Proceedings of the
47th IEEE Conference on Decision and Control
Cancun, Mexico, Dec. 9-11, 2008

Low-Rank Approximations with Applications to Principal Singular Component Learning Systems

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

Abstract

In this paper, we present several dynamical systems for efficient and accurate computation of optimal low rank approximation of a real matrix. The proposed dynamical systems are gradient flows or weighted gradient flows derived from unconstrained optimization of certain objective functions. These systems are then modified to obtain power-like methods for computing a few dominant singular triplets of very large matrices simultaneously rather than just one at a time, by incorporating upper-triangular and diagonal matrices. The validity of the proposed algorithms was demonstrated through numerical experiments.

Keywords: SVD, Dynamical system, asymptotic stability, principal singular flow, Stiefel manifold, global convergence, constrained optimization

1 Introduction

Many engineering problems can be formulated so that their solutions are obtained from solving high dimensional singular value decomposition problems. In many applications such as signal processing,, image processing, and computational physics, the matrices involved are usually large and sparse. Therefore, there is a practical need for computing a few singular triplets of large matrices efficiently and accurately.

Theoretically, bases for principal subspace can be obtained via the singular value decomposition (SVD) of the data matrix. However, the cost of computing SVD directly may be too high for real-time applications where the data dimension is large. Therefore, efficient principal subspace are needed to track or estimate the desired subspaces.

There are many adaptive methods in the literature to obtain SVD of a rectangular matrix. SVD dynamical systems are developed in [1]-[11]. Algorithms for computing smallest singular triplets are proposed in [12]. Generalization of Oja's algorithm for obtaining the principal singular subspaces of a rectangular matrix is considered in [13, 14]. Cross-correlation neural network for extracting the cross-correlation features between two high-dimensional data streams is developed in [15]-[17]. A number of power-based subspace algorithms are presented in [18].

The motivation for studying power-like methods for computing principal singular components or subspaces is that they are simple to implement and always converge when all nonzero singular values are distinct. In general, if the nonzero singular values of a data matrix $A \in \mathbb{R}^{n\times m}$, where m, n are positive integers with $n \geq m$, are $\sigma_1 \geq \cdots \geq \sigma_p > \sigma_{p+1} \geq \cdots \geq \sigma_m$, then the speed of convergence of a power like method for computing the principal p-dimensional subspace of A is dependent on the ratio $\frac{\sigma_p+1}{\sigma_p}$. Slower convergence occurs when this ratio approaches unity.

The following notation will be used throughout. The notation $\mathbb{R}$, and $\mathbb{N}$ denote the set of real numbers, and the set of positive integers, respectively. The transpose of a real matrix 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 $tr(B)$ denotes the trace of B. The identity matrix of appropriate dimension is expressed with the symbol I. Finally, the derivative of $V(x, y)$ with respect to time is denoted by $\dot{V}$. For any vector or matrix x, the notation $||x||$ denotes the Euclidean norm of x. In the subsequent development, an algorithm will be said to converge to the true singular value components if it produces a sequence $(x(k), y(k))$ such that $x(k)^T x(k)$, $y(k)^T y(k)$, and $x(k)^T Ay(k)$ converge to diagonal matrices.

2 Low-Rank Approximation

Let $A \in \mathbb{R}^{n\times m}$, where $m, n \in \mathbb{N}$ with $n \geq m$, be a real matrix with singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p > \sigma_{p+1} \geq \cdots \geq \sigma_m \geq 0$ and the corresponding orthonormal left and right singular vectors are $U = [u_1, \cdots, u_m]$ and $V = [v_1, \cdots, v_m]$, respectively. The matrices U and V are orthogonal, i.e., $U^T U = I$ and $V^T V = I$. Thus A can be expressed as $A = \sum_{k=1}^{m} \sigma_k u_k v_k^T = U\Sigma V^T$, where Σ is a diagonal matrix with diagonal elements $\sigma_1, \sigma_2, \cdots, \sigma_m$. The expression $A_p = \sum_{k=1}^{p} \sigma_k u_k v_k^T = U_p \Sigma_p V_p^T$, $p \leq m$, is known as the low-rank p approximation of A in the sense of Frobenius norm, where $U_p = [u_1, \cdots, u_p]$, $\Sigma_p = \text{diag}(\sigma_1, \sigma_2, \cdots, \sigma_p)$, and $V_p = [v_1, \cdots, v_p]$.

Optimal low-rank approximation can be obtained by minimizing the unconstrained cost function [10]

$$F_1(x, y) = \frac{1}{2} tr(A - xy^T)^T (A - xy^T), \tag{1}$$

where $x \in \mathbb{R}^{n\times p}$ and $y \in \mathbb{R}^{m\times p}$. The gradient of F_1 is

$$\nabla F_1 = \frac{1}{2} \begin{bmatrix} -2Ay + 2xy^T y \\ -2A^T x + 2yx^T x \end{bmatrix}. \tag{2a}$$

The set of nonzero equilibrium point of this system consists of points $\hat{x} = U_{\bar{p}}\alpha$ and $\hat{y} = V_{\bar{p}}\beta$ for some nonsingular matrices α and β. Here $U_{\bar{p}} = [u_{i_1}, \cdots, u_{i_p}]$, $V_{\bar{p}} = [v_{i_1}, \cdots, v_{i_p}]$,

978-1-4244-3124-3/08/$25.00 ©2008 IEEE

and $U_{\bar{p}}^T A V_{\bar{p}} = \Sigma_{\bar{p}}$, where $i_1, \cdots, i_p \in \{1, 2, \cdots, m\}$. Moreover, $\Sigma_{\bar{p}}\beta = \alpha\beta^T\beta$ and $\Sigma_{\bar{p}}\alpha = \beta\alpha^T\alpha$. This implies that $\Sigma_{\bar{p}} = \alpha\beta^T = \beta\alpha^T$. There is no guarantee that α or β is diagonal as in the following example. Let $\alpha = \begin{bmatrix} 1 & 1 \\ 3 & 2 \end{bmatrix}$ and $\beta = \begin{bmatrix} 2 & 1 \\ -3 & -1 \end{bmatrix}$. Clearly $a\beta^T = \beta\alpha^T = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}$, but neither α nor β is diagonal.

The corresponding dynamical system

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = -\begin{bmatrix} -Ay + xy^T y \\ -A^T x + yx^T x \end{bmatrix} = \begin{bmatrix} Ay - xy^T y \\ A^T x - yx^T x \end{bmatrix}, \tag{2b}$$

converges to the low-rank approximation of order p for the matrix A. Clearly, the system (2b) is stable since the function $F_1(x, y)$ may be chosen as a Lyapunov function [19]. In this case $F_1(x, y) \geq 0$ and $\dot{F}_1 = -(\nabla F_1)^T(\nabla F_1) \leq 0$. Under mild conditions on the initial matrices, $x(0)$ and $y(0)$, one can show that $x(t) \to \hat{x}$ and $y(t) \to \hat{y}$, where $\hat{x}$ and $\hat{y}$ have same matrix rank as $x(0)$ and $y(0)$, respectively, and $\nabla F_1(\hat{x}, \hat{y}) = 0$. The solution $(\hat{x}, \hat{y})$ is not unique and is dependent on the initial matrices. Clearly, $\hat{x} = U\alpha$ and $\hat{y} = V\beta$ for some nonsingular matrices α and β. After some manipulations, it follows that $(\hat{x}^T\hat{x})^{\frac{-1}{2}}\hat{x}^T A\hat{y}(\hat{y}^T\hat{y})^{\frac{-1}{2}}(\hat{x}^T\hat{x})^{\frac{-1}{2}}\hat{x}^T A\hat{y}(\hat{y}^T\hat{y})^{\frac{-1}{2}} = (\hat{x}^T\hat{x})^{\frac{1}{2}}(\hat{y}^T\hat{y})^{\frac{1}{2}} = (\alpha^T\alpha)^{\frac{1}{2}}(\beta^T\beta)^{\frac{1}{2}}$. Hence the singular values of the matrix $(\hat{x}^T\hat{x})^{\frac{1}{2}}(\hat{y}^T\hat{y})^{\frac{1}{2}}$ are the largest p singular values of the matrix A. However, since the matrices $(\hat{x}^T\hat{x})^{\frac{-1}{2}}\hat{x}^T A\hat{y}(\hat{y}^T\hat{y})^{\frac{-1}{2}}$, $(\hat{x}^T\hat{x})^{\frac{1}{2}}$, and $(\hat{y}^T\hat{y})^{\frac{1}{2}}$ are generally not diagonal, additional computations involving $p \times p$ matrices are needed to determine the singular values. This is summarized in the following result.

Proposition 1. *Let $(x(t), y(t))$ be a solution of (2b) in the interval $[0, \infty)$, where $x(0) = x_0$ and $y(0) = y_0$. Let P, Q, and $\hat{A}$ be defined as $P = \lim_{t\to\infty} x(t)^T x(t)$, $Q = \lim_{t\to\infty} y(t)^T y(t)$, and $\hat{A} = \lim_{t\to\infty} x(t)^T Ay(t)$. Then,*

$$\begin{aligned} \hat{A} &= PQ, \\ \hat{A}^T &= QP, \\ \hat{x}^T AA^T\hat{x} &= PQP, \\ \hat{y}^T A^T A\hat{y} &= QPQ. \end{aligned}$$

Moreover, there exist nonsingular $p \times p$ matrices α and β such that $\Sigma_p = \alpha\beta^T = \beta\alpha^T$, $\alpha^T\Sigma_p\beta = \alpha^T\alpha\beta^T\beta = \alpha^T\beta\alpha^T\beta = (\alpha^T\beta)^2$, and

$$(\alpha^T\alpha)^{\frac{-1}{2}}\alpha^T\Sigma_p\beta(\alpha^T\alpha)^{\frac{-1}{2}} = (\alpha^T\alpha)^{\frac{1}{2}}(\beta^T\beta)^{\frac{1}{2}}.$$

Remark 1: The dynamical system (2b) can also be obtained by maximizing the unconstrained cost function

$$F_2(x, y) = tr(x^T Ay) - \frac{1}{2}tr\{x^T xy^T y\} \tag{3}$$

over full rank matrices $x \in \mathbb{R}^{n\times p}$ and $y \in \mathbb{R}^{m\times p}$.

Remark 2: In (3), if x is chosen to be orthogonal, i.e., $x^T x = I_p$, one may use optimization theory over Stiefel manifold [20] to obtain the dynamical system:

$$\begin{aligned} x' &= Ay - xy^T y - xy^T A^T x + xy^T y, \\ y' &= A^T x - y. \end{aligned} \tag{4}$$

Similarly, if y is chosen to be orthogonal in (3), then we obtain the dynamical system:

$$\begin{aligned} x' &= Ay - x, \\ x' &= A^T x - yx^T x - yx^T Ay + yx^T x. \end{aligned} \tag{5}$$

Let $x(0) = x_0$ and $y(0) = y_0$ are full rank matrices, and let $(x(t), y(t))$ be a solution of (4) or (5) in the interval $[0, \infty)$, and let $\hat{x} = \lim_{t\to\infty} x(t)$ and $\hat{y} = \lim_{t\to\infty} y(t)$, then $\hat{x}^T A\hat{y} = \hat{y}^T\hat{y}$ in (4), or $\hat{x}^T A\hat{y} = \hat{x}^T\hat{x}$ in (5). This shows that $\hat{x}^T A\hat{y}$ is symmetric and positive definite. Thus the dynamical systems (4) and (5) converge to the p largest singular values of A given by the eigenvalues of $(\hat{x}^T\hat{x})^{\frac{1}{2}}$ or $(\hat{y}^T\hat{y})^{\frac{1}{2}}$.

3 Power-Like Methods

Since the matrices $x^T x$ and $y^T y$ are positive definite, it follows from the theory of gradient dynamical systems, that the convergence behavior of the system

$$\begin{aligned} x' &= Ay(y^T y)^{-1} - x, \\ y' &= A^T x(x^T x)^{-1} - y, \end{aligned} \tag{6}$$

are similar to that of the system (2b), i.e., both systems (2b) and (6) have same equilibrium points. Using Euler's method, a discrete version of the system (6) is

$$\begin{aligned} x(k+1) &= x(k) + \gamma\{Ay(k)(y(k)^T y(k))^{-1} - x(k))\}, \\ y(k+1) &= y(k) + \gamma\{A^T x(k)(x(k)^T x(k))^{-1} - y(k)\}, \end{aligned} \tag{7}$$

where $0 < \gamma \leq 1$ is a stepsize. If $\gamma = 1$ is used in (7), the following algorithm is obtained:

$$\begin{aligned} x(k+1) &= Ay(k)(y(k)^T y(k))^{-1}, \\ y(k+1) &= A^T x(k)(x(k)^T x(k))^{-1}. \end{aligned} \tag{8}$$

This is a power-like method which converges from any full rank initial matrices (x_0, y_0). Numerical simulations have indicated that (8) converges even if the initial matrices (x_0, y_0) are not full rank, provided the inverse operations in (8) are replaced with generalized Moore-Penrose inverses. In other words, if $x(t) \to \hat{x}$ and $y(t) \to \hat{y}$, then $\hat{x}$ and $\hat{y}$ have same matrix rank as $x(0)$ and $y(0)$, respectively. In practical implementation of (8), one may start with iteration (7) using $0 < \gamma < 1$ for the first few iterations, then switch to $\gamma = 1$ to speed up convergence.

The solution $(\hat{x}, \hat{y}) = (x(\infty), y(\infty))$ is not unique in that it is dependent on the initial matrices. Additionally, the iteration (8) only produces an arbitrary basis of the p-dimensional principal singular subspace. With a slight modification of (8), this power-like method could produce the actual low rank SVD. The power method for SVD is given as in the following algorithm:

$$\begin{aligned} x_{k+1} &= Ay(k)Tri((y(k)^T y(k))^{-1}), \\ y_{k+1} &= A^T x(k)Tri((x(k)^T x(k))^{-1}). \end{aligned} \tag{9}$$

Here the notation $Tri(X)$ represents the upper triangular part of X, i.e., $X = Tri(X) + L$, where L is lower diagonal matrix with zero elements on its diagonal. Simulations have shown that $x(k)^T x(k)$, $y(k)^T y(k)$ and $x(k)^T Ay(k)$ converge to diagonal matrices as $k \to \infty$, i.e., the system (9) converges to the true singular value components of A.

To prove this property for (9), let $(x(k), y(k))$ be a sequence generated by (9) with initial matrices $(x(0), y(0))$. Assume also that $x(0)^T U_p$ and $y(0)^T V_p$ are nonsingular. Let $P = \lim_{k\to\infty} x(k)^T x(k)$, $Q = \lim_{k\to\infty} y(k)^T y(k)$, and $\hat{A} = \lim_{k\to\infty} x(k)^T Ay(k)$. Assuming that P and Q are invertible, then the matrices P and Q may be expressed as a sum of lower and upper triangular matrices as follows:

$$P^{-1} = U_1 + L_1 = U_1^T + L_1^T,$$

$$Q^{-1} = U_2 + L_2 = U_2^T + L_2^T,$$

where $U_1 = Tri(P^{-1})$ and $U_2 = Tri(Q^{-1})$. The matrices $L_1 = P^{-1} - Tri(P^{-1})$ and $L_2 = Q^{-1} - Tri(Q^{-1})$ are strickly lower triangular. From (9), we have

$$P = \hat{A}U_2,$$
$$Q = \hat{A}^T U_1.$$

Since P and Q are symmetric,

$$P^{-1} = U_2^{-1}\hat{A}^{-1} = \hat{A}^{-T}U_2^{-T},$$
$$Q^{-1} = U_1^{-1}\hat{A}^{-T} = \hat{A}^{-1}U_1^{-T}.$$

Therefore, the following equations hold

$$U_1 + L_1 = \hat{A}^{-T}U_2^{-T},$$
$$U_2 + L_2 = \hat{A}^{-1}U_1^{-T},$$
$$U_1U_2^T + L_1U_2^T = \hat{A}^{-T},$$
$$U_2U_1^T + L_2U_1^T = \hat{A}^{-1}.$$

The last two equations imply that

$$U_1U_2^T + L_1U_2^T = U_1U_2^T + U_1L_2^T,$$

or equivalently,

$$L_1U_2^T = U_1L_2^T.$$

Since $L_1U_2^T$ and $U_1L_2^T$ are upper and lower triangular matrices, respectively, and L_1 and L_2 are strickly lower triangular matrices, then

$$L_1U_2^T = U_1L_2^T = 0.$$

Since U_1 and U_2 are invertible by the assumption that $\hat{A}+\hat{A}^T$ is positive definite, then $L_1 = 0$, and $L_2 = 0$. Consequently, $P = D_1$, $Q = D_2$, and $\hat{A} = D_1D_2$, where D_1 and D_2 are diagonal matrices. Assume that $\hat{x} = \lim_{t\to\infty} x(t)$, $\hat{y} = \lim_{t\to\infty} y(t)$, then $\hat{x} = U_pD_1^{\frac{1}{2}}$, $\hat{y} = V_pD_2^{\frac{1}{2}}$, and $\Sigma_p = D_1^{\frac{1}{2}}D_2^{\frac{1}{2}}$.

Remark 3: Another gradient dynamical system follows from the optimization problem

$$\text{Maximize } F_3(x,y) = tr\{(x^TAy - \frac{1}{2}(x^Tx + y^Ty)^2\}, \tag{10}$$

where $x \in \mathbb{R}^{n\times p}$ and $y \in \mathbb{R}^{m\times p}$. Note that F_3 is a slight modification of F_1. The corresponding gradient dynamical system is modified and given by

$$\begin{aligned} x' &= AyTri((x^Tx + y^Ty)^{-1}) - x, \\ y' &= A^TxTri((x^Tx + y^Ty)^{-1}) - y. \end{aligned} \tag{11}$$

Let $x(t)$ and $y(t)$ be a solution of (11) in the interval $[0, \infty)$, where $x(0) = x_0$ and $y(0) = y_0$ are full rank. Assume that P, Q, and $\hat{A}$ are as defined previously. Then

$$P = \hat{A}U_1,$$
$$Q = \hat{A}^TU_1.$$

where $U_1 = Tri((P+Q)^{-1})$, i.e.,

$$(P+Q)^{-1} = U_1 + L_1 = U_1^T + L_1^T,$$

where L_1 is lower diagonal matrix. This imply

$$P + Q = (\hat{A} + \hat{A}^T)U_1,$$
$$(P+Q)^{-1} = U_1^{-1}(\hat{A} + \hat{A}^T)^{-1} = U_1^T + L_1^T,$$
$$(\hat{A} + \hat{A}^T)^{-1} = U_1U_1^T + U_1L_1^T = U_1U_1^T + L_1U_1^T.$$

Consequently,

$$U_1L_1^T = L_1U_1^T.$$

Since $U_1L_1^T$ and $L_1U_1^T$ are upper- and lower-triangular matrices, respectively, it follows that

$$L_1 = 0, (P+Q)^{-1} = U_1.$$

The symmetry of $P + Q$ yields

$$(P+Q)^{-1} = D.$$

Here D is a diagonal matrix whose diagonal elements are those of $(P+Q)^{-1}$. Now, $D^{-1} = P+Q = (\hat{A}+\hat{A}^T)U_1 = (\hat{A}+\hat{A}^T)D = D(\hat{A}+\hat{A}^T)$. This implies that

$$\hat{A} + \hat{A}^T = D_1 = D^{-2},$$

for some diagonal matrix D_1. To show that $\hat{A}$ is diagonal, we have

$$\hat{A}D = D\hat{A}^T = D(D_1 - \hat{A}),$$

or

$$\hat{A}D + D\hat{A} = DD_1.$$

Therefore,

$$\hat{A} = D_2,$$

for some diagonal matrix D_2. Hence

$$\hat{A} = \frac{D_1}{2},$$

where

$$D_1 = \frac{D^{-2}}{2}.$$

A discrete version of (11) is given as

$$\begin{aligned} x(k+1) &= x(k) + \gamma\{Ay(k)Tri((x(k)^Tx(k) + y(k)^Ty(k))^{-1}) \\ &\quad - x(k)\}, \\ y(k+1) &= y(k) + \gamma\{A^Tx(k)Tri((x(k)^Tx(k) + y(k)^Ty(k))^{-1}) \\ &\quad - y(k)\}, \end{aligned} \tag{12}$$

where $0 < \gamma \le 1$ is a stepsize. When $\gamma = 1$, (12) transforms into a power-like method:

$$\begin{aligned} x(k+1) &= Ay(k)Tri((x(k)^Tx(k) + y(k)^Ty(k))^{-1}), \\ y(k+1) &= A^Tx(k)Tri((x(k)^Tx(k) + y(k)^Ty(k))^{-1}). \end{aligned} \tag{13}$$

4 Diagonalization Using A Weight Matrix

The cost function (1) may be modified so that $(\hat{x}^T\hat{x})^{\frac{-1}{2}}\hat{x}^TA\hat{y}(\hat{y}^T\hat{y})^{\frac{-1}{2}}$, $(\hat{x}^T\hat{x})^{\frac{1}{2}}$, and $(\hat{y}^T\hat{y})^{\frac{1}{2}}$ converge to diagonal matrices. This can be accomplished by incorporating a weight matrix D which is diagonal and all its diagonal elements are distinct. Thus consider the cost function F_4 defined as

$$F_4(x,y) = tr(x^TAyD) - \frac{1}{2}tr\{x^Txy^Ty\}, \tag{14}$$

where $x \in \mathbb{R}^{n\times p}$ and $y \in \mathbb{R}^{m\times p}$. D is a diagonal matrix whose eigenvalues are distinct and positive.

The gradient of F_4 is

$$\nabla F_4 = \begin{bmatrix} AyD - xy^Ty \\ A^TxD - yx^Tx \end{bmatrix}, \tag{15}$$

from which we obtain the gradient dynamical system

$$\begin{aligned} x' &= AyD - xy^Ty, \\ y' &= A^TxD - yx^Tx. \end{aligned} \tag{16}$$

If $x(0)$ and $y(0)$ are full rank, the system (16) converges to the low-rank approximation of order p for the matrix A. Clearly, the system (16) is stable since the function $-F_4(x,y)$ is bounded below and radially unbounded. The main difference between the systems (2b) and (16) is that the one in (16) converges to the true singular triplets.

Let $x(t)$ and $y(t)$ be a solution of (16) in the interval $[0, \infty)$, where $x(0) = x_0$ and $y(0) = y_0$ are full rank. Let $P = \lim_{t\to\infty} x(t)^T x(t)$, $Q = \lim_{t\to\infty} y(t)^T y(t)$, and $\hat{A} = \lim_{t\to\infty} x(t)^T Ay(t)$. Note that P, Q, and $\hat{A}$ exists since the system (16) is stable. Then

$$\hat{A}D = PQ,$$
$$\hat{A}^T D = QP.$$

Since P and Q are symmetric, then

$$\hat{A}D = D\hat{A}.$$

From the assumption that all eigenvalues of D are distinct, it follows from Proposition 2 that $\hat{A} = D_1$ for some diagonal matrix D_1. Consequently, $PQ = QP = D_1$. If all eigenvalues of $\hat{A}$ are distinct, Proposition 4 guarantees that

$$P = D_3, Q = D_4,$$

for some diagonal matrices D_3 and D_4. This means that P, Q and $\hat{A}$ are diagonal and therefore, $U_p = \hat{x}D_3^{\frac{-1}{2}}$, $V_p = \hat{y}D_4^{\frac{-1}{2}}$, and $\Sigma_p = D_3^{\frac{-1}{2}}\hat{x}^T A\hat{y}D_4^{\frac{-1}{2}}$.

One may use the equation $\nabla F_4 = 0$ to derive the following power-like method:

$$\begin{aligned} x_{k+1} &= Ay(k)D(y(k)^T y(k))^{-1}, \\ y_{k+1} &= A^T x(k)D(x(k)^T x(k))^{-1}. \end{aligned} \tag{17}$$

Let $\{(x(k), y(k)\}_{k=0}^{\infty}$ be a sequence generated by the system (17) where $x(0) = x_0$ and $y(0) = y_0$ are given to be full rank. Also, let $P = \lim_{t\to\infty} x(t)^T x(t)$, $Q = \lim_{t\to\infty} y(t)^T y(t)$, and $\hat{A} = \lim_{t\to\infty} x(t)^T Ay(t)$. Then, (17) implies that

$$\hat{A}D = PQ,$$
$$\hat{A}^T D = QP.$$

The last two equations yield

$$\hat{A}D = D\hat{A}.$$

From the assumption that all eigenvalues of D are distinct, it follows from Proposition 2 that $\hat{A} = D_1$ for some diagonal matrix D_1. Hence,

$$PQ = QP = D_1.$$

If all diagonal elements of D_1 are distinct, then

$$P = D_3, Q = D_4,$$

and hence,

$$\hat{x} = U_p D_3^{\frac{1}{2}},$$
$$\hat{y} = V_p D_4^{\frac{1}{2}}.$$

Remark 4: Another gradient dynamical system follows from the optimization problem

$$\text{Maximize } F_5(x,y) = tr\{(x^T AyD - \frac{1}{2}(x^T x + y^T y)^2\}, \tag{18}$$

where $x \in \mathbb{R}^{n\times p}$ and $y \in \mathbb{R}^{m\times p}$. Here D is a diagonal matrix and all its eigenvalues are distinct. Note that F_5 is a slight modification of F_3. The corresponding gradient dynamical system is given by

$$\begin{aligned} x' &= AyD - x(x^T x + y^T y), \\ y' &= A^T xD - y(x^T x + y^T y). \end{aligned} \tag{19}$$

Let $\hat{x} = \lim_{t\to\infty} x(t)$, $\hat{y} = \lim_{t\to\infty} y(t)$, and let $P, Q, \hat{A}$ be as defined above. Then

$$\begin{aligned} \hat{A}D &= P(P+Q), \\ \hat{A}^T D &= Q(P+Q). \end{aligned} \tag{20}$$

The equations of (20) imply that

$$\hat{A}\hat{A}^{-T} = PQ^{-1}.$$

Let R be a matrix defined by $R = PQ^{-1}$, then

$$\hat{A} = R\hat{A}^T. \tag{21a}$$

Equation (21a) yields

$$\hat{A}(I - R^T) + (I - R)\hat{A}^T = 0. \tag{21b}$$

The next step is to show that $R = I$ under the assumption that $\hat{A} + \hat{A}^T$ is positive definite. Since P and Q are positive definite, then the eigenvalues of R are all real with corresponding real eigenvectors. Thus assume that λ is an eigenvalue of R^T with associative right eigenvector z, then $R^T z = \lambda z$. Pre- and post-multiplying the left and right sides of the equation (21b) by z^T and z respectively, give

$$z^T\hat{A}z(1-\lambda) + z^T\hat{A}^T z(1-\lambda) = 0,$$

and hence

$$(1-\lambda)\{z^T\hat{A}z + z^T\hat{A}^T z\} = 0.$$

Since $\hat{A} + \hat{A}^T$ is positive definite by assumption, it follows that $\lambda = 1$. i.e., each eigenvalue of R is equal to 1. The eigenvalues of $PQ^{-1} = R$ are same as those of $P^{\frac{1}{2}}Q^{-1}P^{\frac{1}{2}}$. Since each eigenvalue of the symmetric matrix $P^{\frac{1}{2}}Q^{-1}P^{\frac{1}{2}}$ is 1, then $P^{\frac{1}{2}}Q^{-1}P^{\frac{1}{2}} = I$ and hence $P = Q$, $R = I$. This shows that $\hat{A}$ is symmetric. Now, the equations of (20) simplify to $\hat{A}D = 2P^2 = D\hat{A}$. Since all eigenvalues of D are distinct, it follows from Proposition 2 that $\hat{A} = D_1$, and consequently, $P = Q = \frac{\sqrt{DD_1}}{2}$. This shows that P, Q, and $\hat{A}$ are diagonal.

5 Numerical Experiments

In order to verify that the dynamical systems and power-like methods, which are proposed in the previous sections, converge to the true singular vectors of the data matrix A, a few numerical examples are provided. The first experiment is to compute the largest six singular values of a matrix A of size 70×70 using the power like method (9). The matrix A is generated randomly with singular values (in decreasing order) as given.
14.5318 14.2656 13.7690 13.3242 13.0242 12.5197 12.5064 12.1766 11.9105 11.6099 11.3957 11.2249 10.8717 10.6580 10.5067 10.3144 10.1320 9.8869 9.6993 9.3254 9.0150 8.7069 8.5200 8.3468 8.2616 7.8747 7.8367 7.5584 7.3931 7.2216 7.0510 6.9850 6.7580 6.4006 6.2288 6.0383 5.8459 5.5437 5.4306 5.2177 5.0972 4.9270 4.8632 4.7497 4.4700 4.1268 4.0707 3.8745 3.6793 3.2429 3.0531 2.9417 2.7391 2.5303 2.3154 2.1925 2.1571 1.9277 1.6832 1.3921 1.1618 1.0600 0.9830 0.8741 0.5887 0.4864 0.3642 0.1808 0.0000 0.0000.

Note there is very little separation between two adjacent singular values. Figure 1 shows the convergence to the six largest singular values of A. As can be seen in the Figure, larger singular values require less number of iterations to converge. We also examined the matrices $x^T x, y^T y$ and $x^T Ay$ and noted that they are diagonal after convergence. The number of iterations was 920.

The values to which the algorithm converges to are: 14.5318, 14.2656, 13.7690, 13.3242, 13.0242, 12.2793.

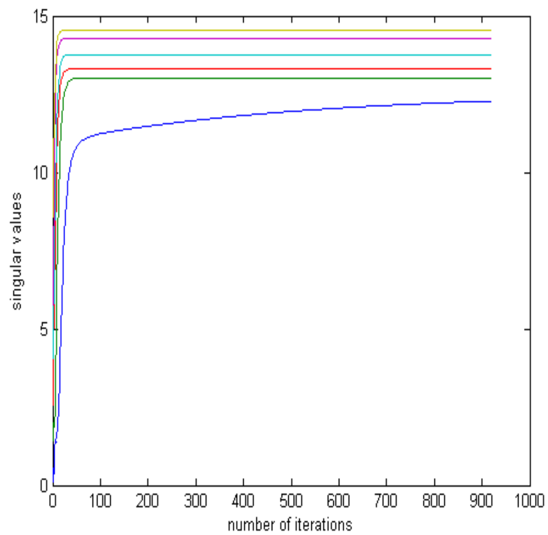

Figure 1: A plot showing the number of iterations versus six singular value approximation. These are obtained via algorithm (11)

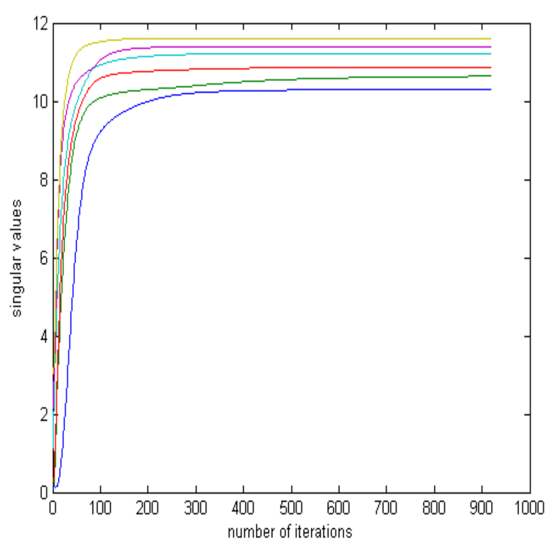

Figure 2: A plot showing the number of iterations versus six singular value approximation. These are obtained via algorithm (19)

The second experiment involves computing the largest six singular values of a 70×70 randomly generated matrix B with singular values (in decreasing order) are as given below.

11.6099 11.3957 11.2249 10.8717 10.6580 10.5067 10.3144 10.1320 9.8869 9.6993 9.3254 9.0150 8.7069 8.5200 8.3468 8.2616 7.8747 7.8367 7.5584 7.3931 7.2216 7.0510 6.9850 6.7580 6.4006 6.2288 6.0383 5.8459 5.5437 5.4306 5.2177 5.0972 4.9270 4.8632 4.7497 4.4700 4.1268 4.0707 3.8745 3.6793 3.2429 3.0531 2.9417 2.7391 2.5303 2.3154 2.1925 2.1571 1.9277 1.6832 1.3921 1.1618 1.0600 0.9830 0.8741 0.5887 0.4864 0.3642 0.1808 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000

This experiment is carried out with 6-dimensional vector using the power-like method (19). The stepsize is $\gamma = 0.09$ and matrix D is given below:

```
D=
0.7635         0         0         0         0         0
     0    1.3055         0         0         0         0
     0         0    0.8288         0         0         0
     0         0         0    0.9740         0         0
     0         0         0         0    0.8840         0
     0         0         0         0         0    0.4626
```

Figure 2 shows the convergence to the six largest singular values of B. As can be seen in the Figure, larger singular values require less number of iterations to converge. This algorithm converges to the following values. $\{11.6099, 11.3957, 11.2249, 10.8707, 10.6350, 10.3149\}$.

We also examined the matrices $x^T x, y^T y$ and $x^T Ay$ and noted that the off-diagonal elements are small in magnitude but are not substantially close to zero as those in Experiment 1. In both experiments, $x(0)$ and $y(0)$ are randomly generated.

6 Conclusion

A number of principal singular subspace methods are derived and analyzed. These methods are based on dynamical systems which are derived using constrained and unconstrained optimization methods. Different dynamical systems are obtained by weighting a given system with a diagonal matrix, or by using upper triangular matrices. Some of the proposed flows generalize Oja's principal component flow and other known flows for singular value decomposition. Further analysis is needed to explore numerical stability and convergence. Extension of the proposed rules to complex data and matrices can be achieved with minor modifications.

7 Appendix

Finally, we state a few results which are essential for the derivations of the proposed methods.

Proposition 2. *Let $D, C \in \mathbb{R}^{n\times n}$ such that D is diagonal having distinct eigenvalues. If $CD = DC$, then C is diagonal.*

Proposition 3. *Let $A, B \in \mathbb{R}^{n\times n}$ be real matrices such that $A^T = A$, and all eigenvalues of A are distinct. If $AB = BA$, then $B^T = B$.*

Proof. Post-multiplying both sides of the equation $AB = BA$ by B^T, yields

$$ABB^T = BAB^T.$$

Thus ABB^T is symmetric, i.e.,

$$ABB^T = BB^T A.$$

Let

$$A = Z\Sigma_1 Z^T,$$

where Z is orthogonal and Σ_1 is diagonal. This implies that

$$Z\Sigma_1 Z^T BB^T = BB^T Z\Sigma_1 Z^T,$$

or equivalently,

$$\Sigma_1 Z^T BB^T Z = Z^T BB^T Z\Sigma_1.$$

Since all eigenvalues of Σ_1 are distinct, Proposition 2 guarantees that $Z^T BB^T Z = \Sigma_2^2$, where Σ_2 is diagonal. Hence

$$BB^T = Z\Sigma_2^2 Z^T,$$

and therefore,

$$B = Z\Sigma_2\alpha,$$

for some orthogonal matrix α, i.e., $\alpha^T\alpha = I_p$. Now $AB = BA$ implies that

$$Z\Sigma_1 Z^T Z\Sigma_2\alpha = Z\Sigma_2\alpha Z\Sigma_1 Z^T,$$

or

$$\Sigma_1\alpha Z = \alpha Z\Sigma_1.$$

Since all eigenvalues of Σ_1 are distinct, Proposition 2 guarantees that $\alpha Z = \Sigma_3$, where Σ_3 is diagonal. Note that $\alpha Z = \Sigma_3$ is orthogonal matrix and thus

$$\Sigma_3^2 = I.$$

The matrix α is then determined as

$$\alpha = \Sigma_3 Z^T,$$

and consequently,

$$B = Z\Sigma_2\Sigma_3 Z^T.$$

Thus B is symmetric.

Proposition 4. *Let $A, B, D \in \mathbb{R}^{n\times n}$ be symmetric matrices such that D is diagonal and $AB = D$. If all eigenvalues of D are distinct, then A and B are diagonal.*

Proof. Clearly, $AD = A^2B = BA^2 = DA$ and $BD = B^2A = AB^2 = DB$. Since all eigenvalues of D are distinct, Proposition 2 implies that A and B are diagonal.

References

[1] A. Cichocki, "Neural network for singular value decomposition," Electron. Lett., vol. 28, pp. 784-786, 1992.

[2] A. Cichocki and R. Unbehauen, Neural Networks for Optimization and Signal Processing. New York: Wiley, 1994.

[3] H. Chen and R.-W. Liu, "An online unsupervised learning machine for adaptive feature extraction," IEEE Trans. Circuits Syst., vol. 41, pp.87-98, Feb. 1994.

[4] U. Helmke and J. B. Moore, "Singular value decomposition via gradient and self-equivalent flows," Linear Algebra Appl., vol. 169, pp. 223-248, 1992.

[5] J. B. Moore, R. E. Mahony, and U. Helmke, "Numerical gradient algorithms for eigenvalues and singular value calculations," SIAM J. Matrix Anal. Appl., vol. 15, pp. 881902, 1994.

[6] S. T. Smith, "Dynamic system that perform the singular value decomposition," Syst. Control Lett., vol. 15, pp. 319-327, 1991.

[7] M. T. Chu and K. R. Dressel, "The projected gradient method for least squares matrix approximations with spectral constraints," SIAM J. Numer. Anal., vol. 27, pp. 1050-1060, 1990.

[8] Hori, G., "A general framework for SVD flows and joint SVD flows," Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003 (ICASSP '03), Volume: 2, 6-10, April 2003, pp:II-693-696.

[9] P. Strobach, "Bi-iteration SVD subspace tracking algorithms," IEEE Trans. Signal Processing, vol.45, no.5, pp.1222-1240, 1997.

[10] Shan Ouyang; Yingbo Hua, "Bi-iterative least square versus bi-iterative singular value decomposition for subspace tracking," Proceedings, Acoustics, Speech, and Signal Processing, 2004 (ICASSP '04), Volume: 2, 17-21 May 2004, pp- 353-356.

[11] A. Cichocki, "Neural network for singular value decomposition," Electron. Lett., vol. 28, pp. 784786, 1992.

[12] E. Kokiopoulou, C. Bekas, E. Gallopoulos, "Computing smallest singular triplets with implicitly restarted Lanczos bidiagonalization," Applied Numerical Mathematics archive, Volume 49 , Issue 1, April 2004, Pages: 39-61

[13] A. L. Yuile, D. M. Kammen, and D. S. Cohen, Quadrature and develop-ment of orientation selective cortical cells by Hebb rules, Biol. Cybern., vol. 61, pp. 183194, 1989.

[14] N. Samardzija and R. L. Waterland, "A neural network for computing eigenvectors and eigenvalues," Biol. Cybern., vol. 65, no. 4, pp. 211214.

[15] K. I. Diamantaras and S. Y. Kung, "Cross-correlation neural network models," IEEE Trans. Signal Processing, vol. 42, pp. 32183223, Nov. 1994.

[16] Da-Zheng Feng; Xian-Da Zhang; Zheng Bao, "A neural network learning for adaptively extracting cross-correlation features between two high-dimensional data streams," IEEE Transactions on Neural Networks, Volume: 15, Issue: 6, pp. 1541-1554, Nov. 2004.

[17] Shan Ouyang; Zheng Bao; Gui-Sheng Liao; Ching, P.C., "Adaptive minor component extraction with modular structure," IEEE Transactions on Signal Processing, Volume: 49, Issue: 9, pp:2127- 2137, Sept. 2001.

[18] Hasan, M.A., "Constrained quadratic optimization problems with applications," Proceedings of the American Control Conference, June 8-10, 2005 Page(s):243-248.

[19] J. J. E. Slotine and W. Li. Applied nonlinear Comtrol. Prentice Hall, 1991.

[20] A. Edelman, T. A. Arias and S. T. Smith, "The geometry of algorithms with orthogonality constraints" SIAM J. Matrix Anal. Appl., 20(2):303-353, 1998.