambda_2 Z_2^T$, where $Z_1 = [z_1, \cdots, z_p]$, $\Lambda_1 = diag\{\lambda_1, \cdots, \lambda_p\}$, $\Lambda_2 = diag\{\lambda_{p+1}, \cdots, \lambda_n\}$, and $Z_2 = [z_{p+1}, \cdots, z_n]$. To find the p $(1 \leq p \leq n)$ smallest eigenvalues $\lambda_1, \cdots, \lambda_p$, and corresponding eigenvectors $z_1, \cdots, z_p$ we define the following criterion:

$$\text{Maximize}\{F(U) = tr\{(U^TU)(U^TBU)^{-1}D\} \tag{1}$$

over all full rank matrices $U \in \mathbb{R}^{n\times p}$. Here $tr(X)$ denotes the trace of a square matrix X, $(.)^T$ denotes matrix transpose, D is a diagonal matrix of size p having distinct positive eigenvalues. It will be assumed that $D = \text{diag}(\mu_1, \cdots, \mu_p)$ and that $\mu_1 > \mu_2 > \cdots > \mu_p > 0$.

In the sequel, we use WRQ to denote the weighted Rayleigh quotient defined by $WRQ(U, B, D) = (U^TBU)(U^TU)^{-1}D$ and $WIRQ$ to denote the weighted inverse Rayleigh quotient defined by $WIRQ(U, B, D) = (U^TU)(U^TBU)^{-1}D$. Also, unless otherwise stated the gradient is defined with respect to the Euclidean inner product $\langle X, Y\rangle = tr(X^TY)$ on matrix space.

The most relevant properties of the WRQ and $WIRQ$ are summarized in the following remarks:

1. **Homogeneity:** In general $F(U) = F(UE)$ for any diagonal matrix E. The converse is also true, i.e., if $F(U) = F(UE)$ and D is diagonal matrix with distinct eigenvalues, then E is essentially diagonal. This property explains why we consider the weighted forms of Rayleigh quotient for computing MCA. However, if $D = I_p$, where I_p denotes a $p \times p$ identity matrix, then $F(U) = F(UE)$ holds for any nonsingular matrix E. The last property implies that any solution of (1) with $D = I_p$ is determined upto a multiplication by a nonsingular matrix. However, this indeterminacy is removed if a diagonal matrix D is incorporated.
2. **Boundedness:** As $U \in \mathbb{R}^{n\times p}$ ranges over all full rank matrices, $F(U)$ remains bounded from above and below.
3. **Orthogonality:** If $D = I_p$ the following orthogonality property holds

$$\begin{aligned} U^T\{U - BU(U^TBU)^{-1}U^TU\} &= 0 \\ U^T\{BU - U(U^TU)^{-1}U^TBU\} &= 0. \end{aligned} \tag{2}$$

   These properties are equivalent to performing the Gram-Schmidt process to the matrices U and BU with respect to some dot products.

The following proposition deals with the WIRQ critical points.

Proposition 1 (Stationarity). *Let D be a diagonal matrix such that the diagonal entries of D are positive, distinct, and arranged in descending order and let B be a real symmetric n-dimensional matrix with eigenvalues $0 < \lambda_1 < \cdots < \lambda_p < \lambda_{p+1} < \cdots < \lambda_n$ and the corresponding orthonormal eigenvectors $z_1, \cdots, z_n$. Then*

$$\max\{F(U)\} = \sum_{k=1}^{p} \frac{d_k}{\lambda_k} \tag{3a}$$

$$\min\{F(U)\} = \sum_{k=1}^{p} \frac{d_k}{\lambda_{n-k+1}}. \tag{3b}$$

Moreover, the global minimum and the global maximum are attained if and only if $U = Z_1\Pi_1$ and $U = Z_2\Pi_2$ respectively, where

$Z_1 = [z_1 \cdots z_p]$ and $Z_2 = [z_{p+1} \cdots z_n]$ and Π_1, Π_2 are permutation matrices. More generally, the critical points and critical values of $F(U)$ are of the form $U = [z_{i_1} \cdots z_{i_p}]$, where $(i_1, \cdots, i_p)$ is a pth order permutation of $\{1, \cdots, n\}$. All other critical points are saddles.

Outline of Proof: Let $U = ZE$, where Z is any matrix consisting of p eigenvectors, and E is a nonsingular matrix, then

$$\begin{aligned} F(ZE) &= tr((E^T E)(E^T \Lambda E)^{-1} D) \\ &= tr(E^T \Lambda^{-1} E^{-T} D), \end{aligned} \tag{4}$$

where $\Lambda = diag(\lambda_{i_1} \cdots \lambda_{i_p})$ and $(i_1, \cdots \lambda_{i_p})$ is a permutation of $\{1, \cdots, n\}$. The possible maximum or minimum of $F(U)$ occurs when $\nabla_E tr(WIRQ(ZE, B, D)) = \nabla_E F(ZE) = 0$. It can be shown (see Appendix, Proposition 5) that

$$\nabla_E F(ZE) = -E^{-T} D E^T \Lambda^{-1} E^{-T} + \Lambda^{-1} E^{-T} D. \tag{5}$$

This implies that

$$E^T \Lambda^{-1} E^{-T} D = D E^T \Lambda^{-1} E^{-T}.$$

Since D is diagonal with distinct eigenvalues, it follows from Proposition 6 that $E^T \Lambda_p^{-1} E^{-T}$ is diagonal. Thus the only possible solution of $\nabla_E tr(F(ZE)) = 0$ is that $E = D_1 P$, where D_1 is diagonal, and P is a permutation matrix. Now, at stationarity points the objective function is given by

$$F(E) = F(D_1 P) = tr(P \Lambda^{-1} P^T D). \tag{6}$$

Clearly, since the diagonal entries of D are in descending order, then among all possible Λ and all possible permutations P, the maximum of $F(U)$ occurs at $\Lambda = \Lambda_1$ and $P = I_p$. Similarly, the minimum occurs at $\Lambda = diag(\lambda_{p+1}, \cdots, \lambda_n)$ and $P = J$, where J is the interchange matrix given by $J = [e_p, e_{p-1} \cdots e_1]$ where e_i is the ith column of a $p \times p$ identity matrix I_p. To examine the critical points for maxima and minima, we have to show that the Hessian matrix defined (see Appendix) as $H\phi(U) = \frac{\partial}{(vecU)^T} \left(\frac{\partial \phi(U)}{\partial (vecU)^T} \right)^T$, where $\phi(U) = tr(F(U))$, is positive definite at $U = Z_1$ and negative definite at $U = Z_2$. After some derivations, we have

$$\begin{aligned} \frac{d^2 F}{dU^2} &= D(U^T BU)^{-1} \otimes I_p + (U^T BU)^{-1} D \otimes I_p \\ &- D(U^T BU)^{-1} \otimes U(U^T BU)^{-1} U^T B + (U^T BU)^{-1} D \\ &\otimes BU(U^T BU)^{-1} U^T - K_{pn}\{BU(U^T BU)^{-1} D \\ &\otimes (U^T BU)^{-1} U^T - D(U^T BU)^{-1} U^T B \otimes U(U^T BU)^{-1}. \end{aligned} \tag{7}$$

Here vec stands for the operation of stacking the columns of a matrix into one column, and $\otimes$ denotes the Kronecker matrix product. The matrix K_{rn} denotes the $pn \times pn$ commutation matrix; $K_{pn}^T = K_{pn}^{-1} = K_{pn}$ and $K_{rm}(A \otimes C) = (C \otimes A) K_{qn}$, where $A \in \mathbb{R}^{m \times n}$ and $C \in \mathbb{R}^{r \times q}$. After long but straightforward calculations (not included due to space limitation), it can be shown that $Htr(FU)$ is negative definite at $U = Z_1$ and positive definite at $U = Z_2$. It is non definite at any other critical points.

Q. E. D.

Proposition 1 indicates that with a properly chosen D, $F(U)$ has exactly one global minima and one global maxima.

3. Unconstrained Optimization Procedure

In effect, Equation (1) is an unconstrained optimization problem on the set $\Omega = \{U : U^T BU > 0\}$. Deriving $F(U)$ with respect to U, we obtain the gradient equation (see Appendix)

$$\begin{aligned} \nabla F &= U(U^T BU)^{-1} D + UD(U^T BU)^{-1} - BU(U^T BU)^{-1} \\ &\times DU^T U(U^T BU)^{-1} - BU(U^T BU)^{-1} U^T UD(U^T BU)^{-1}. \end{aligned} \tag{8}$$

If $D = I_p$, then ∇F simplifies to

$$\nabla F = U(U^T BU)^{-1} - BU(U^T BU)^{-1} U^T U(U^T BU)^{-1}. \tag{9}$$

It can easily be seen that an orthogonality property holds $U^T \nabla F = 0$ or $U \perp \nabla F$. This property turns out to be significant in convergence analysis (see Section 4) by analyzing the ordinary differential equation (ODE) associated with the gradient system (8):

$$\begin{aligned} U' &= \nabla F = U\{(U^T BU)^{-1} D + D(U^T BU)^{-1}\} \\ &- BU(U^T BU)^{-1}\{DU^T U + U^T UD\}(U^T BU)^{-1}, \end{aligned} \tag{10}$$

where $U'(t) = \frac{dU(t)}{dt}$. To alleviate matrix inversion, the quadratic constraint $U^T BU = I_p$ is imposed so that for any U satisfying $U^T BU = I_p$ we have

$$U' = \nabla F = 2UD - BU\{DU^T U + U^T UD\}. \tag{11}$$

In the next proposition, we show that under a mild condition, the gradient ascent with sufficiently small step-size converges to MCA.

Proposition 2. *Let D and B be as in Proposition 1 and let U_∞ be the solution of the difference equation*

$$U_{k+1} = U_k + \alpha\{U_k D - \frac{1}{2} BU_k (DU_k^T U_k + U_k^T U_k D\}, \tag{12}$$

for some learning step size $\alpha \in (0, 1)$. Assume that $DU_\infty^T U_\infty + U_\infty^T U_\infty D$ is non-singular. Then the limiting solution U_∞ of the gradient ascent iteration (12) satisfies the following:

1. $U_\infty^T BU_\infty = I_p$
2. $U_\infty^T U_\infty$ *is diagonal*
3. $F(U_\infty) = \sum_{k=1}^p \frac{d_k}{\lambda_k}$
4. $U_\infty = Z_1 \Lambda_1^{\frac{-1}{2}}$

Outline of Proof: Since there is only one maxima, then for any initial matrix U_0 satisfying $U_0^T BU_0 = I_p$ the gradient ascent (12) converges globally to system's equilibrium point. Assume that U_∞ is the limiting solution of the gradient ascent iteration (12), then $U_\infty^T U_\infty D = U_\infty^T BU_\infty H$, where $H = U_\infty^T U_\infty D + DU_\infty^T U_\infty$. Clearly,

$$2H = HU_\infty^T BU_\infty + U_\infty^T BU_\infty H. \tag{13}$$

We show next that each eigenvalue of $U_\infty^T BU_\infty$ is equal to 1. Let λ be an eigenvalue of $U_\infty^T BU_\infty$ with corresponding eigenvector x, then $U^T BUx = \lambda x$. By post-multiplying and pre-multiplying both sides of (13) by x and x^T, respectively we obtain $2x^T Hx = \lambda x^T Hx + \lambda x^T Hx$ and thus $(1 - \lambda) x^T Hx = 0$. The non-singularity of H implies that $\lambda = 1$. Since $U_\infty^T BU_\infty$ is symmetric then $B = I_n$. Consequently, $U_\infty^T U_\infty D = DU_\infty^T U_\infty$. Since D is diagonal with distinct eigenvalues, we have from Proposition 6 that $U_\infty^T U_\infty$ is diagonal. This shows that $U_\infty^T U_\infty = \Lambda_1^{-1}$ and $BU_\infty^T = U_\infty (U_\infty^T U_\infty)^{-1} = U_\infty \Lambda_1^{-1}$. Consequently, $U_\infty = Z_1 \Lambda_1^{\frac{-1}{2}}$.

Q. E. D.

Remarks: It should be noted that although both matrices D and $U_\infty^T U_\infty$ are positive definite, the symmetric matrix $H = U_\infty^T U_\infty D + DU_\infty^T U_\infty$ generally may not be nonsingular or positive definite even when D is positive definite. Another observation is that since $U_\infty^T U_\infty D = DU_\infty^T U_\infty$, one may use the approximation $U_k^T U_k D \approx DU_k^T U_k$, for sufficiently large k so that (12) can be approximately expressed as

$$U_{k+1} = U_k + \alpha\{U_k D - BU_k DU_k^T U_k\}. \tag{14a}$$

Note that the last learning rule can be seen as the MCA counterpart of Xu's PCA learning method [11]. Now let us examine the special case $D = I_p$ and $U^T BU = I_p$. The ODE (10) reduces to a minor subspace gradient flow

$$U' = U - BUU^T U. \tag{14b}$$

However, this flow is only capable of extracting the minor subspace of B but not the minor eigenvectors. The limiting solution is orthonormal with respect to the Euclidean inner product $\langle X, Y \rangle =$

$tr(X^TBY)$ on matrix space, i.e., $U_\infty^T B U_\infty = I_p$. To extract the true MCA, a modification of (14b) can be made so that

$$U' = U - BUg(U^TU), \tag{15}$$

where $g(X)$ is the upper triangular part of X. Another variation of (9) is obtained by choosing $U^TBU = D$, in which case the ODE: $U' = U(U^TBU)^{-1} - BU(U^TBU)^{-1}U^TU(U^TBU)^{-1}$, reduces to

$$U' = UD - BUDU^TUD. \tag{16a}$$

Numerical simulations suggested that this learning rule performs only minor subspace analysis (MSA) but not the true MCA. If B is assumed symmetric, then $B + \alpha I$ is positive definite for some $\alpha \geq 0$. Thus system (14b) simplifies to

$$U' = U(I - \alpha U^TU) - BUU^TU. \tag{16b}$$

Numerous MCA/MSA variations can be obtained by a change of variables approach. For example by considering the change of variables: $U = \mu^{\frac{1}{2}} GVD^{\frac{1}{2}}$, where $\mu > 0$, D is diagonal mayrix, and G is positive definite matrix, then (14b) transforms into

$$\mu^{\frac{1}{2}} GV'D^{\frac{1}{2}} = \mu^{\frac{1}{2}} GVD^{\frac{1}{2}} - \mu^{\frac{3}{2}} BGVDV^TG^2VD^{\frac{1}{2}},$$

or equivalently,

$$V' = V - \mu G^{-1}BGVDV^TG^2V.$$

The matrix G can be chosen to be a diagonal preconditioner for cases where B is near singular. If $\mu = 1$ and $G = B^{\frac{1}{2}}$, we obtain the following gradient flow: $V' = V - BVDV^TBV$.

PCA Flows: The MCA flow of (10) can be converted to a PCA flow by merely changing the sign of the gradient. This follows directly from Proposition 1, Eq. (3b). Finally, by considering the matrix $\lambda I_n - B$ instead of B, where $\lambda > \lambda_n$, all the MCA rules in this paper convert to PCA learning rules.

Remarks: If the matrix B in (14b) is replaced with the identity matrix, we obtain the following orthonormalization dynamical system:

$$U' = U - UU^TU.$$

Clearly this system is a gradient system since it can be derived from the gradient of the cost function $\frac{1}{2}tr(U^TU) - \frac{1}{4}tr\{(U^TU)^2\}$. This cost function may be modified so that $F(U) = \frac{1}{2}tr(U^TU) - \frac{1}{2s}tr\{(U^TU)^s\}$. The corresponding gradient system is given by

$$U' = U - U(U^TU)^s, \text{ s is positive integer.}$$

The stability and invariant sets of this system can be examined using $V(U) = \frac{1}{4}tr\{(U^TU - I)^2\}$, in which case

$$\begin{aligned}\dot{V} &= tr\{(U^TU - I)(U^TU - (U^TU)^s)\} \\ &= -tr\{(U^TU - I)^2U^TU\sum_{k=0}^{s-1}(U^TU)^k\} \leq 0.\end{aligned}$$

Clearly, $\dot{V} = 0$ only if $(U^TU - I)U^T = 0$, or $(U^TU)^2 = U^TU$, i.e., U^TU is a projection. Thus from Lyapunov stability theory, the system is stable for any positive integer s. In any of the above orthonormalization flows, if a solution $U(t)$ esists for $t \geq 0$ satisfying $U(0) = U_0$, where U_0 is full rank, then $\lim_{t\infty} U(t) = U_0(U_0^TU_0)^{\frac{-1}{2}}$. Here $(U_0^TU_0)^{\frac{-1}{2}}$ represents the principal square root of $(U_0^TU_0)^{-1}$.

4. Dependence on Initial Conditions

Convergence properties of the proposed algorithm can be studied by considering the gradient rule (10). For simplicity, we only consider the case $D = I_p$. Thus, we consider the continuous-time dynamical system described by the (ODE)

$$U' = U(U^TBU)^{-1} - BU(U^TBU)^{-1}U^TU(U^TBU)^{-1}. \tag{17}$$

It should be noted that the system (17) is a minor subspace analyzer. It can be shown that by switching the sign of B, System (17) converts into a principal subspace analyzer. We show in the next result some properties of solutions of (17) and their dependence on the initial condition.

Proposition 3. *Let $U(t)$ be the solution of the ODE (17) in the interval $t \in [0, \infty)$, and assume that $U(0)^TZ_1$ is nonsingular. Then, for all $t \in [0, \infty)$, we have $U(t)^TU(t) = U(0)^TU(0)$, $\operatorname{rank} U(t) = \operatorname{rank} U(0)$, and $||U(t)|| = ||U(0)||$.*

Outline of Proof: From (17), $U(t)$ satisfies the following ODE:

$$\frac{d(U^TU)}{dt} = U'^TU + U^TU' = 0. \tag{19}$$

This shows that $U(t)^TU(t)$ is constant, or $U(t)^TU(t) = U(0)^TU(0)$. The rest of conclusion follows from the last observation.

Q. E. D.

Proposition 3 establishes that the solution of (17) is of the form $U(t) = U(0)Z(t)$ where $Z(t) \in \mathcal{R}^{p\times p}$ is orthogonal for all $t \in [0, \infty)$. It can easily be seen that $Z(t)$ satisfies the ODE:

$$\begin{aligned}&Z'(t) = (U^T(0)BU(0))^{-1}Z(t) - \\ &(U^T(0)U(0))^{-1}Z(t)U^T(0)BU(0)Z^T(t)(U^T(0)BU(0))^{-1}Z(t).\end{aligned} \tag{20}$$

The significance of the last ODE is that solving n-dimensional equation can be obtained by solving a p-dimensional equation (20) which has much less computational demand when $p << n$.

5. Simulation Results

In this section, we present a simulation result to demonstrate the behavior of the WIRQ algorithm. This simulation serves to show the transient behavior of the learning in the minor eigenvectors. By solving the associated ODE (11) using the Euler method, the difference equation (12) is obtained. A random vector sequence from an ideal 8×8 covariance matrix B is generated with these eigenvalues: $8.7822, 8.8189, 8.1260, 9.8730, 10.6173, 7.3788, 11.2716, 18.2215$. The gradient ascent (12) is applied with $p = 4$ and learning step size $\alpha = 0.02$. The matrix D is chosen as $D = \operatorname{diag}(0.54, 0.35, 0.25, 0.1)$. One hundred random experiments are conducted each using different initial matrix U_0. The parameters α, and D are kept the same. The convergence behavior in this example is measured in two ways. Both ways measure how fast $U_k^TU_k$ and $U_k^TBU_k$ converge to Λ^{-1} and I_4, respectively. The range of k is $k = 1, \cdots, 3000$. The first accuracy measure is computed as the Frobenius norm of the matrix of the off-diagonal elements of $(U_k^TU_k)^{-1}$ shown in Figure 1, and the second is given by $||U_k^TBU_k - I_4||_F$, as shown in Figure 2. Here $||X||_F$ denotes the Frobenius norm of X. It is observed from the Figures that the WIRQ algorithm converges to the true minor subspace.

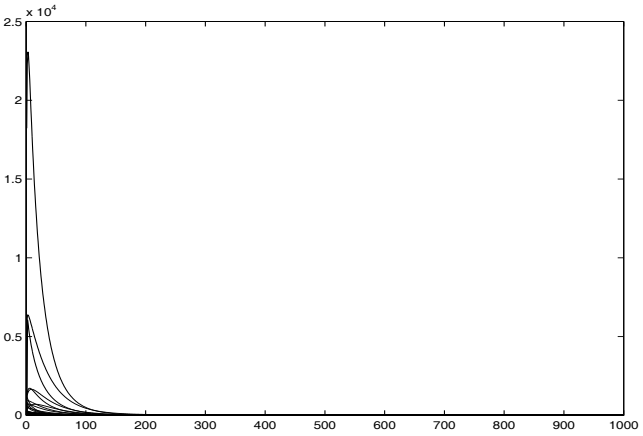

Figure 1: *Convergence behavior of $(U_k^TU_k)^{-1}$. The x-axis is the number of iterations, and the y-axis contains the magnitudes of the off-diagonal entries of $(U_k^TU_k)^{-1}$.*

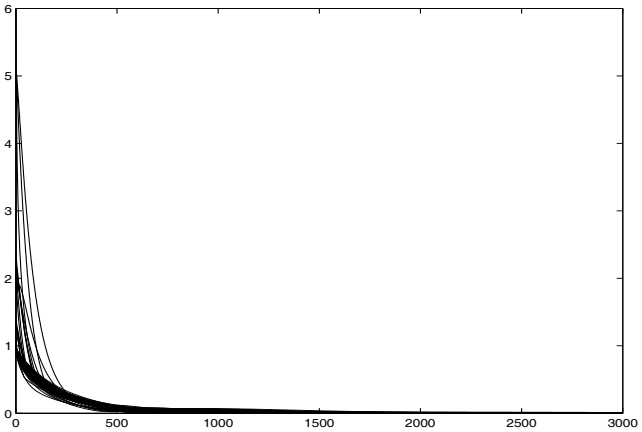

Figure 2: *Convergence behavior of $U_k^T BU_k$. The x-axis is the number of iterations, and the y-axis contains the Frobenius norm of $U_k^T BU_k - I_4$.*

6. Conclusion

This paper proposes an unconstrained optimization criterion using a weighted inverse Rayleigh quotient algorithm for extracting multiple minor components. Based on the gradient-ascent method, we derive several WIRQ algorithms for performing the true MCA recursively. Many issues remained to be analyzed including numerical complexity, global convergence, and comparisons with existing methods. The invariant sets of the system (14b) is another area of interest. Additional simulations not included here have shown that the system (14b) always converges even when the initial condition has very small nonzero norm. Generally, it is noticed that the system (14b) converges for any nonzero generic initial condition. This suggests that the system (14b) is globally convergent. Also the convergence behavior of of the systems (14a) and (16b) for symmetric matrices needs to analyzed. Some of these issues will be addressed in a forthcoming paper.

Appendix

We briefly present some facts from matrix differential calculus. The computation of derivatives can be performed simply based on the following lemma [14].

Lemma 4. *Let ϕ be a twice differentiable real-valued function of an $n \times p$ matrix. Then, the following relationships hold:*

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

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

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

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

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

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

where vec is the vector operator.

Proposition 5. *Let $g(X) = tr((XAX^{-1}D)$, where A and D are square matrices and X is a nonsingular square matrix, then*

$$\frac{dg}{dX} = -X^{-T}AX^T DX^{-T} + DX^{-T}A$$

The proof of this result is a direct application of A.2 and A.3.

Proposition 6. *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 $D = \text{diag}(\mu_1, \cdots, \mu_n)$, 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.

Theorem 7. *Let $D, A \in \mathbb{R}^{n\times n}$ be positive definite matrices and assume that D is diagonal having distinct eigenvalues. If $PDP = D$ for some positive definite matrix P, then $P = I_p$.*

Proof: From the assumption that $PDP = D$, it follows that $DP = P^{-1}D$, and thus $D^2P = DP^{-1}D = PD^2$, i.e., P and D^2 commute. Proposition (6) implies that P is diagonal. It follows that $P^2 = I_p$, or equivalently $P = diag(d_1, \cdots, d_p)$, where $d_i = \pm 1$. Since P is positive definite, we have $P = I_p$.

Q. E. D.

Corollary 8. *Let D be a diagonal matrix with distinct eigenvalues and let P be positive definite. If PDP is diagonal, then P is diagonal.*

Proof: Assume that $PDP = D_1$ for some diagonal matrix D_1, and let $D_2 = (DD_1^{-1})^{\frac{1}{4}}$. Then $\bar{P}D_3\bar{P} = D_3$, where $\bar{P} = D_2PD_2$ and $D_3 = (DD_1)^{\frac{1}{2}}$. Theorem 7 implies that $\bar{P}$ is diagonal and hence P is diagonal.

Q. E. D.

References

[1] G. Mathew and V. Reddy, "Development and analysis of a neural net-work approach to Pisarenkos harmonic retrieval method," IEEE Trans. Signal Processing, vol. 42, pp. 663-667, 1994.

[2] G. Mathew and V. Reddy, "Orthogonal eigensubspace estimation using neural networks, IEEE Trans. Signal Processing, vol. 42, pp. 1803-1811, 1994.

[3] R. Schmidt, "Multiple emitter location and signal parameter estimation," IEEE Trans. Antennas Propagation, vol. 34, pp. 276-280, 1986.

[4] J. W. Griffiths, "Adaptive array processing, a tutorial," Proc. Inst. Elect. Eng., pt. F, vol. 130, pp. 3-10, 1983.

[5] K. Gao, M. O. Ahmad, and M. N. Swamy, "Learning algorithm for total least-squares adaptive signal processing," Electron. Lett., vol. 28, no. 4, pp. 430-432, Feb. 1992.

[6] K. Gao, M. O. Ahmad, and M. N. Swamy, "A constrained anti-Hebbian learning algorithm for total least squares estimation with applications to adaptive fir and iir filtering," IEEE Trans. Circuits Syst. Part II, vol. 41, pp. 718-729, Nov. 1994.

[7] E. Oja, and J. Karhunen, "On stochastic approximation of the eigenvectors and eigenvalues of the expectation of a random matrix", J. Math. Anal. Appl., 1985, 106, pp. 69-84.

[8] G. Cirrincione, M. Cirrincione, J. Herault, and S. Van Huffel, "The MCA EXIN Neuron for the Minor Component Analysis," IEEE Trans. on Neural Networks, Vol. 13, No. 1, pp. 160-187, January 2002.

[9] T. D. Sanger, "Optimal unsupervised learning in a single-layer linear feedforward network," Neural Networks, 2:459-473, 1989.

[10] E. Oja, "Principal components, minor components, and linear neural networks. Neural Networks," 5:927-935, November 1992.

[11] L. Xu, "Least mean square error recognition principle for self organizing neural nets," Neural Networks, 6:627-648, 1993.

[12] J. H. Manton, Uwe Helmke, and Iven M. Y. Mareels, "Dynamical Systems for Principal and Minor Component Analysis," Proceedings of the 42nd IEEE Conference on Decision and Control, Maui, Hawaii USA, pp. 1863-1868, December 2003.

[13] R. W. Brockett, "Dynamical systems that sort lists, diagonalise matrices, and solve linear programming problems," Linear Algebra Appl., 146:79-91, 1991.

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