Adaptive Machine Learning Algorithms

Proofs of Convergence

Chanchal Chatterjee
14521 Weeth Drive
San Jose, CA 95124

Vwani P. Roychowdhury
6371C Boelter Hall,
UCLA, Los Angeles, CA 90095

August 30, 2021

Table of Contents

Adaptive Machine Learning Algorithms 1
Proofs of Convergence 1
1. Introduction to the Proofs of Convergence 4
1.1 Objective Functions for Adaptive Algorithms 4
1.2 Common Methodology for Derivations and Convergence Proofs 6
1.3 Stochastic Approximation Theory 7
1.4 Solution of Riccati Differential Equation 11
2. Proofs for Chapter 2 12
2.1 Convergence Proof for Adaptive Stationary Mean 12
2.2 Convergence Proof for Adaptive Stationary Correlation 12
2.3 Adaptive Normalized Mean Algorithm 13
2.4 Convergence Proof of Adaptive Median 14
2.5 Brief Review of Optimization Theory 15
2.6 Matrix Operations 17
2.7 Derivatives of Matrix Functions 18
3. Proofs for Chapter 3 19
3.2 Adaptive Square Root Algorithm – Method 1 19
3.3 Adaptive Square Root Algorithm – Method 2 21
3.4 Adaptive Square Root Algorithm – Method 3 22
3.5 Adaptive Inverse Square Root Algorithm – Method 1 23
3.6 Adaptive Inverse Square Root Algorithm – Method 2 24
3.7 Adaptive Inverse Square Root Algorithm – Method 3 25
4. Proofs for Chapter 4 27
4.2. Algorithms and Objective Functions 27
4.3. OJA Algorithm 27
4.4. RQ, OJAN, and LUO Algorithms 30
4.5. IT Algorithm 34
4.6. XU Algorithm 35
4.7. Penalty Function (PF) Algorithm 37
4.8. Augmented Lagrangian 1 (AL1) Algorithm 39
4.9. Augmented Lagrangian 2 (AL2) Algorithm 40
5. Proofs for Chapter 5 43
5.2. Algorithms and Objective Functions 43

- 5.5. PF Algorithms ... 43
- 5.7. AL2 Algorithm ... 49

6. Proofs for Chapter 6 ... 54

7. Proofs for Chapter 7 ... 54
- 7.2 Algorithms and Objective Functions ... 54
- 7.4 XU GEVD Algorithms ... 55

8. Proofs for Chapter 8 ... 60

1. Introduction to the Proofs of Convergence

1.1 Objective Functions for Adaptive Algorithms

In order to compute matrix functions by adaptive algorithms, we need a convenient methodology or strategy to derive them. One such approach is to identify an *objective or cost function J*, which is a function of a parameter matrix *W* and the asymptotic matrix *A*. Given an objective function *J*(*W*;*A*), we minimize it to obtain W^* as:

$$W^* = \arg\min_{W} J(W;A) \tag{1.1}$$

where W^* is the desired matrix function of *A*. For example, if we are computing the inverse square root ($A^{-½}$) of *A*, then $W^* = A^{-½}$. A crucial step in this strategy is to identify an objective function *J*(*W*;*A*) such that its minimizer W^* is the desired matrix function of *A*. Obtaining such objective functions offers several benefits as described below.

Derive Adaptive Algorithms

Since the minimizer W^* of the objective function *J*(*W*;*A*) is the desired matrix function of *A*, we can obtain several adaptive algorithms to compute W^* by applying standard optimization techniques [Luenberger 84] to the objective function *J*(*W*;*A*). Examples of optimization techniques are: (1) gradient descent, (2) steepest descent, (3) conjugate direction, (4) Newton-Raphson, and (5) Recursive Least Squares (RLS). An adaptive algorithm for a data sequence $\{A_k\}$ can be obtained by replacing *A* with A_k, although a rigorous convergence proof is necessary for this substitution.

For example, the standard gradient descent algorithm for minimizing *J*(*W*;*A*) is:

$$W_{k+1} = W_k - \eta_k \nabla_W J(W_k;A)\,, \tag{1.2}$$

where η_k is a positive gain sequence, and W_0 is a starting matrix. In order to obtain an adaptive algorithm that uses the online data sequence $\{A_k\}$ instead of its asymptotic value *A*, we modify (1.2) as:

$$W_{k+1} = W_k - \eta_k \nabla_W J(W_k;A_k)\,. \tag{1.3}$$

While the convergence of (1.2) is well-established, the convergence of (1.3) is not guaranteed by classical optimization theory. Hence, we use an alternative method due to Stochastic Approximation Theory [Kushner&Clark 78; Ljung 77,78,84,92; Benveniste *et al.* 90] to prove the convergence of (1.3).

Speedup of Adaptive Algorithms

The availability of the objective function *J*(*W*;*A*) allows us to speed up the basic gradient descent stochastic approximation algorithm (1.3). By applying speedup techniques in optimization theory such as steepest descent, conjugate direction, Newton-Raphson and recursive least squares, we can enhance the speed of the stochastic approximation algorithm (1.3). Details of these methods for principal component analysis are given in Chapter 6.

Convergence Analysis

Although the adaptive algorithms are derived by following standard optimization techniques on the objective function *J*(*W*;*A*), their derivations do not constitute a proof of convergence. For example, algorithm (1.3) is obtained from (1.2). Although the convergence of (1.2) is guaranteed by standard optimization theory, the convergence of (1.3) should be proven rigorously. Hence, it is important to provide a convergence analysis for our adaptive algorithms.

In simple terms, our derivations of the algorithms from objective function *J*(*W*;*A*) show that the descent direction of *J* is the same as the average evolution direction of the adaptive algorithms. We still need to show that the global minimum of the objective function *J*(*W*;*A*) is the desired matrix function of *A*. Hence, a study of the landscape of the objective function is necessary to determine the convergence of the stochastic approximation algorithm (1.3). The objective function offers an energy or Lyapunov function for such analysis.

The stationary points *W** of the objective function *J*(*W*;*A*) are given by the *Kuhn-Tucker conditions*:

$$\nabla_W J(W^*;A) = 0\,. \tag{1.4}$$

In almost all our algorithms, this equation (1.4) is same as equating the ODE (see (1.7) below) to 0. Some of these stationary points may be stable equilibrium or local minimum points, whereas some may be unstable equilibrium or saddle points. In order to determine the stable equilibrium points, we compute the *Hessian H* of the objective function at the equilibrium points *W**:

$$H = \nabla^2_{WW} J(W^*;A)\,. \tag{1.5}$$

Stable equilibrium points that are local minima have positive definite Hessians, whereas unstable equilibrium points have indefinite Hessians, and local maxima have negative definite Hessians.

1.2 Common Methodology for Derivations and Convergence Proofs

The literature for adaptive algorithms for matrix computation offers a wide range of techniques (including ad hoc methods) and various types of convergence procedures. In this study, we present a *common methodology* to derive and prove the convergence of our adaptive algorithms.

In the stochastic approximation (adaptive) algorithm (1.3), for each time instant k, let W_k be an instantaneous estimate of the desired matrix function W of the asymptotic matrix A. Thus, if we are computing the inverse square root ($A^{-½}$) of A, then W_k is an instantaneous estimate of $A^{-½}$ at time instant k. For each algorithm, we offer an update rule for W_k for each new observation A_k, such that W_k converges to the desired matrix function of A. Steps for the derivation and convergence analysis of each adaptive algorithm is discussed below:

1. We first present an *Objective Function* $J(W;A_k)$ such that the minimizer W^* of J is the desired matrix function of the asymptotic data matrix A.
2. Derive an *Adaptive Update Rule* for matrix W by applying the gradient descent technique on the objective function $J(W;A_k)$. Note that other methods of nonlinear optimization such as steepest descent, conjugate direction, Newton-Raphson or Recursive Least Squares (RLS) can also be used on the objective function. The adaptive gradient descent update rule is:
$$W_{k+1} = W_k - \eta_k \nabla_W J(W_k, A_k) = W_k + \eta_k h(W_k, A_k)\,, \tag{1.6}$$
where the function $h(W_k,A_k)$ follows certain continuity and regularity properties, and η_k is a decreasing gain sequence.
3. In order to prove the convergence of W_k in (1.6) to the desired matrix function of A, we employ the well-known theory of *Stochastic Approximation* [Kushner&Clark 78; Ljung 77,78,84,92; Benveniste *et al.* 90]. In most instances, we assume a stationary data sequence $\{A_k\}$ in order to prove the convergence of the adaptive algorithm by stochastic approximation theory, although practical implementations of the same algorithms on non-stationary sequences can be achieved and yield good results.

4. As an intermediate step of the Stochastic Approximation convergence analysis, we obtain an *Ordinary Differential Equation (ODE)* $dW(t)/dt = \lim_{k\to\infty} E[h(W, A_k)]$, where $W(t)$ is the continuous time counterpart of W_k. The stable stationary solution of the ODE is a convergence point of the adaptive algorithm (1.6). In solving the ODE, we use the well-known solution for the continuous time *Riccati Differential Equation* [Anderson&Moore 90]. If it is difficult to solve the ODE, we study the landscape of the corresponding Lyapunov function $J(W;A)$. We discuss its stable stationary points and show the convergence properties of the ODE for initial states close to the stable points.
5. In order to estimate the *rate of convergence* of the stochastic approximation algorithm (1.6), we compute the time constants (τ) for some of our algorithms. If we can solve the ODE, we get a solution for $W(t,W(0))$ as a function of $e^{-t/\tau}$. Comparatively smaller time constants (τ) indicate faster convergence of the adaptive algorithm.

In the following two sections, we briefly describe the Stochastic Approximation Theory and a solution to the Riccati Differential Equation.

1.3 Stochastic Approximation Theory

An important property of the algorithms discussed here is that they are stochastic approximation algorithms i.e., they are adaptive and of the type (1.6). In contrast to conventional processes, our data arrives in temporal succession i.e., in vector or matrix sequences $\{\mathbf{x}_k\}$ or $\{A_k\}$, instead of vectors or matrices $\mathbf{x}$ or A respectively. For every data sample $\mathbf{x}_k$ or A_k, we update the estimates of the targeted parameters W_k, and we require that the estimates converge strongly to the desired solution. We also require that the statistical procedure keep pace with the incoming data so that, at any instant, the estimates fully reflect all of the currently available data. Such stochastic approximation procedures are also useful when a given estimate has to adapt to small changes in the data (e.g., a few incoming samples). Thus, if the features are computed with conventional methods from some initial samples, then these estimates can be used as starting values for the stochastic approximation procedure when a few additional samples are available. In this situation, the adaptive techniques allow direct updating in a computationally inexpensive way.

An important advantage of stochastic approximation algorithms is that there is a significant amount of theory associated with them, which can be used to analyze and prove their convergence. We use the stochastic approximation theory due to Ljung [Ljung 77,78,84,92; Benveniste *et al*. 90] which deals with stationary sequences. An alternative proof by a similar approach due to Kushner and Clark [Kushner&Clark 78] can also be used. In a somewhat looser language, stochastic approximation theory states the following:

1. Matrix W_k can converge only to stable stationary points of the Ordinary Differential Equation (ODE):

   $$\frac{dW}{dt} = \lim_{k\to\infty} E\left[h(W, A_k)\right], \tag{1.7}$$

   where $W(t)$ is the continuous time counterpart of W_k with t denoting continuous time.
2. If W_k belongs to the domain of attraction of a stable stationary point W^* of the ODE infinitely often with probability one (w.p.1), then W_k converges w.p.1 to W^* as $k \to \infty$.
3. The trajectories of the ODE are the "asymptotic paths" of W_k generated by (1.6). The convergence proof requires the following steps:
   a. defining a set of assumptions,
   b. finding the stable stationary points of the ODE, and
   c. showing that W_k visits the domain of attraction of a stable stationary point infinitely often.

Assumptions

In order to prove the convergence of (1.6), we use Theorem 1 of Ljung [Ljung 77; Benveniste *et al*. 90]. The following is a general set of assumptions for the convergence proof of algorithm (1.6):

Assumption (A1.1). The sequence {$\mathbf{x}_k$} consists of real random vectors, where each $\mathbf{x}_k$ is uniformly bounded with probability one (w.p.1); i.e., $\|\mathbf{x}_k\| < \alpha < \infty$, and $\lim_{k\to\infty} E[\mathbf{x}_k\mathbf{x}_k^T] = A$ where A is positive definite.

Assumption (A1.2). The gain sequence {$\eta_k \in \Re^+$} is decreasing such that $\sum_{k=0}^{\infty}\eta_k = \infty$, $\sum_{k=0}^{\infty}\eta_k^r < \infty$ for some $r > 1$, and $\lim_{k\to\infty}\sup(\eta_k^{-1} - \eta_{k-1}^{-1}) < \infty$.

Assumption A2.1 is reasonable for most practical implementations, where {$\mathbf{x}_k$} are kept bounded either by deliberate measures or automatically. Methods to keep {$\mathbf{x}_k$} bounded are discussed in Ljung [Ljung 77]. The physical meaning of A2.2 can be described as follows. Condition $\eta_k \to 0$ allows the process to settle down in the limit whereas, $\sum_{k=0}^{\infty}\eta_k = \infty$ insures that there is enough corrective action to avoid stopping short of the solution. Conditions $\sum_{k=0}^{\infty}\eta_k^r < \infty$ and $\lim_{k\to\infty}\sup(\eta_k^{-1} - \eta_{k-1}^{-1}) < \infty$ guarantee that the variance of the accumulated noise is finite so that we can correct for the effect of noise. Assumption A2.2 holds for η_k = $ck^{-\delta}$ where $c \in \Re^+$,and 0<$\delta \leq$ 1. The choice of δ= 1 is a leading case. Another alternative used in this study extensively is η_k = $c_1(k+c_2)^{-1}$, where c_1, $c_2 \in \Re^+$.

In the literature for stochastic approximation proofs, there are many assumptions that are usually made on the statistical properties of {$\mathbf{x}_k$}, such as statistically independent and independent and identically distributed (i.i.d). However, Ljung [Ljung 77; Benveniste *et al*. 90] permits far less restrictive choices for {$\mathbf{x}_k$}. Specifically, we can assume that {$\mathbf{x}_k$} is generated by a linear structure:

$$\tilde{\mathbf{x}}_k = A(W_k)\tilde{\mathbf{x}}_{k-1} + B(W_k)\mathbf{e}_k \text{ and } \mathbf{x}_k = C(W_k)\tilde{\mathbf{x}}_k, \tag{1.8}$$

or a nonlinear variant:

$$\tilde{\mathbf{x}}_k = g(k, W_k, \tilde{\mathbf{x}}_{k-1}, \mathbf{e}_k) \text{ and } \mathbf{x}_k = h(k, W_k, \tilde{\mathbf{x}}_{k-1}) \tag{1.9}$$

has also been postulated [Ljung 77]. Here {$\mathbf{e}_k$} is a uniformly bounded sequence of independent (not necessarily stationary or with zero means) random vectors. These structures are treated at length in [Ljung 77,92; Benveniste *et al*. 90]. In light of these models, we state the following general assumption for {$\mathbf{x}_k$}:

> **Assumption (A1.3).** Sequence {$\mathbf{x}_k$} is generated by (1.8) or (1.9) satisfying the stability and regularity conditions of Ljung [Ljung 77].

The main assumptions of Theorem 1 of Ljung [Ljung 77] are:

> **L1.** The function $h(W,A)$ is continuously differentiable with respect to W and A. The derivatives are, for fixed W and A, bounded in k.
>
> **L2.** The so-called *mean vector field* $\bar{h}(W) = \lim_{k\to\infty} E[h(W, A_k)]$ exists and is regular; i.e., locally Lipschitz. The expectation is with respect to the distribution of A_k for a fixed value of W.

Formulation and Solution of the ODE

We modify the results given by Ljung [Ljung 77] in Theorem 1 to suit our adaptive algorithms in the following Theorem:

Theorem 1.1. Let A2.1-A2.3 hold. Let W* be a locally asymptotically stable (in the sense of Liapunov) solution for the ordinary differential equation (ODE):

$$\frac{dW}{dt} = \bar{h}(W) = \lim_{k\to\infty} E[h(W, A_k)] \tag{1.10}$$

with domain of attraction $D(W^*)$. If there is a compact subset $S \subset D(W^*)$ such that $W_k \in S$ infinitely often, then we have $W_k \to W^*$ with probability one as $k \to \infty$.

Proof. The proof can be obtained by using Ljung's Theorem 1 [Ljung 77]. ■

The solution for the ODE (1.10) varies from problem to problem. However, there are two common methods of analyzing the ODE. These are:

1. **Global Analysis:** Formulate the ODE as a Riccati equation [Anderson&Moore 90] and use its well-known solution, whereby we obtain the continuous time solution $W(t,W_0)$, where W_0 is the initial condition at t=0. This is discussed in detail in Section 1.4. We then obtain the asymptotically stable solution $W^*=\lim_{t\to\infty}W(t,W_0)$ where W^* is the desired matrix function of A. Conditions for the existence of the solution for $t \geq 0$ gives us the domain of attraction $D(W^*)$. This explicit solution is useful in studying the global convergence properties of the ODE (1.10) for initial states far from the equilibrium solutions.
2. **Local Analysis:** When it is difficult to directly solve the ODE, we analyze the solution in two steps. We first compute all the equilibrium points of the ODE from the equation $\bar{h}(W)=0$. We then prove that all these equilibrium points of the ODE are unstable equilibrium points, except for the desired matrix function W^*, which is the stable equilibrium point. This analysis allows us to study convergence properties of the ODE (1.10) for initial states close to the stable equilibrium points.

We use both these methods (as needed) to solve the ODE corresponding to the adaptive algorithms presented here.

W_k Visits S Infinitely Often

Although the above analyses give us the limiting values of the ODE, they are not sufficient for the complete convergence proof. One must, in addition, prove that the adaptive algorithm (1.6) is stable; i.e., the weight matrix W_k must remain bounded on some realistic conditions. Such boundedness condition is also necessary for W_k to visit a compact subset S of the domain of attraction of W^* infinitely often. In practical implementation, we can hard-limit the entries of W_k so that their magnitudes remain below a certain limit ρ and thus within a compact region *A.* An alternative analytical approach also exists. It turns out that, for most of our algorithms, there exists a uniform upper bound for η_k such that W_k is uniformly bounded.

The final convergence of the adaptive algorithm (1.6) is guaranteed by Theorem 1 of Ljung and stated in Theorem 2.1.

1.4 Solution of Riccati Differential Equation

Consider the Riccati differential equation (RDE) [Anderson&Moore 90] with time-varying coefficient matrices:

$$\frac{dP(t)}{dt} = PF + F^T P - PGP + Q\,,\ P(0) = P_0\,. \tag{1.11}$$

Associated with the RDE is the linear differential equation:

$$\begin{bmatrix} \dot{X} \\ \dot{Y} \end{bmatrix} = \begin{bmatrix} -F & G \\ Q & F^T \end{bmatrix} \begin{bmatrix} X \\ Y \end{bmatrix},\ \begin{bmatrix} X(0) \\ Y(0) \end{bmatrix} = \begin{bmatrix} I \\ P_0 \end{bmatrix}. \tag{1.12}$$

Some important properties regarding (1.11) and (1.12) are summarized in the Theorem below:

Theorem 2.2. Consider two initial value problems (1.11) and (1.12). Then, the solution of (1.11) exists on [0,T) if and only if X(t) is nonsingular on [0,T). Moreover, the solution to (1.11) is unique and is given by:

$$P(t) = Y(t)X(t)^{-1}\,. \tag{1.13}$$

Exponential Formula for Time-Invariant Problem

Suppose that *F*, *G*, and *Q* are constant matrices or scalars. One can define the so-called *Hamiltonian* matrix *H* as:

$$H = \begin{bmatrix} -F & G \\ Q & F^T \end{bmatrix}. \tag{1.14}$$

It has no imaginary eigenvalue, given detectability and stabilizability, and if λ is an eigenvalue, so is $-\lambda$. Thus, there exists a real *M* such that

$$H = M \begin{bmatrix} \Lambda_1 & 0 \\ 0 & \Lambda_2 \end{bmatrix} M^{-1} \tag{1.15}$$

and Λ_1, Λ_2 are real Jordan matrices such that the real parts of all eigenvalues are respectively negative and positive. It follows that:

$$\begin{bmatrix} X(t) \\ Y(t) \end{bmatrix} = M \begin{bmatrix} e^{\Lambda_1 t} & 0 \\ 0 & e^{\Lambda_2 t} \end{bmatrix} M^{-1} \begin{bmatrix} I \\ P_0 \end{bmatrix} = \begin{bmatrix} M_{11}e^{\Lambda_1 t} & M_{12}e^{\Lambda_2 t} \\ M_{21}e^{\Lambda_1 t} & M_{22}e^{\Lambda_2 t} \end{bmatrix} \begin{bmatrix} L \\ RL \end{bmatrix}$$

where *L* is an unimportant matrix and

$$R = -\left[M_{22} - P_0 M_{12}\right]^{-1}\left[M_{21} - P_0 M_{11}\right]. \tag{1.16}$$

Moreover,

$$P(t,P_0)=\left[M_{21}+M_{22}e^{\Lambda_2 t}\,\mathrm{R}e^{-\Lambda_1 t}\right]\left[M_{11}+M_{12}e^{\Lambda_2 t}\,\mathrm{R}e^{-\Lambda_1 t}\right]^{-1}. \tag{1.17}$$

Evaluating the Asymptotic Solution

It follows from (1.17) and because $e^{\Lambda_2 t}$ and $e^{-\Lambda_1 t}$ decay to 0 as $t\to\infty$, that

$$\lim_{t\to\infty}P(t,P_0)=\overline{P}=M_{21}M_{11}^{-1}. \tag{1.18}$$

Notice that the limit is approached at an exponential rate equal to twice the smallest real part of any eigenvalue of Λ_2 and is independent of the boundary condition of P_0.

2. Proofs for Chapter 2

2.1 Convergence Proof for Adaptive Stationary Mean

The objective function $J(\mathbf{w}_k;\mathbf{x}_k)$ whose minimizer $\mathbf{w}^*$ is the asymptotic mean $\mathbf{m}=\lim_{k\to\infty}E[\mathbf{x}_k]$ is:

$$J(\mathbf{w}_k;\mathbf{x}_k)=\left\|\mathbf{x}_k-\mathbf{w}_k\right\|^2. \tag{2.1}$$

The gradient of $J(\mathbf{w}_k;\mathbf{x}_k)$ with respect to $\mathbf{w}_k$ is:

$$(1/2)\nabla_{\mathbf{w}_k}J(\mathbf{w}_k;\mathbf{x}_k)=-(\mathbf{x}_k-\mathbf{w}_k). \tag{2.2}$$

From the gradient in (2.2), we obtain the adaptive gradient descent algorithm

$$\mathbf{m}_k=\frac{1}{k}\sum_{i=1}^{k}\mathbf{x}_i=\mathbf{m}_{k-1}+\frac{1}{k}(\mathbf{x}_k-\mathbf{m}_{k-1}). \tag{2.3}$$

where $\eta_k=1/k$. The ODE corresponding to (2.3) is:

$$d\mathbf{w}(t)/dt=\mathbf{m}-\mathbf{w}\text{, where }\mathbf{w}_0=\mathbf{w}(0). \tag{2.4}$$

The solution of (2.4) is:

$$\mathbf{w}(t)=\mathbf{m}+(\mathbf{w}_0-\mathbf{m})e^{-t}\to\mathbf{m}\text{, as }t\to\infty. \tag{2.5}$$

The domain of attraction $D(\mathbf{w}^*)=\{\Re^n\}$. The time constant τ for convergence is 1.

2.2 Convergence Proof for Adaptive Stationary Correlation

The objective function $J(W;A)$ whose minimizer W^* is the asymptotic correlation matrix $A=\lim_{k\to\infty}E[\mathbf{x}_k\mathbf{x}_k^T]$ is:

$$J(W;A) = tr\left((A-W)^T(A-W)\right) \tag{2.6}$$

The gradient of *J*(*W*;*A*) with respect to *W* is:

$$(1/2)\nabla_W J(W;A) = -(A-W). \tag{2.7}$$

From the gradient in (2.7), we obtain the following adaptive gradient descent algorithm:

$$A_k = \frac{1}{k}\sum_{i=1}^{k}\mathbf{x}_i\mathbf{x}_i^T = A_{k-1} + \frac{1}{k}\left(\mathbf{x}_k\mathbf{x}_k^T - A_{k-1}\right). \tag{2.8}$$

for η_k=1/*k*. The ODE corresponding to (2.8) is:

$$dW/dt = A - W, \tag{2.9}$$

which is a Riccati differential equation (2.9) with *F*= –*I*/2, *Q*=*A*, *G*=0, and Hamiltonian matrix *H*:

$$H = \begin{bmatrix} -I/2 & 0 \\ A & I/2 \end{bmatrix} = \begin{bmatrix} 0 & I \\ I & A \end{bmatrix}\begin{bmatrix} -I/2 & 0 \\ 0 & I/2 \end{bmatrix}\begin{bmatrix} 0 & I \\ I & A \end{bmatrix}^{-1}. \tag{2.10}$$

We have $R = (W_0 - A)^{-1}$, where W_0 is the initial value of *W* at *t*=0 From (2.33), we obtain:

$$W(t,W_0) = A + (W_0 - A)e^{-t} \to A \text{ as } t \to \infty. \tag{2.11}$$

The domain of attraction $D(W^*)=\{\Re^{n\mathrm{X}n},\ A=A^T\}$. The time constant is 1, and the rate of convergence is independent of the eigen-structure of *A*.

2.3 Adaptive Normalized Mean Algorithm

The most obvious choice for adaptive normalized mean algorithm is to use (2.3) and normalize each $\mathbf{m}_k$. However, a more efficient algorithm can be obtained from the following cost function whose minimizer $\mathbf{w}^*$ is the asymptotic normalized mean **m**/‖**m**‖, where $\mathbf{m} = \lim_{k\to\infty} E[\mathbf{x}_k]$:

$$J(\mathbf{w}_k;\mathbf{x}_k) = \left\|\mathbf{x}_k - \mathbf{w}_k\right\|^2 + \alpha\left(\mathbf{w}_k^T\mathbf{w}_k - 1\right), \tag{2.12}$$

where α is a Lagrange multiplier that enforces the constraint that the mean is normalized. The gradient of $J(\mathbf{w}_k;\mathbf{x}_k)$ with respect to $\mathbf{w}_k$ is:

$$(1/2)\nabla_{\mathbf{w}_k} J(\mathbf{w}_k;\mathbf{x}_k) = -(\mathbf{x}_k - \mathbf{w}_k) + \alpha\mathbf{w}_k. \tag{2.13}$$

Multiplying (2.13) by $\mathbf{w}_k^T$, and applying the constraint $\mathbf{w}_k^T\mathbf{w}_k = 1$, we obtain:

$$\alpha = \mathbf{w}_k^T\mathbf{x}_k - 1. \tag{2.14}$$

Using this α in (2.14), we obtain the adaptive gradient descent algorithm for normalized mean:

$$\mathbf{w}_{k+1} = \mathbf{w}_k + \eta_k\left(\mathbf{x}_k - \mathbf{w}_k^T\mathbf{x}_k\mathbf{w}_k\right), \tag{2.15}$$

where η_k follows assumption A1.2. The ODE corresponding to (2.15) is:

$$d\mathbf{w}/dt = \mathbf{m} - \mathbf{w}^T\mathbf{m}\mathbf{w}. \tag{2.16}$$

Theorem 2.3. For the ODE (2.16), the point $\mathbf{w}^*=\mathbf{m}/\|\mathbf{m}\|$ is (uniformly) asymptotically stable. The domain of attraction of $\mathbf{w}^*$ is $D(\mathbf{w}^*)=\{\Re^n\}$.

Proof. Defne $v=\mathbf{m}^T\mathbf{w}$. From (2.16), we get $dv/dt = \|\mathbf{m}\|^2-v^2$, whose solution is:

$$v(t)=\|\mathbf{m}\|\left[\frac{\left(\|\mathbf{m}\|+v(0)\right)-\left(\|\mathbf{m}\|-v(0)\right)e^{-2\|\mathbf{m}\|t}}{\left(\|\mathbf{m}\|+v(0)\right)+\left(\|\mathbf{m}\|-v(0)\right)e^{-2\|\mathbf{m}\|t}}\right]\rightarrow\|\mathbf{m}\| \text{ as } t\rightarrow\infty. \tag{2.17}$$

The domain of attraction $D(\mathbf{w}^*)$ is:

$$D(\mathbf{w}^*)=\left\{\left(\|\mathbf{m}\|+v(0)\right)e^{2\|\mathbf{m}\|t}+\left(\|\mathbf{m}\|-v(0)\right)\neq 0 \quad \forall t\geq 0\right\}.$$

Since $(v(0)-\|\mathbf{m}\|)<(v(0)+\|\mathbf{m}\|)$, the above inequality is valid for all $t \geq 0$. Thus, the domain of attraction $D(\mathbf{w}^*)=\{\Re^n\}$. The time constant is $1/(2\|\mathbf{m}\|)$. Clearly, $v=\mathbf{m}^T\mathbf{w}\rightarrow\|\mathbf{m}\|$ as $t\rightarrow\infty$.

We next define $u=\mathbf{w}^T\mathbf{w}$. Then, from (2.16) and $\mathbf{m}^T\mathbf{w}\rightarrow\|\mathbf{m}\|$, we obtain $du/dt = 2\|\mathbf{m}\|(1-u)$. The solution of this ODE is:

$$u(t)=1+(u(0)-1)e^{-2\|\mathbf{m}\|t}\rightarrow 1 \text{ as } t\rightarrow\infty$$

The domain of attraction $D(\mathbf{w}^*)=\{\Re^n\}$. The time constant is $1/(2\|\mathbf{m}\|)$. Clearly, $u=\mathbf{w}^T\mathbf{w}\rightarrow 1$ as $t\rightarrow\infty$. Considering that $\mathbf{m}^T\mathbf{w}\rightarrow\|\mathbf{m}\|$ and $\mathbf{w}^T\mathbf{w}\rightarrow 1$ as $t\rightarrow\infty$, we conclude $\mathbf{w}(t)\rightarrow\mathbf{m}/\|\mathbf{m}\|$ as $t\rightarrow\infty$. ■

In order to prove the convergence of the adaptive algorithm (2.15), we also need to prove that $\mathbf{w}_k$ is bounded. For this, we determine an upper bound of η_k such that $\mathbf{w}_k$ is bounded for all k. The following Theorem gives the result without proof.

Theorem 2.4. For the adaptive algorithm (2.39) let A2.1 and A2.2 hold. Then there exists a uniform upper bound of η_k such that $\mathbf{w}_k$ is almost surely uniformly bounded. Furthermore, if $\|\mathbf{x}_k\| \leq \alpha$ (Assumption A2.1) and θ is the almost sure upper bound of $\|\mathbf{w}_k\|$, then $\|\mathbf{w}_{k+1}\| \leq \|\mathbf{w}_k\|$ if:

$$\eta_k \leq \frac{2}{\alpha\theta}. \qquad ■$$

The convergence of (2.15) is now a direct corollary of the above theorems and an application of Lemma 2.1

2.4 Convergence Proof of Adaptive Median

Given a sequence $\{\mathbf{x}_k\}$, its asymptotic median μ satisfies the following:

$$\lim_{k\rightarrow\infty}\mathrm{P}(\mathbf{x}_k \geq \mu)=\lim_{k\rightarrow\infty}\mathrm{P}(\mathbf{x}_k < \mu)=0.5, \tag{2.18}$$

where P(E) is the probability measure of event E, and 0 ≤ P(E) ≤ 1.
Adaptive median algorithm:

$$\mathbf{w}_{k+1} = \mathbf{w}_k + \eta_k \operatorname{sgn}(\mathbf{x}_k - \mathbf{w}_k), \tag{2.19}$$

The ODE corresponding to (2.19) is:

$$\frac{d\mathbf{w}(t)}{dt} = \lim_{k\to\infty} E[\operatorname{sgn}(\mathbf{x}_k - \mathbf{w})]. \tag{2.20}$$

Clearly, a stationary point $\mathbf{w}^*$ of (2.20) is:

$$\lim_{k\to\infty} \mathrm{P}(\mathbf{x}_k \geq \mathbf{w}^*) = \lim_{k\to\infty} \mathrm{P}(\mathbf{x}_k < \mathbf{w}^*), \tag{2.21}$$

which satisfies the condition (2.18) of asymptotic median μ. Besides, we can consider the objective function $J(\mathbf{w}_k;\mathbf{x}_k)$

$$J(\mathbf{w}_k;\mathbf{x}_k) = \|\mathbf{x}_k - \mathbf{w}_k\| \tag{2.22}$$

as an energy function, and it has been proven [Bickel&Doksum 77] that this energy function is minimized for $\mathbf{w}^*=\mu$.

2.5 Brief Review of Optimization Theory

Most of the methods described in this section are from [Luenberger 84].

Unconstrained Minimization Problem

We first consider the unconstrained optimization problem:

$$\text{Minimize } f(\mathbf{x}) \text{ subject to } \mathbf{x}\in\Re^n. \tag{2.23}$$

Definition 2.1. If a real-valued function f is continuous on $\Re$, and has continuous partial derivatives of order p, we write $f\in C^p$.

Definition 2.2. A point $\mathbf{x}^*\in\Re^n$ is said to be a *local minimum point* of f over $\Re^n$ if there is an ε>0 such that $f(\mathbf{x}) \geq f(\mathbf{x}^*)$ for all $\mathbf{x}\in\Re^n$ within a distance ε of $\mathbf{x}^*$. If $f(\mathbf{x}) > f(\mathbf{x}^*)$ for all $\mathbf{x}\in\Re^n$, $\mathbf{x}\neq\mathbf{x}^*$, within a distance ε of $\mathbf{x}^*$, then $\mathbf{x}^*$ is said to be the *strict local minimum point* of f over $\Re^n$.

Theorem 2.5. (*First Order Necessary Conditions*). Let $f\in C^1$ be a function of $\mathbf{x}$ on $\Re^n$. If $\mathbf{x}^*\in\Re^n$ is a local minimum point of f over $\Re$, then $\nabla_x f(\mathbf{x}^*)=0$.

Theorem 2.6. (*Second Order Sufficient Conditions*). Let $f \in C^2$ be a function of $\mathbf{x}$ on $\Re^n$, and let $\mathbf{x}^* \in \Re^n$. Suppose in addition that (1) $\nabla_x f(\mathbf{x}^*)=0$, and (2) $\nabla^2_{\mathbf{xx}} f(\mathbf{x}^*)$ is positive definite, then $\mathbf{x}^*$ is the strict local minimum point of f.

We next describe the gradient-based methods for iteratively solving the unconstrained minimization problem (2.23).

1. Gradient Descent Method: $\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha \nabla_{\mathbf{x}} f(\mathbf{x}_k)$,

   where $\alpha > 0$ is a small scalar constant.
2. Steepest Descent Method:

   $$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \nabla_{\mathbf{x}} f(\mathbf{x}_k),$$

   where the scalar $\alpha_k \geq 0$ is obtained by minimizing $f(\mathbf{x}_k - \alpha \nabla_x f(\mathbf{x}_k))$ for α.
3. Conjugate Direction Method:

   $$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{d}_k,$$

   where $\alpha_k \geq 0$ is obtained by minimizing $f(\mathbf{x}_k + \alpha \mathbf{d}_k)$ for α, and

   $$\mathbf{d}_{k+1} = -\nabla_{\mathbf{x}} f(\mathbf{x}_{k+1}) + \beta_k \mathbf{d}_k,$$

   where β_k is obtained by one of several methods, where $\mathbf{g}_k = \nabla_{\mathbf{x}} f(\mathbf{x}_k)$:

   Hestenes-Stiefel: $\beta_k = \mathbf{g}_{k+1}^T(\mathbf{g}_{k+1} - \mathbf{g}_k) / \mathbf{d}_k^T(\mathbf{g}_{k+1} - \mathbf{g}_k)$,

   Polak-Ribiere: $\beta_k = \mathbf{g}_{k+1}^T(\mathbf{g}_{k+1} - \mathbf{g}_k) / \mathbf{g}_k^T \mathbf{g}_k$,

   Fletcher-Reeves: $\beta_k = \mathbf{g}_{k+1}^T \mathbf{g}_{k+1} / \mathbf{g}_k^T \mathbf{g}_k$,

   Powell: $\beta_k^i = \max\left[0, \mathbf{g}_{k+1}^T(\mathbf{g}_{k+1} - \mathbf{g}_k) / \mathbf{g}_k^T \mathbf{g}_k\right]$.

   This method accelerates the typically slow convergence of the gradient or steepest descent methods.
4. Newton's Method: $\mathbf{x}_{k+1} = \mathbf{x}_k - \left[\nabla^2_{\mathbf{xx}} f(\mathbf{x}_k)\right]^{-1} \nabla_{\mathbf{x}} f(\mathbf{x}_k)$.

   Here we assume that the Hessian $H_k = \nabla^2_{\mathbf{xx}} f(\mathbf{x}_k)$ is nonsingular.
5. Quasi-Newton Method:

   Although there are a variety of quasi-Newton methods [Luenberger 84], we consider a simple case where the Hessian H_k is recursively updated by a rule like

   $$A_k^{-1} = \frac{k}{\beta(k-1)}\left(A_{k-1}^{-1} - \frac{A_{k-1}^{-1}\mathbf{x}_k\mathbf{x}_k^T A_{k-1}^{-1}}{\beta(k-1) + \mathbf{x}_k^T A_{k-1}^{-1}\mathbf{x}_k}\right)$$

   for a new sample $\mathbf{x}_k$. We use the Sherman-Morrison formula in Section 2.11 to create an update rule for H_k^{-1} similar to those in Section 2.7. We then replace H_k^{-1} in Newton's method by this update rule.

Constrained Minimization Problem

We consider the constrained optimization problem:

$$\text{Minimize } f(\mathbf{x})\in C^2 \text{ subject to } \mathbf{h}(\mathbf{x})=\mathbf{0}, \qquad (2.24)$$

where and $\mathbf{h}=(h_1,h_2,\ldots,h_m)\in C^2$, and $\mathbf{x}\in\Re^n$.

1. Lagrange Multiplier Method:

   We introduce the Lagrangian associated with the constrained problem as:

   $$l(\mathbf{x},\boldsymbol{\lambda}) = f(\mathbf{x}) + \boldsymbol{\lambda}^T\mathbf{h}(\mathbf{x}), \text{ where } \boldsymbol{\lambda}\in\Re^m, \qquad (2.25)$$

   which reduces the constrained problem to an unconstrained one. By the first order necessary conditions (Theorem 2.5), if $\mathbf{x}^*\in\Re^n$ is a local minimum point of f over $\Re$ subject to $\mathbf{h}(\mathbf{x})=\mathbf{0}$, then

   $$\nabla_{\mathbf{x}} l(\mathbf{x}^*,\boldsymbol{\lambda}) = \nabla_{\mathbf{x}} f(\mathbf{x}^*) + \boldsymbol{\lambda}^T\nabla_{\mathbf{x}}\mathbf{h}(\mathbf{x}^*) = \mathbf{0} \text{ and } \nabla_{\boldsymbol{\lambda}} l(\mathbf{x}^*,\boldsymbol{\lambda}) = \mathbf{h}(\mathbf{x}^*) = \mathbf{0}\,. \qquad (2.26)$$

   By the second order sufficient conditions (Theorem 2.6), if we denote by M the tangent plane $M=\{\mathbf{y}:\mathbf{y}^T\nabla_{\mathbf{x}}\mathbf{h}(\mathbf{x}^*)=\mathbf{0}\}$, then if the matrix:

   $$\nabla^2_{\mathbf{xx}} l(\mathbf{x}^*,\boldsymbol{\lambda}) = \nabla^2_{\mathbf{xx}} f(\mathbf{x}^*) + \boldsymbol{\lambda}^T\nabla^2_{\mathbf{xx}}\mathbf{h}(\mathbf{x}^*), \qquad (2.27)$$

   is positive definite on M, then $\mathbf{x}^*$ is the strict local minimum point of f subject to $\mathbf{h}(\mathbf{x})=\mathbf{0}$.

2. Penalty Method:

   Here the unconstrained problem is:

   $$\text{Minimize } f(\mathbf{x}) + \mu\|\mathbf{h}(\mathbf{x})\|^2, \text{ for some large } \mu>0. \qquad (2.28)$$

   Here $\|\mathbf{h}(\mathbf{x})\|^2$ is a penalty function used in literature.

3. Augmented Lagrangian Method:

   This method can be viewed as a combination of the lagrangian and penalty function methods. The unconstrained problem is:

   $$\text{Minimize } f(\mathbf{x}) + \boldsymbol{\lambda}^T\mathbf{h}(\mathbf{x}) + \mu\|\mathbf{h}(\mathbf{x})\|^2, \text{ for some large } \mu>0. \qquad (2.29)$$

2.6 Matrix Operations

We use several matrix operations in deriving and analyzing our adaptive algorithms. Given a real, symmetric matrix $A\in\Re^{n\text{X}n}$, where $A=[a_{ij}]$, let the eigen-decomposition of A be $A=\Phi\Lambda\Phi^T$, where Λ = $diag(\lambda_1,\ldots,\lambda_n)$ is the diagonal eigenvalue matrix with $\lambda_1\geq\ldots\geq\lambda_n$, and $\Phi\in\Re^{n\text{X}n}$ is the eigenvector matrix. Some matrix operations are:

1. *Transpose*: Transpose(A)=A^T,

2. *Trace*: $tr[A] = tr[A^T] = \sum_{i=1}^{n} a_{ii} = \sum_{i=1}^{n} \lambda_i$,
3. *Determinant*: $\det(A) = \prod_{i=1}^{n} \lambda_i$,
4. *Frobenius Norm*: $\|A\|_F^2 = tr[AA^T] = tr[A^T A] = \lambda_1^2 + ... + \lambda_n^2$,
5. *Euclidean Norm*: $\|A\| = \lambda_1$,
6. *Sherman-Morrison Formula*: $\left[A + \mathbf{ab}^T\right]^{-1} = A^{-1} - \dfrac{A^{-1}\mathbf{ab}^T A^{-1}}{1 + \mathbf{b}^T A^{-1}\mathbf{a}}$.

For positive definite square matrix, *A*: $tr(\log(A)) = \log(\det(A))$

2.7 Derivatives of Matrix Functions

In formulating the adaptive algorithms from a scalar objective function, we use gradients of these functions with respect to a matrix. Some useful gradients are listed below. We assume that $C, D \in \Re^{nXn}$ are symmetric and positive definite, $A \in \Re^{nXn}$ is symmetric, and $X \in \Re^{nXn}$.

1. $\nabla_X tr[X] = I$,
2. $\nabla_X tr[e^X] = e^{X^T}$,
3. $\nabla_X tr[CX] = \nabla_X tr[X^T C] = C$,
4. $\nabla_X tr[CX^T] = \nabla_X tr[XC] = C$,
5. $\nabla_X tr[XCX^T] = \nabla_X tr[CX^T X] = \nabla_X tr[X^T XC] = 2XC$,
6. $\nabla_X tr[CXX^T] = \nabla_X tr[X^T CX] = \nabla_X tr[XX^T C] = 2CX$,
7. $\nabla_X tr[XX^T XX^T] = \nabla_X tr[X^T XX^T X] = 4XX^T X$,
8. $\nabla_X tr[CXCX] = \nabla_X tr[X^T CX^T C] = 2C^2 X$,
9. $\nabla_X tr[CXX^T C] = \nabla_X tr[X^T C^2 X] = 2CX^T C$,
10. $\nabla_X tr[XCX^T XCX^T] = 4XCX^T XC$,
11. $\nabla_X tr[X^T CXX^T CX] = 4CXX^T CX$,
12. $\nabla_X tr[XAX^T C] = \nabla_X tr[AX^T CX] = \nabla_X tr[X^T CXA] = 2CXA$,
13. $\nabla_X tr[X^T XX^T CX] = \nabla_X tr[XX^T CXX^T] = \nabla_X tr[X^T CXX^T X] = 2XX^T CX + 2CXX^T X$,
14. $\nabla_X tr[log(X^T CX)] = 2CX(X^T CX)^{-1}$,
15. $\nabla_X tr[log(X^T CXD)] = 2CXD(X^T CX)^{-1} D^{-1}$.

3. Proofs for Chapter 3

3.2 Adaptive Square Root Algorithm – Method 1

Let $\{\mathbf{x}_k \in \Re^n\}$ be a sequence of data vectors, whose online data correlation matrix $A_k \in \Re^{n\mathrm{X}n}$ is given by:

$$A_k = \frac{1}{k}\sum_{i=1}^{k} \beta^{k-i}\mathbf{x}_i\mathbf{x}_i^T \ . \tag{3.1}$$

Here $\mathbf{x}_k$ is an observation vector at time k, and $0<\beta\le 1$ is a forgetting factor used for non-stationary sequences. If the data is stationary, the asymptotic correlation matrix A is:

$$A = \lim_{k\to\infty} E[A_k]. \tag{3.2}$$

Objective Function

Following the methodology described in Section 2.3, we present the algorithm by first showing an objective function J, whose minimum with respect to matrix W gives us the square root of the asymptotic data correlation matrix A. The objective function is:

$$J(W) = \left\| A - W^T W \right\|_F^2 . \tag{3.3}$$

The gradient of $J(W)$ with respect to W is:

$$\nabla_W J(W) = -4W(A - W^T W). \tag{3.4}$$

Adaptive Algorithm

From the gradient in (3.4), we obtain the following adaptive gradient descent algorithm:

$$W_{k+1} = W_k - \eta_k (1/4)\nabla_W J(W_k; A_k) = W_k + \eta_k (W_k A_k - W_k W_k^T W_k) , \tag{3.5}$$

where η_k follows assumption A2.2 in Chapter 2. There are several models for generating the random sequence $\{A_k\}$ from observations $\{\mathbf{x}_k\}$. We can represent A_k simply as its instantaneous value $\mathbf{x}_k\mathbf{x}_k^T$ or by its recursive formulae in (2.21, 2.23).

Solution of the Ordinary Differential Equation (ODE)

We use Stochastic Approximation Theory described in Chapter 2.4 to prove the convergence of (3.5). From (3.3) and (3.5), we obtain the ODE below:

$$\frac{dW}{dt} = -\frac{1}{4}\nabla_W J(W;A) = WA - WW^T W \,. \tag{3.6}$$

Let us define $P=W^TW$. Then from (3.6), we obtain:

$$\frac{dP}{dt} = AP + PA - 2P^2 \,. \tag{3.7}$$

which is a Riccati differential equation (2.12) with $F=A$, $G=2I$, $Q=0$. Let $A=\Phi\Lambda\Phi^T$ be the eigen-decomposition of A. Here Λ = diag($\lambda_1,\ldots,\lambda_n$) is the diagonal eigenvalue matrix with $\lambda_1 \geq \ldots \geq \lambda_n > 0$, and $\Phi \in \Re^{nXn}$ is the eigenvector matrix. The Hamiltonian matrix H is:

$$H = \begin{bmatrix} -A & 2I \\ 0 & A \end{bmatrix} = \begin{bmatrix} \Phi\Lambda^{-1} & \Phi \\ \Phi & 0 \end{bmatrix} \begin{bmatrix} \Lambda & 0 \\ 0 & -\Lambda \end{bmatrix} \begin{bmatrix} 0 & \Phi^T \\ \Phi^T & -\Lambda^{-1}\Phi^T \end{bmatrix}. \tag{3.8}$$

We have $R = \Phi^T(P_0^{-1} - A^{-1})\Phi$, where $P_0 = W_0^T W_0$ is the initial condition of the ODE at t = 0. Clearly, R exists if P_0 is nonsingular, and A is nonsingular by assumption A2.1. Using the standard solution [Anderson&Moore 90] for the RDE (3.7), we obtain a trajectory $P(t,P_0)$ from (2.18) as:

$$P(t,P_0) = \left(A^{-1} + e^{-At}\left(P_0^{-1} - A^{-1}\right)e^{-At}\right)^{-1} = W(t,W_0)^T W(t,W_0) \,. \tag{3.9}$$

It follows from (3.9) that:

$$W(t,W_0) = U(t)\left(A^{-1} + e^{-At}\left((W_0^T W_0)^{-1} - A^{-1}\right)e^{-At}\right)^{-½} \,. \tag{3.10}$$

As $t\to\infty$, we have $W(t) \to UA^{½}$ where $U(t)$ is an orthonormal matrix, and $A^{½} = \Phi\Lambda^{½}\Phi^T$, i.e., the *symmetric positive definite* square root of A. The domain of attraction $D(W^*)$ is:

$$D(W^*) = \left\{\det\left((W_0^T W_0)^{-1} + e^{At}A^{-1}e^{At} - A^{-1}\right) \neq 0 \quad \forall t \geq 0\right\}. \tag{3.11}$$

The domain attraction can be further simplified to:

$$\det\left((W_0^T W_0)^{-1} + \left(e^{2At} - I\right)A^{-1}\right) \neq 0 \quad \forall t \geq 0 \,. \tag{3.12}$$

Since $\det((e^{2At} - I)A^{-1}) \geq 0 \quad \forall t \geq 0$, a sufficient condition for (3.12) is:

$$\det(W_0^T W_0) > 0 \text{ or } \det(W_0) \neq 0 \quad \forall t \geq 0 \,. \tag{3.13}$$

The time constant for (3.10) is A^{-1}. The rate of convergence is dependent on the eigen-structure of A, i.e., larger eigenvalues (Λ) of A lead to faster convergence.

Boundedness of W_k

For the complete convergence proof of (3.5) according to Theorem 2.1 (See Section 2.4.2), we need to find conditions under which W_k is bounded. For this, we determine an upper bound of η_k such that W_k is bounded for all k. From (3.5), we obtain a sufficient condition for $\| W_{k+1} \|_F < \| W_k \|_F$ as:

$$\eta_k < \frac{2tr(W_k^T W_k (W_k^T W_k - A_k))}{tr(W_k^T W_k (W_k^T W_k - A_k)^2)}. \tag{3.14}$$

Thus, if $\| W_k \|_F^2 = tr(W_k^T W_k) \le \alpha$ and $\| A_k \| < \beta$ (see Assumption A2.1), then

$$\eta_k < \frac{2}{(\alpha + \beta)} \tag{3.15}$$

Convergence of adaptive algorithm (3.5) is a direct application of Theorem 1.

3.3 Adaptive Square Root Algorithm – Method 2

Objective Function

The objective function *J*(*W*), whose minimum with respect to *W* gives us the square root of *A* is:

$$J(W) = \left\| A - WW^T \right\|_F^2 . \tag{3.16}$$

The gradient of *J*(*W*) with respect to *W* is:

$$\nabla_W J(W) = -4(A - WW^T)W. \tag{3.17}$$

Adaptive Algorithm

We obtain the following adaptive gradient descent algorithm for square root of *A*:

$$W_{k+1} = W_k - \eta_k (1/4) \nabla_W J(W_k; A_k) = W_k + \eta_k (A_k W_k - W_k W_k^T W_k), \tag{3.18}$$

where η_k follows the assumption A2.2 in Chapter 2.

Convergence Analysis

From (3.18), we obtain the ODE below:

$$\frac{dW}{dt} = -\frac{1}{4} \nabla_W J(W; A) = AW - WW^T W . \tag{3.19}$$

Let us define *P*=*WW*^{*T*}. Then from (3.19), we obtain:

$$\frac{dP}{dt} = AP + PA - 2P^2 . \tag{3.20}$$

which is the same Riccati differential equation as (3.7). Hence the solution for *P*(*t*) in (3.9) applies. The solution for *W*(*t*) is:

$$W(t, W_0) = \left(A^{-1} + e^{-At}\left((W_0^T W_0)^{-1} - A^{-1}\right)e^{-At}\right)^{-1/2} U(t) . \tag{3.21}$$

As $t \rightarrow \infty$, we have $W(t) \rightarrow A^{½}U$ where $U(t)$ is an orthonormal matrix, and $A^{½} = \Phi\Lambda^{½}\Phi^T$, i.e., the *symmetric positive definite* square root of *A*. The domain of attraction $D(W^*)$ is the same as (3.11-3.13). The time constant is A^{-1}, and the rate of convergence is dependent on the eigen-structure of *A*.

3.4 Adaptive Square Root Algorithm – Method 3

Adaptive Algorithm

Following the adaptive algorithms (3.5) and (3.18), we now present an algorithm for the computation of a symmetric positive definite square root of *A*:

$$W_{k+1} = W_k + \eta_k (A_k - W_k^2)\,, \tag{3.22}$$

where η_k follows the assumption A2.2 in Chapter 2 and W_k is symmetric.

Convergence Analysis

From (3.22), we obtain the ODE below:

$$\frac{dW}{dt} = A - W^2\,, \tag{3.23}$$

which is a Riccati differential equation (2.12) with *F*=0, *G*=*I*, *Q*=*A*. The Hamiltonian matrix *H* is:

$$H = \begin{bmatrix} 0 & I \\ A & 0 \end{bmatrix} = \begin{bmatrix} ½\Phi\Lambda^{-½} & ½\Phi\Lambda^{-½} \\ ½\Phi & -½\Phi \end{bmatrix} \begin{bmatrix} \Lambda^{½} & 0 \\ 0 & -\Lambda^{½} \end{bmatrix} \begin{bmatrix} ½\Phi\Lambda^{-½} & ½\Phi\Lambda^{-½} \\ ½\Phi & -½\Phi \end{bmatrix}^{-1}. \tag{3.24}$$

We have $R = \Phi^T (I + W_0 A^{-½})^{-1} (I - W_0 A^{-½})\Phi$, where W_0 is the initial condition of the ODE at *t*=0. Clearly *R* exists if $(I+W_0A^{-½})$ is positive definite, i.e., $W_0 \neq -A^{½}$. From (2.18), we obtain:

$$W(t, W_0) = \left(I - e^{-A^{½}t}\Phi R \Phi^T e^{-A^{½}t}\right)\left(I + e^{-A^{½}t}\Phi R \Phi^T e^{-A^{½}t}\right)^{-1} A^{½}\,. \tag{3.25}$$

In other words,

$$W(t, W_0) = \left(I - e^{-A^{½}t}(I + W_0A^{-½})^{-1}(I - W_0A^{-½})e^{-A^{½}t}\right). \\ \left(I + e^{-A^{½}t}(I + W_0A^{-½})^{-1}(I - W_0A^{-½})e^{-A^{½}t}\right)^{-1} A^{½} \tag{3.26}$$

As $t \rightarrow \infty$, we have $W(t) \rightarrow A^{½} = \Phi\Lambda^{½}\Phi^T$, i.e., the symmetric positive definite square root of *A*. The domain of attraction $D(W^*)$ is:

$$D(W^*) = \left\{W = W^T, W_0 \neq -A^{½}, \det\left(e^{At} + (I + W_0A^{-½})^{-1}(I - W_0A^{-½})\right) \neq 0 \quad \forall t \geq 0\right\}. \tag{3.27}$$

The domain of attraction can be simplified as:

$$\det\left((e^{At}+I)+W_0A^{-1/2}(e^{At}-I)\right)\neq 0 \quad \forall t\geq 0\,. \tag{3.28}$$

Since $\det(e^{At}\pm I)\geq 0\;\forall t\geq 0$, a sufficient condition for (3.28) is:

$$\det(W_0)>0 \;\;\forall t\geq 0\,. \tag{3.29}$$

The rate of convergence of (26) is dependent on the eigen-structure of *A*.

3.5 Adaptive Inverse Square Root Algorithm – Method 1

Objective Function

The objective function *J*(*W*), whose minimizer *W** gives us the inverse square root of *A* is:

$$J(W)=\left\|I-W^TAW\right\|_F^2\,. \tag{3.30}$$

The gradient of *J*(*W*) with respect to *W* is:

$$\nabla_W J(W)=-4AW(I-W^TAW). \tag{3.31}$$

Adaptive Algorithm

From the gradient in (3.31), we obtain the following adaptive gradient descent algorithm:

$$W_{k+1}=W_k-\eta_k(1/4)A_k^{-1}\nabla_W J(W_k;A_k)=W_k+\eta_k(W_k-W_kW_k^TA_kW_k)\,. \tag{3.32}$$

Convergence Analysis

From (3.31) and (3.32), we obtain the ODE below:

$$\frac{dW}{dt}=-\frac{1}{4}A^{-1}\nabla_W J(W;A)=W-WW^TAW\;\;. \tag{3.33}$$

Let us define $P=W^TAW$. Then from (3.33), we obtain:

$$\frac{dP}{dt}=2P-2P^2\,. \tag{3.34}$$

which is a Riccati differential equation (2.12) with *F*=*I*, *G*=2*I*, *Q=0*. The Hamiltonian matrix *H* is:

$$H=\begin{bmatrix}-I & 2I\\ 0 & I\end{bmatrix}=\begin{bmatrix}I & I\\ I & 0\end{bmatrix}\begin{bmatrix}I & 0\\ 0 & -I\end{bmatrix}\begin{bmatrix}0 & I\\ I & -I\end{bmatrix}. \tag{3.35}$$

We have $R=(P_0^{-1}-I)$, where $P_0=W_0^TAW_0$ is the initial condition of the ODE at *t*=0. Clearly, *R* exists if P_0 is nonsingular. From (2.18), we obtain:

$$P(t,P_0)=\left(I+e^{-\Lambda t}\left(P_0^{-1}-I\right)e^{-\Lambda t}\right)^{-1}=W(t,W_0)^TAW(t,W_0)=\left(A^{1/2}W(t,W_0)\right)^T\left(A^{1/2}W(t,W_0)\right). \tag{3.36}$$

It follows from (3.36) that:

$$W(t,W_0) = A^{-½}U(t)\left(I + e^{-\Lambda t}\left((W_0^T AW_0)^{-1} - I\right)e^{-\Lambda t}\right)^{-½}. \tag{3.37}$$

As $t \to \infty$, we have $W(t) \to A^{-½}U$ where $U(t)$ is an orthonormal matrix, and $A^{-½} = \Phi\Lambda^{-½}\Phi^T$, i.e., the symmetric positive definite inverse square root of A. The domain of attraction $D(W^*)$ is:

$$D(W^*) = \left\{\det\left(e^{2\Lambda t} - I + (W_0^T AW_0)^{-1}\right) \neq 0 \quad \forall t \geq 0\right\}. \tag{3.38}$$

Since $\det(e^{2\Lambda t} - I) \geq 0 \; \forall t \geq 0$, a sufficient condition for (3.38) is:

$$\det(W_0^T AW_0) > 0 \quad \forall t \geq 0. \tag{3.39}$$

The <u>time constant</u> for (37) is A^{-1}, and the rate of convergence is dependent on the eigen-structure of A.

Boundedness of W_k

For the complete convergence proof of (3.32) according to Theorem 2.1 (See Section 2.4.2), we need to find conditions under which W_k is bounded. For this, we determine an upper bound of η_k such that W_k is bounded for all k. From (3.32), we obtain a sufficient condition for $\| W_{k+1} \|_F < \| W_k \|_F$ as:

$$\eta_k < \frac{2tr(W_k^T W_k (W_k^T A_k W_k - I))}{tr(W_k^T W_k (W_k^T A_k W_k - I)^2)}. \tag{3.40}$$

Thus, if $\| W_k \|_F^2 = tr(W_k^T W_k) \leq \alpha$ and $\| A_k \| < \beta$ (see Assumption A2.1), then

$$\eta_k < \frac{2}{(\alpha\beta + 1)} \tag{3.41}$$

The convergence of adaptive algorithm (32) is a direct application of Theorem 1.

3.6 Adaptive Inverse Square Root Algorithm – Method 2

Objective Function

The objective function $J(W)$, whose minimum with respect to W gives us the inverse square root of A is:

$$J(W) = \left\| I - WAW^T \right\|_F^2. \tag{3.42}$$

The gradient of $J(W)$ with respect to W is:

$$\nabla_W J(W) = -4(I - WAW^T)WA. \tag{3.43}$$

Adaptive Algorithm

We obtain the following adaptive algorithm for inverse square root of *A*:

$$W_{k+1} = W_k - \eta_k (1/4) \nabla_W J(W_k ; A_k) A_k^{-1} = W_k + \eta_k (W_k - W_k A_k W_k^T W_k) \,, \tag{3.44}$$

where η_k follows the assumption A2.2 in Chapter 2.

Convergence Analysis

From (3.44), we obtain the ODE below:

$$\frac{dW}{dt} = W - WAW^T W \,. \tag{3.45}$$

Let us define *P*=*WAW*^{T}. Then from (3.45), we obtain:

$$\frac{dP}{dt} = 2P - 2P^2 \,. \tag{3.46}$$

which is the a Riccati differential equation as (3.34). Hence the solution for *P*(*t*) in (3.36) applies. The solution for *W*(*t*) is:

$$W(t, W_0) = \left(I + e^{-\Lambda t}\left((W_0 A W_0^T)^{-1} - I\right) e^{-\Lambda t}\right)^{-½} U(t) A^{-½} \,. \tag{3.47}$$

As *t*→∞, we have $W(t) \rightarrow UA^{-½}$ where *U*(*t*) is an orthonormal matrix, and $A^{-½} = \Phi \Lambda^{-½} \Phi^T$. The domain of attraction *D*(*W**) is the same as (3.38).

The <u>time constant</u> for (47) is *A*$^{-1}$, and the rate of convergence is dependent on the eigen-structure of *A*.

3.7 Adaptive Inverse Square Root Algorithm – Method 3

Adaptive Algorithm

By extending the adaptive algorithms (3.32) and (3.44), we now present an adaptive algorithm for the computation of a symmetric positive definite inverse square root of *A*:

$$W_{k+1} = W_k + \eta_k (I - W_k A_k W_k) \,, \tag{3.48}$$

where η_k follows the assumption A2.2 in Chapter 2 and W_k is symmetric.

Convergence Analysis

From (3.46), we obtain the ODE below:

$$\frac{dW}{dt} = I - WAW \,, \tag{3.49}$$

which is a Riccati differential equation (2.12) with *F=0*, *G=A*, *Q=I*. The Hamiltonian matrix *H* as:

$$H=\begin{bmatrix}0 & A\\ I & 0\end{bmatrix}=\begin{bmatrix}\Phi\Lambda^{½} & \Phi\Lambda^{½}\\ \Phi & -\Phi\end{bmatrix}\begin{bmatrix}\Lambda^{½} & 0\\ 0 & -\Lambda^{½}\end{bmatrix}\begin{bmatrix}½\Lambda^{-½}\Phi^{T} & ½\Phi^{T}\\ ½\Lambda^{-½}\Phi^{T} & -½\Phi^{T}\end{bmatrix}. \tag{3.50}$$

We have $R=\Phi^{T}(I+W_0A^{½})^{-1}(I-W_0A^{½})\Phi$, where W_0 is the initial condition of the ODE at t=0. Clearly R exists if (I+$W_0A^{½}$) is positive definite, i.e., $W_0 \neq -A^{-½}$. From (2.18), we obtain:

$$W(t,W_0)=\left(I-e^{-A^{½}t}\Phi R\Phi^{T}e^{-A^{½}t}\right)\left(I+e^{-A^{½}t}\Phi R\Phi^{T}e^{-A^{½}t}\right)^{-1}A^{-½}\,. \tag{3.51}$$

In other words,

$$W(t,W_0)=\left(I-e^{-A^{½}t}(I+W_0A^{½})^{-1}(I-W_0A^{½})e^{-A^{½}t}\right).$$
$$\left(I+e^{-A^{½}t}(I+W_0A^{½})^{-1}(I-W_0A^{½})e^{-A^{½}t}\right)^{-1}A^{-½} \tag{3.52}$$

As $t\to\infty$, we have $W(t)\to A^{-½}=\Phi\Lambda^{-½}\Phi^{T}$, i.e., the symmetric positive definite inverse square root of A. The domain of attraction $D(W^{*})$ is:

$$D(W^{*})=\left\{W=W^{T},W_0\neq -A^{-½},\det\left((e^{At}+I)+W_0A^{½}(e^{At}-I)\right)\neq 0 \quad \forall t\geq 0\right\}. \tag{3.53}$$

Once again, since $\det(e^{At}\pm I)\geq 0\ \forall t\geq 0$, a sufficient condition for (3.53) is:

$$\det(W_0)>0 \quad \forall t\geq 0\,. \tag{3.54}$$

The rate of convergence of (3.52) is dependent on the eigen-structure of A.

4. Proofs for Chapter 4

4.2. Algorithms and Objective Functions

Proofs of Convergence and Assumptions

In order to prove the convergence of the adaptive algorithms to the first principal eigenvector ϕ_1 of $A=\lim_{k\to\infty}E[A_k]$ by using stochastic approximation theory, we require the following assumptions:

Assumption (A4.1). Same as A1.1 in Section 1.3.

Assumption (A4.2). The largest eigenvalue of *A* has unit multiplicity.

Assumption (A4.3). Same as A1.2 in Section 1.3.

For each algorithm, we discuss the convergence results. However, due to the repetitive nature of the convergence proofs and since some of these theorems have been proven by other practitioners, we state the results wherever available, and prove the convergence results for some algorithms. We note that the remaining proofs are similar to those proven here.

4.3. OJA Algorithm

This algorithm was given by Oja *et al.* [Oja 85,89,92]. Intuitively, the OJA algorithm is derived from the Rayleigh quotient criterion by representing it as a Lagrange function, which minimizes $-\mathbf{w}_k^T A_k \mathbf{w}_k$ under the constraint $\mathbf{w}_k^T \mathbf{w}_k = 1$.

Objective Function

In terms of the data samples $\mathbf{x}_k$, the objective function for OJA algorithm can be written as:

$$J(\mathbf{w}_k;\mathbf{x}_k) = -\left\|\mathbf{x}_k^T\left(\mathbf{x}_k - \mathbf{w}_k\mathbf{w}_k^T\mathbf{x}_k\right)\right\|^2 \tag{4.3}$$

If we represent the data correlation matrix A_k by its instantaneous value $\mathbf{x}_k\mathbf{x}_k^T$, then (4.3) is equivalent to the following objective function:

$$J(\mathbf{w}_k;A_k) = -\left\|A_k^{½}\left(I - \mathbf{w}_k\mathbf{w}_k^T\right)A_k^{½}\right\|_F^2 . \tag{4.4}$$

We see, from (4.3), that the objective function $J(\mathbf{w}_k;\mathbf{x}_k)$ represents the difference between the sample $\mathbf{x}_k$ and its transformation due to a matrix $\mathbf{w}_k\mathbf{w}_k^T$. In neural networks, this transform is called *auto-association*[1] [Haykin 94]. Figure 4.1 shows a two-layer auto-associative network.

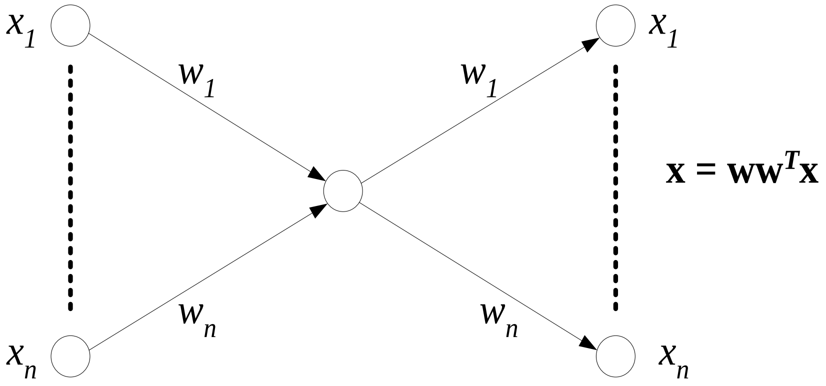

Figure 4.1. Two-Layer Linear Auto-associative Neural Network for First Principal Eigenvector.

Adaptive Algorithm

The gradient of (4.4) with respect to $\mathbf{w}_k$ is:

$$\nabla_{\mathbf{w}_k} J(\mathbf{w}_k; A_k) = -4A_k\left(A_k\mathbf{w}_k - \mathbf{w}_k\mathbf{w}_k^T A_k\mathbf{w}_k\right). \tag{4.5}$$

The adaptive gradient descent OJA algorithm for PCA is:

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \eta_k A_k^{-1}\nabla_{\mathbf{w}_k} J(\mathbf{w}_k; A_k) = \mathbf{w}_k + \eta_k(A_k\mathbf{w}_k - \mathbf{w}_k\mathbf{w}_k^T A_k\mathbf{w}_k)\,, \tag{4.6}$$

where η_k follows assumption A4.3 in Section 4.2.3.

Solution of the ODE

The ordinary differential equation (ODE) associated with the adaptive algorithm (4.6) is:

$$\frac{d}{dt}\mathbf{w}(t) = A\mathbf{w} - \mathbf{w}\mathbf{w}^T A\mathbf{w}\,, \tag{4.7}$$

where $\mathbf{w}(t)$ is the continuous time counterpart of $\mathbf{w}_k$ with t denoting continuous time. In the following theorem, we look for the stable equilibrium points of the ODE (4.7), and for convergence properties of (4.3) for initial states close to the stable points.

[1] In the auto-associative mode, the output of the network is desired to be same as the input.

Theorem 4.1. *Let A4.1 and A4.2 hold, and let* $\mathbf{w}(t)=\sum_{i=1}^{n}a_i(t)\boldsymbol{\phi}_i$ *be a solution to the ODE* (4.7) *in terms of the entire orthonormal set of eigenvectors* $\{\phi_1, \phi_2, \ldots, \phi_n\}$ *of A. Then for any initial condition* $\mathbf{w}(0)=\mathbf{w}_0\in\Re^n$, $\mathbf{w}_0^T\phi_1\neq 0$, *the solutions for the coefficients* $a_i(t)$ *for* $t \geq 0$ *are given by:*

$$\frac{a_i(t)}{a_1(t)}=\frac{a_i(0)}{a_1(0)}e^{-(\lambda_1-\lambda_i)t} \text{ for } i=2,\ldots,n \tag{4.8}$$

and

$$a_1(t)=\pm\sqrt{\frac{c_0}{e^{-2\lambda_1 t}+c_0}}\text{, where } c_0=\frac{a_1(0)^2}{1-a_1(0)^2}. \tag{4.9}$$

The points $\pm\phi_1$ *are* (*uniformly*) *asymptotically stable. The domain of attraction of* ϕ_1 *is* $D(\phi_1)=\{\mathbf{w}\in\Re^n \mid \mathbf{w}^T\phi_1>0\}$ *and that of* $-\phi_1$ *is* $D(-\phi_1)=\{\mathbf{w}\in\Re^n \mid \mathbf{w}^T\phi_1<0\}$.

Proof. Proof similar to Theorem 4.3 and [Oja&Karhunen 85, Lemma 2]. ■

It is clear from (4.8) and (4.9) that $a_1(t)\to\pm1$ and $a_i(t)\to 0$ for $i=2,\ldots,n$ as $t\to\infty$. Thus, $\mathbf{w}(t)\to\pm\phi_1$ as $t\to\infty$.

$\mathbf{w}_k$ *Visits Domain of Attraction Infinitely Often*

In order to complete the proof of convergence, we need to establish that $\mathbf{w}_k$ visits one of the domains of attraction $D(\phi_1)$ or $D(-\phi_1)$ infinitely often. Theorem below summarizes the result.

Theorem 4.2. *For the adaptive algorithm* (4.6) *let A4.1 and A4.3 hold. Assume that* $\mathbf{w}_0$ *is almost surely bounded. Then there exists a uniform upper bound of* η_k *such that* $\mathbf{w}_k$ *is almost surely uniformly bounded. Furthermore, if* $\|\mathbf{w}_0\|\leq\mu+1$ *and* α *is the almost sure upper bound of* $\|A_k\|$, *then* $\|\mathbf{w}_k\|\leq\mu+1$ *if:*

$$\eta_k \leq \frac{2}{\alpha\mu}. \tag{4.10}$$

Proof. See [Oja&Karhunen 85, Lemma 5] and Theorem 4.5. ■

The convergence of (4.6) is now a direct corollary of the above theorems and an application of Theorem 1 of Ljung [Ljung 77; Benveniste *et al.* 90].

Rate of Convergence

The rate of convergence of the OJA algorithm can be obtained from (4.8) and (4.9). The time constants for $a_i(t)$ are $1/(\lambda_1-\lambda_i)$ for $i=2,\ldots,n$, and for $a_1(t)$ is $1/\lambda_1$. The time constants are dependent on the eigen-structure of *A*.

4.4. RQ, OJAN, and LUO Algorithms

Objective Function

All three algorithms are different derivations of the Rayleigh Quotient (RQ) objective function given below:

$$J(\mathbf{w}_k;A_k)=-\left(\frac{\mathbf{w}_k^T A_k \mathbf{w}_k}{\mathbf{w}_k^T \mathbf{w}_k}\right). \tag{4.11}$$

These algorithms were initially presented by Luo *et al.* [Luo *et al.* 97; Taleb *et al.* 99; Cirrincione *et al.* 00], and Oja *et al.* [Oja *et al.* 92]. Variations of the RQ algorithm have been presented by many practitioners [Chauvin 89; Sarkar *et al.* 89; Yang *et al.* 89; Fu&Dowling 95; Taleb *et al.* 99; Cirrincione *et al.* 00].

Adaptive Algorithms

The gradient of (4.11) with respect to $\mathbf{w}_k$ is:

$$\nabla_{\mathbf{w}_k} J(\mathbf{w}_k;A_k)=\frac{-1}{\mathbf{w}_k^T \mathbf{w}_k}\left(A_k\mathbf{w}_k-\mathbf{w}_k\frac{\mathbf{w}_k^T A_k \mathbf{w}_k}{\mathbf{w}_k^T \mathbf{w}_k}\right). \tag{4.12}$$

The adaptive gradient descent RQ algorithm for PCA is:

$$\mathbf{w}_{k+1}=\mathbf{w}_k-\eta_k\nabla_{\mathbf{w}_k} J(\mathbf{w}_k;A_k)=\mathbf{w}_k+\eta_k\frac{1}{\mathbf{w}_k^T \mathbf{w}_k}\left(A_k\mathbf{w}_k-\mathbf{w}_k\frac{\mathbf{w}_k^T A_k \mathbf{w}_k}{\mathbf{w}_k^T \mathbf{w}_k}\right). \tag{4.13}$$

The adaptive gradient descent OJAN algorithm for PCA is:

$$\mathbf{w}_{k+1}=\mathbf{w}_k-\eta_k\left(\mathbf{w}_k^T \mathbf{w}_k\right)\nabla_{\mathbf{w}_k} J(\mathbf{w}_k;A_k)=\mathbf{w}_k+\eta_k\left(A_k\mathbf{w}_k-\mathbf{w}_k\frac{\mathbf{w}_k^T A_k \mathbf{w}_k}{\mathbf{w}_k^T \mathbf{w}_k}\right). \tag{4.14}$$

The adaptive gradient descent LUO algorithm for PCA is:

$$\mathbf{w}_{k+1}=\mathbf{w}_k-\eta_k\left(\mathbf{w}_k^T \mathbf{w}_k\right)^2\nabla_{\mathbf{w}_k} J(\mathbf{w}_k;A_k)=\mathbf{w}_k+\eta_k\left(\mathbf{w}_k^T \mathbf{w}_k\right)\left(A_k\mathbf{w}_k-\mathbf{w}_k\frac{\mathbf{w}_k^T A_k \mathbf{w}_k}{\mathbf{w}_k^T \mathbf{w}_k}\right). \tag{4.15}$$

Convergence Analysis

The ordinary differential equations (ODEs) associated with the adaptive algorithms (4.13)-(4.15) are:

RQ:

$$\frac{d}{dt}\mathbf{w}(t)=\frac{1}{\mathbf{w}^T\mathbf{w}}\left(A\mathbf{w}-\mathbf{w}\frac{\mathbf{w}^T A\mathbf{w}}{\mathbf{w}^T\mathbf{w}}\right). \tag{4.16}$$

OJAN:
$$\frac{d}{dt}\mathbf{w}(t)=\left(A\mathbf{w}-\mathbf{w}\frac{\mathbf{w}^T A\mathbf{w}}{\mathbf{w}^T\mathbf{w}}\right). \tag{4.17}$$

LUO:
$$\frac{d}{dt}\mathbf{w}(t)=\mathbf{w}^T\mathbf{w}\left(A\mathbf{w}-\mathbf{w}\frac{\mathbf{w}^T A\mathbf{w}}{\mathbf{w}^T\mathbf{w}}\right). \tag{4.18}$$

Here **w**(*t*) is the continuous time counterpart of $\mathbf{w}_k$ with *t* denoting continuous time. From (4.16)-(4.18), we first observe that:

$$\frac{d}{dt}\left\|\mathbf{w}(t)\right\|^2=2\mathbf{w}(t)^T\frac{d}{dt}\mathbf{w}(t)=0\,. \tag{4.19}$$

Thus, for all three algorithms, we have:

$$\left\|\mathbf{w}(t)\right\|^2=\left\|\mathbf{w}(0)\right\|^2. \tag{4.20}$$

The RQ and LUO ODE’s can be modified as:

RQ:
$$\frac{d}{dt}\mathbf{w}(t)=\frac{1}{\left\|\mathbf{w}(0)\right\|^2}\left(A\mathbf{w}-\mathbf{w}\frac{\mathbf{w}^T A\mathbf{w}}{\mathbf{w}^T\mathbf{w}}\right). \tag{4.21}$$

LUO:
$$\frac{d}{dt}\mathbf{w}(t)=\left\|\mathbf{w}(0)\right\|^2\left(A\mathbf{w}-\mathbf{w}\frac{\mathbf{w}^T A\mathbf{w}}{\mathbf{w}^T\mathbf{w}}\right). \tag{4.22}$$

The solutions of the three ODEs are summarized in the theorem below.

Theorem 4.3. *Let A4.1 and A4.2 hold, and let* $\mathbf{w}(t)=\sum_{i=1}^{n}a_i(t)\phi_i$ *be solutions for the ODEs* (4.16)-(4.18) *in terms of the entire orthonormal set of eigenvectors* $\{\phi_1,\phi_2,\ldots,\phi_n\}$ *of A. Then for any initial condition* $\mathbf{w}(0)=\mathbf{w}_0\in\Re^n$, $\mathbf{w}_0^T\phi_1\neq 0$, *the solutions for the coefficients* $a_i(t)$ *for* $t\geq 0$ *for the RQ, OJAN, and LUO algorithms are:*

RQ:
$$\frac{a_i(t)}{a_1(t)}=\frac{a_i(0)}{a_1(0)}e^{\frac{-(\lambda_1-\lambda_i)t}{\left\|\mathbf{w}_0\right\|^2}}\quad\text{for } i=2,\ldots,n \tag{4.23}$$

OJAN:
$$\frac{a_i(t)}{a_1(t)}=\frac{a_i(0)}{a_1(0)}e^{-(\lambda_1-\lambda_i)t}\quad\text{for } i=2,\ldots,n \tag{4.24}$$

LUO:
$$\frac{a_i(t)}{a_1(t)}=\frac{a_i(0)}{a_1(0)}e^{-\left\|\mathbf{w}_0\right\|^2(\lambda_1-\lambda_i)t}\quad\text{for } i=2,\ldots,n \tag{4.25}$$

and

RQ:
$$a_1(t)=\pm\sqrt{\frac{c_0\left\|\mathbf{w}_0\right\|^2}{e^{-2\lambda_1\left\|\mathbf{w}_0\right\|^{-2}t}+c_0}}\,, \tag{4.26}$$

OJAN:
$$a_1(t)=\pm\sqrt{\frac{c_0\left\|\mathbf{w}_0\right\|^2}{e^{-2\lambda_1 t}+c_0}}\,, \tag{4.27}$$

LUO: $a_1(t) = \pm\sqrt{\dfrac{c_0\|\mathbf{w}_0\|^2}{e^{-2\lambda_1\|\mathbf{w}_0\|^2 t} + c_0}}$, where $c_0 = \dfrac{a_1(0)^2}{\|\mathbf{w}_0\|^2 - a_1(0)^2}$. (4.28)

The points ±‖$\mathbf{w}_0$‖ϕ_1 *are* (*uniformly*) *asymptotically stable. The domain of attraction of* ‖$\mathbf{w}_0$‖ϕ_1 *is* D(‖$\mathbf{w}_0$‖ϕ_1)={$\mathbf{w}\in\Re^n$ | $\mathbf{w}^T\phi_1$>0} *and that of* –‖$\mathbf{w}_0$‖ϕ_1 *is* D(–‖$\mathbf{w}_0$‖ϕ_1)={$\mathbf{w}\in\Re^n$ | $\mathbf{w}^T\phi_1$<0}.

Proof. We show the result for ODE (4.21) and the remaining ODEs are similar. Let $\mathbf{w}(t) = \sum_{i=1}^{n} a_i(t)\boldsymbol{\phi}_i$ be solutions for the ODE (4.21) in terms of the entire orthonormal set of eigenvectors {ϕ_1, ϕ_2, …, ϕ_n} of *A*. Substituting this **w**(*t*) in (4.21) and multiplying on the left by $\boldsymbol{\phi}_i^T$, we obtain:

$$\frac{da_i}{dt} = \frac{1}{\|\mathbf{w}_0\|^2}\left(a_i\lambda_i - a_i\frac{\sum_{i=1}^{n} a_i^2\lambda_i}{\|\mathbf{w}_0\|^2}\right). \tag{4.A.1}$$

Let $b_i = a_i / a_1$ for *i*=2,…,*n*. Then, we can write the (4.A.1) as:

$$\frac{db_i}{dt} = -b_i\left(\frac{\lambda_1 - \lambda_i}{\|\mathbf{w}_0\|^2}\right), \tag{4.A.2}$$

which gives us the solution in (4.23). Since we know from (4.23) that $a_i(t)\to 0$ for *i*=2,…,*n* as $t\to\infty$, we obtain from (4.A.1):

$$\frac{da_1}{dt} = \frac{1}{\|\mathbf{w}_0\|^2}\left(a_1\lambda_1 - \frac{a_1^3\lambda_1}{\|\mathbf{w}_0\|^2}\right), \tag{4.A.3}$$

which gives us the solution in (4.26). If $a_1(0)\neq 0$ then $a_1(t)\neq 0$ for all *t*. Thus the sign of $a_1(t)$ is determined by the sign of $a_1(0)=\mathbf{w}(0)^T\phi_1$. ■

Theorem 4.5. Let A4.1 hold. Then there exists a uniform upper bound for η_k such that $\mathbf{w}_k$ is uniformly bounded w.p.1. Furthermore, if $\|\mathbf{w}_k\|^2 \le \alpha+1$, then $\|\mathbf{w}_{k+1}\|^2 \le \|\mathbf{w}_k\|^2$ if:

$$\eta_k < \frac{2(\alpha+1)}{\alpha}. \tag{4.44}$$

Proof. Let ρ be the principal eigenvector of A_k and θ the corresponding (largest) eigenvalue. We multiply (4.40) on the left by ρ^T and we define $\mathbf{v}_k=\rho^T\mathbf{w}_k$. Then from (4.40), we get:

$$\mathbf{v}_{k+1} = \mathbf{v}_k + \eta_k\left(\frac{\theta\mathbf{v}_k}{\mathbf{w}_k^T A_k \mathbf{w}_k} - \mathbf{v}_k\right). \tag{4.A.4}$$

Taking the norm of the above equation and using $\|\mathbf{v}_{k+1}\|^2 \le \|\mathbf{v}_k\|^2$, we get from (4.A.4):

$$\eta_k < 2\bigg/\left(1 - \frac{\theta}{\mathbf{w}_k^T A_k \mathbf{w}_k}\right). \tag{4.A.5}$$

Using $\mathbf{w}_k^T A_k \mathbf{w}_k \le \theta \|\mathbf{w}_k\|^2$, we get from (4.A.5):

$$\eta_k \le \frac{2}{1-\left(1/\|\mathbf{w}_k\|^2\right)}, \text{ which implies } \eta_k < \frac{2(\alpha+1)}{\alpha}. \qquad \blacksquare$$

It is clear from (4.23)-(4.28) that $a_1(t)\to\pm\|\mathbf{w}_0\|$ and $a_i(t)\to 0$ for i=2,…,n as $t\to\infty$. Thus, if $\|\mathbf{w}_0\|$=1, then $\mathbf{w}(t)\to\pm\phi_1$ as $t\to\infty$. The remainder of the convergence proof is similar to Section 4.3.4, i.e., there exists a uniform upper bound of η_k such that $\mathbf{w}_k$ is almost surely uniformly bounded. Intuitively, we see that all three adaptive algorithms are similar except for a different η_k. Using (4.20), we can apply the approximation $\|\mathbf{w}_k\|^2\approx\|\mathbf{w}_0\|^2$. Thus, the three adaptive algorithms (4.13)-(4.15) can be written as:

RQ: $$\mathbf{w}_{k+1} = \mathbf{w}_k + \frac{\eta_k}{\|\mathbf{w}_0\|^2}\left(A_k\mathbf{w}_k - \mathbf{w}_k\frac{\mathbf{w}_k^T A_k \mathbf{w}_k}{\mathbf{w}_k^T\mathbf{w}_k}\right). \qquad (4.29)$$

OJAN: $$\mathbf{w}_{k+1} = \mathbf{w}_k + \eta_k\left(A_k\mathbf{w}_k - \mathbf{w}_k\frac{\mathbf{w}_k^T A_k \mathbf{w}_k}{\mathbf{w}_k^T\mathbf{w}_k}\right). \qquad (4.30)$$

LUO: $$\mathbf{w}_{k+1} = \mathbf{w}_k + \eta_k\|\mathbf{w}_0\|^2\left(A_k\mathbf{w}_k - \mathbf{w}_k\frac{\mathbf{w}_k^T A_k \mathbf{w}_k}{\mathbf{w}_k^T\mathbf{w}_k}\right). \qquad (4.31)$$

Thus, all three algorithms are same except for different choices of gain constants, i.e., they are $\eta_k\|\mathbf{w}_0\|^{-2}$, η_k, and $\eta_k\|\mathbf{w}_0\|^2$ for the RQ, OJAN and LUO algorithms respectively.

Rate of Convergence

The rates of convergence for the three algorithms can be obtained from (4.23)-(4.28). The time constants for $a_i(t)$ are:

RQ: $\|\mathbf{w}_0\|^2/(\lambda_1-\lambda_i)$ for i=2,…,n. (4.32)

OJAN: $1/(\lambda_1-\lambda_i)$ for i=2,…,n. (4.33)

LUO: $\|\mathbf{w}_0\|^{-2}/(\lambda_1-\lambda_i)$ for i=2,…,n. (4.34)

The time constants for $a_1(t)$ are:

RQ: $\|\mathbf{w}_0\|^2/\lambda_1$. (4.35)

OJAN: $1/\lambda_1$. (4.36)

LUO: $\|\mathbf{w}_0\|^{-2}/\lambda_1$. (4.37)

The time constants are dependent on the eigen-structure of A.

4.5. IT Algorithm

Objective Function

The objective function for the Information Theory (IT) criterion is:

$$J(\mathbf{w}_k;A_k)=\mathbf{w}_k^T\mathbf{w}_k-\ln\left(\mathbf{w}_k^TA_k\mathbf{w}_k\right). \tag{4.38}$$

Plumbley [Pumbley 95] and Miao and Hua [Miao&Hua 98] have studied this objective function.

Adaptive Algorithm

The gradient of (4.38) with respect to **w***k* is:

$$\nabla_{\mathbf{w}_k}J(\mathbf{w}_k;A_k)=\mathbf{w}_k-\frac{A_k\mathbf{w}_k}{\mathbf{w}_k^TA_k\mathbf{w}_k}. \tag{4.39}$$

The adaptive gradient descent IT algorithm for PCA is:

$$\mathbf{w}_{k+1}=\mathbf{w}_k-\eta_k\nabla_{\mathbf{w}_k}J(\mathbf{w}_k;A_k)=\mathbf{w}_k+\eta_k\left(\frac{A_k\mathbf{w}_k}{\mathbf{w}_k^TA_k\mathbf{w}_k}-\mathbf{w}_k\right). \tag{4.40}$$

Convergence Analysis

The ordinary differential equation (ODE) associated with the adaptive algorithms (4.40) is:

$$\frac{d}{dt}\mathbf{w}(t)=\frac{A\mathbf{w}}{\mathbf{w}^TA\mathbf{w}}-\mathbf{w}\ , \tag{4.41}$$

where **w**(*t*) is the continuous time counterpart of **w***k* with *t* denoting continuous time. From (4.41), we get:

$$\frac{d}{dt}\|\mathbf{w}(t)\|^2=2\mathbf{w}(t)^T\frac{d}{dt}\mathbf{w}(t)=2\left(1-\|\mathbf{w}(t)\|^2\right).$$

We obtain:

$$\|\mathbf{w}(t)\|^2=1+(\|\mathbf{w}(0)\|^2-1)\,e^{-2t}. \tag{4.42}$$

Clearly, as $t\to\infty$, $\|\mathbf{w}(t)\|\to\pm1$.

Next, we analyze the stable stationary points of the ODE (4.41). The objective function *J*(**w**,*A*) in (4.38) is also an energy function *E*(**w**) for the adaptive algorithm (4.40) as:

$$E(\mathbf{w})=\mathbf{w}^T\mathbf{w}-\ln(\mathbf{w}^TA\mathbf{w}). \tag{4.43}$$

It is obvious that $\lim_{\mathbf{w}\to\mathbf{0}}E(\mathbf{w})=+\infty$ and $\lim_{\|\mathbf{w}\|\to\infty}E(\mathbf{w})=+\infty$. Therefore, the function *E*(**w**) has global minimum points.

Theorem 4.4. *Let A4.1 and A4.2 hold. Vectors* $\pm\phi_1$*, the two converged points of* (4.40) *are global minimum points of E*(**w**)*, and E*(**w**) *has no other local minimum point. In addition,* $\pm\phi_i$ (*i=2,…,n*) *are saddle points of E*(**w**)*.*

Proof. See [Zhang&Leung 95]. ■

Rate of Convergence

The rate of convergence for (4.40) can be obtained from (4.42). A unique feature of this algorithm is that the time constant for ‖**w**(*t*)‖ is 1, and is *independent* of the eigen-structure of *A*.

Upper Bound of η_k

There exists a uniform upper bound for η_k such that w_k is uniformly bounded w.p.1. Furthermore, if $\|\mathbf{w}_k\|^2 \le \alpha+1$, then $\|\mathbf{w}_{k+1}\|^2 \le \|\mathbf{w}_k\|^2$ if:

$$\eta_k < \frac{2(\alpha+1)}{\alpha}. \tag{4.44}$$

Proof. See Appendix.

■

4.6. XU Algorithm

Objective Function

Originally presented by Xu [Xu 91, 93], the objective function for XU algorithm is:

$$J(\mathbf{w}_k;A_k) = -\mathbf{w}_k^T A_k \mathbf{w}_k + \mathbf{w}_k^T A_k \mathbf{w}_k (\mathbf{w}_k^T \mathbf{w} - 1) = -2\mathbf{w}_k^T A_k \mathbf{w}_k + \mathbf{w}_k^T A_k \mathbf{w}_k \mathbf{w}_k^T \mathbf{w}_k \,. \tag{4.45}$$

The objective function *J*(**w**$_k$;*A*$_k$) represents the mean squared error between the sample $\mathbf{x}_k$ and its transformation due to a matrix $\mathbf{w}_k\mathbf{w}_k^T$. This transform, also known as *auto-association*, is shown in Figure 4.1. We define $A_k = (1/k)\sum_{t=1}^{k}\mathbf{x}_t\mathbf{x}_t^T$. Then, the mean squared error objective function is:

$$J(\mathbf{w}_k;A_k) = \frac{1}{k}\sum_{i=1}^{k}\left\|\mathbf{x}_k - \mathbf{w}_k\mathbf{w}_k^T\mathbf{x}_k\right\|^2 = trA_k - 2\mathbf{w}_k^T A_k \mathbf{w}_k + \mathbf{w}_k^T A_k \mathbf{w}_k \mathbf{w}_k^T \mathbf{w}_k \,, \tag{4.46}$$

which is the same as (4.45).

Adaptive Algorithm

The gradient of (4.45) with respect to $\mathbf{w}_k$ is:

$$\nabla_{\mathbf{w}_k} J(\mathbf{w}_k;A_k) = -(2A_k\mathbf{w}_k - \mathbf{w}_k\mathbf{w}_k^T A_k\mathbf{w}_k - A_k\mathbf{w}_k\mathbf{w}_k^T\mathbf{w}_k)\,. \tag{4.47}$$

The adaptive gradient descent XU algorithm for PCA is:

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \eta_k\nabla_{\mathbf{w}_k} J(\mathbf{w}_k;A_k) = \mathbf{w}_k + \eta_k\left(2A_k\mathbf{w}_k - \mathbf{w}_k\mathbf{w}_k^T A_k\mathbf{w}_k - A_k\mathbf{w}_k\mathbf{w}_k^T\mathbf{w}_k\right). \tag{4.48}$$

Convergence Analysis

The ordinary differential equation (ODE) associated with the adaptive algorithms (4.48) is:

$$\frac{d}{dt}\mathbf{w}(t) = 2A\mathbf{w} - \mathbf{w}\mathbf{w}^T A\mathbf{w} - A\mathbf{w}\mathbf{w}^T\mathbf{w}\,, \tag{4.49}$$

where **w**(*t*) is the continuous time counterpart of $\mathbf{w}_k$ with *t* denoting continuous time.

We next analyze the stable stationary points of the ODE (4.49). The objective function *J* in (4.45) is also an energy function *E*(**w**) for the adaptive algorithm (4.48) as:

$$E(\mathbf{w}) = -2\mathbf{w}^T A\mathbf{w} + \mathbf{w}^T A\mathbf{w}\mathbf{w}^T\mathbf{w}\,. \tag{4.50}$$

The stable stationary points of this energy function are shown in Theorem 4.6 below.

Theorem 4.6. *Let A4.1 and A4.2 hold. Vectors* $\pm\phi_1$*, the two critical points of* (4.48), *are global minimum points of E*(**w**)*, and E*(**w**) *has no other local minimum point. In addition,* $\pm\phi_i$ (*i*=2,…,*n*) *are saddle points of E*(**w**).

Proof. See [Chatterjee *et al.* Mar 00, Theorems 4.1, 4.2]. ■

Rate of Convergence

The time constants for $a_i(t)$ are $1/(\lambda_1 - \lambda_i)$ for *i*=2,…,*n*, and for $a_1(t)$ is $1/\lambda_1$. The time constants are dependent on the eigen-structure of *A*.

Upper Bound of η_k

Theorem 4.7. Let A4.1 hold. Then there exists a uniform upper bound for η_k such that $\mathbf{w}_k$ is uniformly bounded w.p.1. Furthermore, if $\|\mathbf{w}_k\|^2 \le \alpha+1$, and θ is the largest eigenvalue of A_k, then $\|\mathbf{w}_{k+1}\|^2 \le \|\mathbf{w}_k\|^2$ if:

$$\eta_k < \frac{1}{\theta\alpha}. \tag{4.51}$$

Proof. See [Chatterjee et al. Mar 00, Theorem 4.3]. ■

4.7. Penalty Function (PF) Algorithm

Objective Function

Originally given by Chauvin [Chauvin 89], the objective function for Penalty Function (PF) algorithm is:

$$J(\mathbf{w}_k;A_k) = -\mathbf{w}_k^T A_k \mathbf{w}_k + \frac{\mu}{2}\left(\mathbf{w}_k^T \mathbf{w}_k - 1\right)^2, \quad \mu > 0. \tag{4.52}$$

The objective function *J*($\mathbf{w}_k$;A_k) is an implementation of the Rayleigh Quotient criterion (4.1), where the constraint $\mathbf{w}_k^T\mathbf{w}_k = 1$ is enforced by the penalty function method of nonlinear optimization, and μ is a positive penalty constant.

Adaptive Algorithm

The gradient of (4.52) with respect to $\mathbf{w}_k$ is:

$$(1/2)\nabla_{\mathbf{w}_k} J(\mathbf{w}_k;A_k) = -(A_k\mathbf{w}_k - \mu\mathbf{w}_k(\mathbf{w}_k^T\mathbf{w}_k - 1)).$$

The adaptive gradient descent PF algorithm for PCA is:

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \eta_k \nabla_{\mathbf{w}_k} J(\mathbf{w}_k;A_k) = \mathbf{w}_k + \eta_k\left(A_k\mathbf{w}_k - \mu\mathbf{w}_k(\mathbf{w}_k^T\mathbf{w}_k - 1)\right), \tag{4.53}$$

where $\mu > 0$.

Convergence Analysis

The ordinary differential equation (ODE) associated with the adaptive algorithm (4.58) is:

$$\frac{d}{dt}\mathbf{w}(t) = A\mathbf{w} - \mu\mathbf{w}\left(\mathbf{w}^T\mathbf{w} - 1\right), \tag{4.54}$$

where $\mu > 0$ and **w**(*t*) is the continuous time counterpart of $\mathbf{w}_k$ with *t* denoting continuous time.

Theorem 4.8. *Let A4.1 and A4.2 hold, and let* $\mathbf{w}(t) = \sum_{i=1}^n a_i(t)\boldsymbol{\phi}_i$ *be a solution to the ODE* (4.54) *in terms of the entire orthonormal set of eigenvectors* {$\boldsymbol{\phi}_1$, $\boldsymbol{\phi}_2$, ..., $\boldsymbol{\phi}_n$} *of A. Then for any initial condition* **w**(0)=$\mathbf{w}_0 \in \Re^n$, $\mathbf{w}_0^T\boldsymbol{\phi}_1 \neq 0$, *the solutions for the coefficients* $a_i(t)$ *for* $t \geq 0$ *are given by:*

$$\frac{a_i(t)}{a_1(t)} = \frac{a_i(0)}{a_1(0)}e^{-(\lambda_1-\lambda_i)t} \text{ for } i=2,\ldots,n \tag{4.55}$$

and

$$a_1(t) = \pm c\sqrt{\frac{k_0}{e^{-2(\lambda_1+\mu)t} + k_0}} \text{ where } c = \sqrt{1 + \frac{\lambda_1}{\mu}} \text{ and } k_0 = \frac{a_1(0)^2}{c^2 - a_1(0)^2}. \tag{4.56}$$

The points ±cϕ_1 *are* (*uniformly*) *asymptotically stable. The domain of attraction of* cϕ_1 *is* D(cϕ_1)={**w** $\in\Re^n$ | $\mathbf{w}^T\phi_1$>0} *and that of* –cϕ_1 *is* D(–cϕ_1)={**w**$\in\Re^n$ | $\mathbf{w}^T\phi_1$<0}.

Proof. Similar to Theorem 4.3. ■

It is clear from (4.55) and (4.56) that $a_1(t) \to \pm\sqrt{1+\lambda_1/\mu}$ and $a_i(t)$→0 for *i*=2,…,*n* as *t*→∞. Thus, $\mathbf{w}(t) \to \pm\phi_1\sqrt{1+\lambda_1/\mu}$ as *t*→∞.

We next analyze the stable stationary points of the ODE (4.54). The objective function *J* in (4.52) is also an energy function *E*(**w**) for the adaptive algorithm (4.53) as:

$$E(\mathbf{w}) = -\mathbf{w}^T A\mathbf{w} + \frac{\mu}{2}\left(\mathbf{w}^T\mathbf{w} - 1\right)^2. \tag{4.57}$$

The stable stationary points of this energy function are shown in Theorem 4.9 below.

Theorem 4.9. *Let A4.1 and A4.2 hold. The critical points of E*(**w**) *are* **0** *and* $\pm\phi_i\sqrt{1+\lambda_i/\mu}$ for *i*=1,…,*n. The two critical points* $\pm\phi_1\sqrt{1+\lambda_1/\mu}$ *are global minimum points of E*(**w**), *and E*(**w**) *has no other local minimum point. The point* **0** *is a local maximum. In addition,* $\pm\phi_i\sqrt{1+\lambda_i/\mu}$ (*i=2,…,n*) *are saddle points of E*(**w**).

Proof. See Chauvin [Chauvin 89] and similar to Theorem 4.13 below. ■

Rate of Convergence

The rate of convergence of the PF algorithm can be obtained from (4.55) and (4.56). The time constants for $a_i(t)$ are $1/(\lambda_1 - \lambda_i)$ for *i*=2,…,*n*, and for $a_1(t)$ is $1/(\lambda_1 + \mu)$. The time constants are dependent on the eigen-structure of *A*.

Upper Bound of η_k

Theorem 4.10. Let A4.1 hold. Then there exists a uniform upper bound for η_k such that $\mathbf{w}_k$ is uniformly bounded w.p.1. Furthermore, if $\|\mathbf{w}_k\|^2 \le \alpha+1$, and θ is the largest eigenvalue of A_k, then $\|\mathbf{w}_{k+1}\|^2 \le \|\mathbf{w}_k\|^2$ if:

$$\eta_k < \frac{1}{\mu\alpha - \theta}, \text{ assuming } \mu\alpha>\theta. \tag{4.58}$$

Proof. Similar to Theorem 4.5. ■

4.8. Augmented Lagrangian 1 (AL1) Algorithm

Objective Function and Adaptive Algorithm

The objective function for Augmented Lagrangian 1 (AL1) algorithm is obtained by applying the Augmented Lagrangian method of nonlinear optimization to minimize $-\mathbf{w}_k^T A_k \mathbf{w}_k$ under the constraint $\mathbf{w}_k^T \mathbf{w}_k = 1$:

$$J(\mathbf{w}_k;A_k) = -\mathbf{w}_k^T A_k \mathbf{w}_k + \alpha_k\left(\mathbf{w}_k^T \mathbf{w}_k - 1\right) + \frac{\mu}{2}\left(\mathbf{w}_k^T \mathbf{w}_k - 1\right)^2, \tag{4.59}$$

where α_k is a Lagrange multiplier, and μ is a positive penalty constant. The gradient of $J(\mathbf{w}_k;A_k)$ with respect to $\mathbf{w}_k$ is:

$$\nabla_{\mathbf{w}_k} J(\mathbf{w}_k;A_k) = -2\left(A_k\mathbf{w}_k - \alpha_k\mathbf{w}_k - \mu\mathbf{w}_k\left(\mathbf{w}_k^T\mathbf{w}_k - 1\right)\right). \tag{4.60}$$

By equating the gradient to **0** and using the constraint $\mathbf{w}_k^T\mathbf{w}_k = 1$, we obtain $\alpha_k = \mathbf{w}_k^T A_k \mathbf{w}_k$. Replacing this α_k in the gradient, we obtain the AL1 algorithm:

$$\mathbf{w}_{k+1} = \mathbf{w}_k + \eta_k\left(A_k\mathbf{w}_k - \mathbf{w}_k\mathbf{w}_k^T A_k\mathbf{w}_k - \mu\mathbf{w}_k\left(\mathbf{w}_k^T\mathbf{w}_k - 1\right)\right), \tag{4.61}$$

where $\mu > 0$. Note that (4.61) is the same as OJA+ algorithm (see Sec. 4.2.1) for $\mu = 1$.

Convergence Analysis

The ordinary differential equation (ODE) associated with the adaptive algorithm (4.61) is:

$$\frac{d}{dt}\mathbf{w}(t) = A\mathbf{w} - \mathbf{w}\mathbf{w}^T A\mathbf{w} - \mu\mathbf{w}\left(\mathbf{w}^T\mathbf{w} - 1\right), \tag{4.62}$$

where $\mu > 0$ and $\mathbf{w}(t)$ is the continuous time counterpart of $\mathbf{w}_k$ with t denoting continuous time.

Theorem 4.11. Let A4.1 and A4.2 hold, and let $\mathbf{w}(t) = \sum_{i=1}^{n} a_i(t)\boldsymbol{\phi}_i$ be a solution to the ODE (4.62) in terms of the entire orthonormal set of eigenvectors $\{\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \ldots, \boldsymbol{\phi}_n\}$ of A. Then for any initial condition $\mathbf{w}(0) = \mathbf{w}_0 \in \Re^n$, $\mathbf{w}_0^T\boldsymbol{\phi}_1 \neq 0$, the solutions for the coefficients $a_i(t)$ for $t \geq 0$ are given by:

$$\frac{a_i(t)}{a_1(t)} = \frac{a_i(0)}{a_1(0)} e^{-(\lambda_1 - \lambda_i)t} \text{ for } i=2,\ldots,n \tag{4.63}$$

and

$$a_1(t) = \pm\sqrt{\frac{k_0}{e^{-2(\lambda_1+\mu)t} + k_0}} \text{ where } k_0 = \frac{a_1(0)^2}{1 - a_1(0)^2}. \tag{4.64}$$

The points ±ϕ_1 *are* (*uniformly*) *asymptotically stable. The domain of attraction of* ϕ_1 *is* $D(\phi_1)=\{\mathbf{w}\in\Re^n \mid \mathbf{w}^T\phi_1>0\}$ *and that of* $-\phi_1$ *is* $D(-\phi_1)=\{\mathbf{w}\in\Re^n \mid \mathbf{w}^T\phi_1<0\}$.

Proof. Similar to Theorem 4.3. ■

It is clear from (4.63) and (4.64) that $a_1(t)\to\pm1$ and $a_i(t)\to0$ for i=2,…,n as $t\to\infty$. Thus, $\mathbf{w}(t)\to\pm\phi_1$ as $t\to\infty$. It is clear from Theorem 4.11 that the AL1 adaptive algorithm converges to the normalized principal eigenvector of A, as compared to a weighted version in the PF algorithm.

Rate of Convergence

The rate of convergence of the AL1 algorithm can be obtained from (4.63) and (4.64). The time constants for $a_i(t)$ are $1/(\lambda_1-\lambda_i)$ for i=2,…,n, and for $a_1(t)$ is $1/(\lambda_1+\mu)$. The time constants are dependent on the eigen-structure of A.

Upper Bound of η_k

Theorem 4.12. Let A4.1 hold. Then there exists a uniform upper bound for η_k such that $\mathbf{w}_k$ is uniformly bounded w.p.1. Furthermore, if $\|\mathbf{w}_k\|^2 \le \alpha+1$, and θ is the largest eigenvalue of A_k, then $\|\mathbf{w}_{k+1}\|^2 \le \|\mathbf{w}_k\|^2$ if:

$$\eta_k < \frac{1}{(\mu+\theta)\alpha}. \tag{4.65}$$

Proof. Similar to Theorem 4.5. ■

4.9. Augmented Lagrangian 2 (AL2) Algorithm

Objective Function

The objective function for Augmented Lagrangian 2 (AL2) algorithm is:

$$J(\mathbf{w}_k;A_k) = -\mathbf{w}_k^T A_k \mathbf{w}_k + \mathbf{w}_k^T A_k \mathbf{w}_k\left(\mathbf{w}_k^T\mathbf{w}_k - 1\right) + \frac{\mu}{2}\left(\mathbf{w}_k^T\mathbf{w}_k - 1\right)^2, \quad \mu > 0. \tag{4.66}$$

The objective function $J(\mathbf{w}_k;A_k)$ is an application of the Augmented Lagrangian method on the Rayleigh Quotient criterion (4.1). It uses the XU objective function and also uses the Penalty Function (PF) method (4.52), where μ is a positive penalty constant.

Adaptive Algorithm

The gradient of (4.66) with respect to $\mathbf{w}_k$ is:

$$(1/2)\nabla_{\mathbf{w}_k} J(\mathbf{w}_k; A_k) = -(2A_k\mathbf{w}_k - \mathbf{w}_k\mathbf{w}_k^T A_k\mathbf{w}_k - A_k\mathbf{w}_k\mathbf{w}_k^T\mathbf{w}_k - \mu\mathbf{w}_k(\mathbf{w}_k^T\mathbf{w}_k - 1)). \qquad (4.67)$$

The adaptive gradient descent AL2 algorithm for PCA is:

$$\mathbf{w}_{k+1} = \mathbf{w}_k + \eta_k\left(2A_k\mathbf{w}_k - \mathbf{w}_k\mathbf{w}_k^T A_k\mathbf{w}_k - A_k\mathbf{w}_k\mathbf{w}_k^T\mathbf{w}_k - \mu\mathbf{w}_k(\mathbf{w}_k^T\mathbf{w}_k - 1)\right), \qquad (4.68)$$

where $\mu > 0$.

Convergence Analysis

The ordinary differential equation (ODE) associated with the adaptive algorithm (4.68) is:

$$\frac{d}{dt}\mathbf{w}(t) = 2A\mathbf{w} - \mathbf{w}\mathbf{w}^T A\mathbf{w} - A\mathbf{w}\mathbf{w}^T\mathbf{w} - \mu\mathbf{w}\left(\mathbf{w}^T\mathbf{w} - 1\right), \qquad (4.69)$$

where $\mu > 0$, and **w**(*t*) is the continuous time counterpart of $\mathbf{w}_k$ with *t* denoting continuous time.

We next analyze the stable stationary points of the ODE (4.69). The objective function *J* in (4.66) is also an energy function *E*(**w**) for the adaptive algorithm (4.68) as:

$$E(\mathbf{w}) = -2\mathbf{w}^T A\mathbf{w} + \mathbf{w}^T A\mathbf{w}\mathbf{w}^T\mathbf{w} + \frac{\mu}{2}\left(\mathbf{w}^T\mathbf{w} - 1\right)^2. \qquad (4.70)$$

The stable stationary points of the energy function *E*(**w**) are shown in Theorem 4.13 below.

Theorem 4.13. *Let A4.1 and A4.2 hold. The critical points of E*(**w**) *in* (4.70) *are* **0** *and* ±ϕ_i (*i*=1,…,*n*). *The two critical points* ±ϕ_1 *are global minimum points of E*(**w**)*, and E*(**w**) *has no other local minimum point. The point* **0** *is a local maximum. In addition,* ±ϕ_i (*i*=2,…,*n*) *are saddle points of E*(**w**)*.*

Proof. Substituting $\mathbf{w} = \sum_{i=1}^{n} a_i(t)\phi_i$ in the ODE (4.69) (where {ϕ_1,ϕ_2,…,ϕ_n} are orthonormal eigenvectors of *A*), and multiplying it on the left by ϕ_k^T (*k*=1,…,*n*), and then equating it to 0, we get:

$$a_k\left(-(2\lambda_k + \mu) + \sum_{r=1}^{n} a_r^2(\lambda_r + \lambda_k + \mu)\right) = 0 \qquad \text{for } k=1,\ldots,n, \qquad (4.A.6)$$

which gives us:

$$a_k = 0 \text{ or } \sum_{r=1}^{n} a_r^2(\lambda_r + \lambda_k + \mu) = 2\lambda_k + \mu \qquad \text{for } k=1,\ldots,n. \qquad (4.A.7)$$

Rewriting the second equation of (4.A.7), we obtain:

$$\sum_{r=1,r\neq k}^{n} a_r^2(\lambda_r+\lambda_k+\mu)+(a_k^2-1)(2\lambda_k+\mu)=0 \qquad \text{for } k=1,\ldots,n. \tag{4.A.8}$$

Since A is positive definite (i.e., λ_i>0 by Assumption A4.1), μ>0, and since λ_1 unit multiplicity (Assumption A4.2), from (4.A.8) we obtain at most one non-zero a_k=±1. Thus, the equilibrium points of E(**w**) are **w**=$d_{(1)}\phi_{(1)}$, where $\phi_{(1)}$ is a permutation of the eigenvectors $\phi_1,\ldots,\phi_n$, and $d_{(1)}$= 0 or ±1.

Now let us assume that $d_{(1)}$=0, i.e. **w**=0. Then, from (4.70), E(0)=μ/2. Next, we perturb **w** by $\delta\phi_1$. We observe that $E(\delta\phi_1)-E(0) = (2\lambda_1+\mu)(\delta^4-2\delta^2) < 0$ for $\delta<\sqrt{2}$. Clearly, the energy function E(**w**) decreases. We next prove that if $\phi_{(1)}\neq\pm\phi_1$, then the critical points are unstable equilibrium points of E(**w**). We obtain the Hessian of E(**w**) with respect to **w** as

$$\nabla_{\mathbf{w}}^2 E(\mathbf{w}) = A\mathbf{w}^T\mathbf{w}+2\mathbf{w}\mathbf{w}^T A+\mathbf{w}^T A\mathbf{w}I+2A\mathbf{w}\mathbf{w}^T+\mu(\mathbf{w}^T\mathbf{w}-1)I+2\mu\mathbf{w}\mathbf{w}^T-2A. \tag{4.A.9}$$

We note that

$$\nabla_{\mathbf{w}}^2 E(\pm\phi_1) = -A+\lambda_1 I+4\lambda_1\phi_1\phi_1^T+2\mu\phi_1\phi_1^T, \tag{4.A.10}$$

whose eigenvectors are $\phi_1,\ldots,\phi_n$, and eigenvalues are $2(2\lambda_1+\mu)$ for ϕ_1, and $(\lambda_1-\lambda_r)$ for ϕ_r (r>1). Clearly, $\nabla_{\mathbf{w}}^2 E(\pm\phi_1)$ is positive definite. On the other hand, we observe that

$$\nabla_{\mathbf{w}}^2 E(\pm\phi_r) = -A+\lambda_r I+4\lambda_r\phi_r\phi_r^T+2\mu\phi_r\phi_r^T \quad \text{for } r>1. \tag{4.A.11}$$

The eigenvectors of $\nabla_{\mathbf{w}}^2 E(\pm\phi_r)$ are $\phi_1,\ldots,\phi_n$, and the eigenvalues are $-(\lambda_1-\lambda_r)$ for ϕ_1, and $2(2\lambda_r+\mu)$ for ϕ_r (r>1). Clearly, $\nabla_{\mathbf{w}}^2 E(\pm\phi_r)$ for r>1 is an indefinite matrix. Thus, $\pm\phi_1$ is a stable equilibrium point of the energy function E(**w**), whereas $\pm\phi_r$ for r>1 are unstable equilibrium points. ■

It is clear from Theorem 4.13 that the AL2 adaptive algorithm converges to the normalized principal eigenvector of A, as compared to a weighted version in the PF algorithm.

Rate of Convergence

The time constants for $a_i(t)$ are $1/(\lambda_1-\lambda_i)$ for i=2,…,n, and for $a_1(t)$ is $1/(\lambda_1+(\mu/2))$. The time constants are dependent on the eigen-structure of A.

Upper Bound of η_k

Theorem 4.14. Let A4.1 hold. Then there exists a uniform upper bound for η_k such that $\mathbf{w}_k$ is uniformly bounded w.p.1. Furthermore, if $\|\mathbf{w}_k\|^2 \le \alpha+1$, and θ is the largest eigenvalue of A_k, then $\|\mathbf{w}_{k+1}\|^2 \le \|\mathbf{w}_k\|^2$ if:

$$\eta_k < \frac{2}{(2\theta + \mu)\alpha}. \tag{4.71}$$

Proof. Similar to Theorem 4.5. ■

5. Proofs for Chapter 5

5.2. Algorithms and Objective Functions

Proofs of Convergence and Assumptions

In order to prove the convergence, the adaptive algorithms, we use Stochastic Approximation theory [Ljung77,92; Benveniste *et al.* 90]. However, due to the repetitive nature of the proofs and since some of these algorithms have been proven by other practitioners, we state all the results, and prove the convergence for PF and AL2 algorithms. We require the following assumptions:

Assumption (A5.1). Same as A2.1 in Section 2.4.1.

Assumption (A5.2). Same as A2.2 in Section 2.4.1.

Assumption (A5.3). The $p \le n$ largest eigenvalues of *A* are each of unit multiplicity.

In the following discussions, we denote $\Phi=[\phi_1 \; \ldots \; \phi_n] \in \Re^{n\mathrm{X}n}$ as the orthonormal eigenvector matrix of *A*, and $\Lambda= diag(\lambda_1,\ldots,\lambda_n)$ as the eigenvalue matrix, such that $\lambda_1>\lambda_2>\ldots>\lambda_p>\lambda_{p+1}\ge\ldots\ge\lambda_n>0$. We use the subscript (*i*) to denote the *i*th permutation of the indices {1,2,…,*n*}.

5.5. PF Algorithms

PF Homogeneous Algorithm

We obtain the objective function for Penalty Function (PF) Homogeneous PCA algorithm by expressing the Rayleigh quotient criterion as the following penalty function:

$$J(\mathbf{w}_k^i; A_k) = -\mathbf{w}_k^{i\,T} A_k \mathbf{w}_k^i + \mu\left(\sum_{j=1,\, j\ne i}^{p}\left(\mathbf{w}_k^{j\,T}\mathbf{w}_k^i\right)^2 + \frac{1}{2}\left(\mathbf{w}_k^{i\,T}\mathbf{w}_k^i - 1\right)^2\right), \tag{5.18}$$

where $\mu>0$ and $i=1,\ldots,p$. From the gradient of (5.18) with respect to $\mathbf{w}_k^i$, we obtain the PF Homogeneous adaptive gradient descent algorithm for PCA as:

$$W_{k+1} = W_k + \eta_k \left(A_k W_k - \mu W_k \left(W_k^T W_k - I_p\right)\right), \qquad (5.19)$$

where I_p is a pXp identity matrix.

The ODE associated with (5.19) is:

$$\frac{d}{dt} W(t) = AW - \mu W\left(W^T W - I_p\right). \qquad (5.20)$$

We look for the stable equilibrium points of the ODE, and the convergence properties of its initial states close to the stable points. In the following theorems, we recall that $\lambda_1 > \lambda_2 > \ldots > \lambda_p > \lambda_{p+1} \geq \ldots \geq \lambda_n > 0$ are the eigenvalues of A, and ϕ_i as the eigenvector corresponding to λ_i such that $\Phi = [\phi_1 \ \ldots \ \phi_n]$ are orthonormal. We use the subscript (i) to denote the i^{th} permutation of the indices {1,2,…,n}.

Theorem 5.1. For the ordinary differential equation (5.20), let A5.1 and A5.3 hold. Then $W = \Phi$ PDU are equilibrium points of (5.20), where $D = [D_1 | 0]^T \in \Re^{nXp}$ with $D_1 \in \Re^{pXp}$ is diagonal with elements $d_i = 0$ or $\pm\sqrt{1 + (\lambda_{(i)} / \mu)}$, $P \in \Re^{nXn}$ is an arbitrary permutation matrix, and $U \in \Re^{pXp}$ is an arbitrary rotation matrix, i.e., $U^T U = UU^T = I_p$.

Proof. From (5.20), we need to find a $W \in \Re^{nXp}$ such that

$$AW - \mu W(W^T W - I_p) = 0. \qquad (5.21)$$

The trivial solution is $W=0$. We next assume that $W \neq 0$. Let $W=QDU$ be the singular value decomposition of W, where $Q \in \Re^{nXn}$ and $U \in \Re^{pXp}$ are orthonormal, and $D \in \Re^{nXp}$ is diagonal. Replacing QDU for W in (5.21) and defining $B = Q^T A Q$, we get from (5.21):

$$BD - \mu D(D^T D - I_p) = 0. \qquad (5.22)$$

Let $D = [D_1 | 0]^T$ where D_1 is a pXp diagonal matrix with diagonal elements d_i for i=1,…,p. Let $B = \begin{bmatrix} B_1 & B_2 \\ B_2^T & B_3 \end{bmatrix}$ be a partition of B where $B_1 \in \Re^{pXp}$. From (5.22) we get:

$$\left(B_1 - \mu\left(D_1^2 - I_p\right)\right) D_1 = 0 \text{ and } \quad B_2^T D_1 = 0. \qquad (5.23)$$

From (5.23), we get:

$$D_1 = 0 \text{ or } B_1 = \mu(D_1^2 - I_p) \text{ and } B_2 = 0. \qquad (5.24)$$

From (5.24), we conclude that B$_1$ is diagonal, and $d_i^2 = 1 + (b_i / \mu)$ for i=1,…,p, where b_i are the diagonal elements of B_1. Since $B = Q^T A Q$ is diagonal, where Q is orthonormal ($Q^T Q = QQ^T = I_n$), the columns of Q are the n orthonormal eigenvectors of A, i.e., $Q = \Phi P$, where $P \in \Re^{nXn}$ is an arbitrary permutation matrix. Also, $B = P^T \Lambda P$, i.e., a permutation of Λ. Therefore, $b_{(i)} = \lambda_{(i)}$ for i=1,…,p. Combining all results, we get W=ΦPDU, where D=$[D_1|0]^T\in\Re^{\mathrm{nXp}}$ and $D_1\in\Re^{pXp}$ is a diagonal matrix with elements d_i=0 or $\pm\sqrt{1+(\lambda_{(i)}/\mu)}$ for i=1,…,p.

■

Theorem 5.2. Let A5.1 and A5.3 hold. Then W=ΦDU, where D=$[D_1|0]^{\mathrm{T}}\in\Re^{\mathrm{nXp}}$, D_1=diag(d_1,…,d_p)$\in\Re^{pXp}$, d_i=$\pm\sqrt{1+(\lambda_i/\mu)}$ for i=1,…,p, and $U\in\Re^{pXp}$ is an arbitrary rotation matrix, are stable equilibrium points of the ODE (5.20) and strict global minimum points of the objective function (5.18). In addition, W=ΦPDU, where d_i=0 for $i\leq p$ or $P\neq I$, are unstable equilibrium points of the ODE (5.20).

Proof. From (5.18), the energy function $E(W)$ for the PF Homogeneous adaptive algorithm is:

$$E(W)=-tr\left(W^TAW\right)+\frac{\mu}{2}tr\left(\left(W^TW-I_p\right)^2\right). \tag{5.25}$$

From Theorem 5.1, W=ΦPDU=ΨU, where Ψ=ΦPD. Then from (5.A.5):

$$E(W)=\sum_{i=1}^{p}\left(-\boldsymbol{\psi}_i^TA\boldsymbol{\psi}_i+\mu\left(\sum_{j=1,\,j\neq i}^{p}\left(\boldsymbol{\psi}_i^T\boldsymbol{\psi}_j\right)^2+\left(\boldsymbol{\psi}_i^T\boldsymbol{\psi}_i-1\right)^2\right)\right).$$

Here $\boldsymbol{\psi}_i$=$d_{(i)}\boldsymbol{\phi}_{(i)}$ is the i^{th} column of Ψ for i=1,…,p, and $d_{(i)}$=0 or $\pm\sqrt{1+(\lambda_{(i)}/\mu)}$. We first prove that $d_{(i)}$=0 is an unstable equilibrium point of E(W). We perturb $\boldsymbol{\psi}_i$ by $\delta\boldsymbol{\phi}_i$, for δ>0. Then

$$E\left(\boldsymbol{\psi}_i=\delta\boldsymbol{\phi}_i\right)-E\left(\boldsymbol{\psi}_i=0\right)=-\delta^2\lambda_i+\mu(\delta^2-1)^2<0$$

$$\text{for }\left(\lambda+2\mu-\sqrt{\lambda(\lambda+4\mu)}\right)/2\mu<\delta^2<\left(\lambda+2\mu+\sqrt{\lambda(\lambda+4\mu)}\right)/2\mu,$$

i.e., the energy function decreases. We next prove that $\boldsymbol{\psi}_i$=$d_r\boldsymbol{\phi}_r$, d_r=$\pm\sqrt{1+(\lambda_r/\mu)}$, ($r$>$p$) is an unstable equilibrium point of $E(W)$. We perturb $\boldsymbol{\psi}_i$ by $\delta\boldsymbol{\phi}_p$, i.e., $\boldsymbol{\psi}_i=d_r(\boldsymbol{\phi}_r+\delta\boldsymbol{\phi}_p)/\sqrt{(1+\delta^2)}$. Then

$$E\left(d_r(\boldsymbol{\phi}_r+\delta\boldsymbol{\phi}_p)/\sqrt{(1+\delta^2)}\right)-E\left(d_r\boldsymbol{\phi}_r\right)=-\delta^2d_r^2(\lambda_p-\lambda_r)/(1+\delta^2)<0.$$

We next prove that $\boldsymbol{\psi}_i$=$d_r\boldsymbol{\phi}_{\mathrm{r}}$, d_r=$\pm\sqrt{1+(\lambda_r/\mu)}$, ($r\leq p$) is a stable equilibrium point of $E(W)$. We perturb $\boldsymbol{\psi}_i$ by $\delta\boldsymbol{\phi}_s$, i.e., $\boldsymbol{\psi}_i=d_r(\boldsymbol{\phi}_r+\delta\boldsymbol{\phi}_s)/\sqrt{(1+\delta^2)}$ for s>p. Then

$$E\left(d_r(\boldsymbol{\phi}_r+\delta\boldsymbol{\phi}_s)/\sqrt{(1+\delta^2)}\right)-E\left(d_r\boldsymbol{\phi}_r\right)=\delta^2d_r^2(\lambda_r-\lambda_s)/(1+\delta^2)>0,$$

i.e., the energy function increases. Thus, the columns of Ψ consisting of the first p orthonormal eigenvectors $\boldsymbol{\phi}_i$ of A scaled by d_i are the stable minimum points of $E(W)$.

■

Finally, the convergence of algorithm (5.19) can be established by referring to Theorem 1 of Ljung [Ljung 77].

PF Deflation Algorithm

The objective function for PF Deflation PCA algorithm is:

$$J(\mathbf{w}_k^i;A_k) = -\mathbf{w}_k^{i\,T} A_k \mathbf{w}_k^i + \mu\left(\sum_{j=1}^{i-1}\left(\mathbf{w}_k^{j\,T}\mathbf{w}_k^i\right)^2 + \frac{1}{2}\left(\mathbf{w}_k^{i\,T}\mathbf{w}_k^i - 1\right)^2\right), \tag{5.26}$$

where $\mu > 0$ and i=1,…,p. The PF Deflation adaptive gradient descent algorithm for PCA is:

$$W_{k+1} = W_k + \eta_k\left(A_k W_k - \mu W_k \mathrm{UT}\left(W_k^T W_k - I_p\right)\right), \tag{5.27}$$

where UT[·] sets all elements below the diagonal of its matrix argument to zero. The ODE associated with (5.27) is:

$$\frac{d}{dt}W(t) = AW - \mu W \mathrm{UT}\left(W^T W - I_p\right). \tag{5.28}$$

Theorem 5.3. *Let A5.1 and A5.3 hold. Then, all the equilibrium points of the ODE* (5.28) *are up to an arbitrary permutation of the eigenvectors of A weighted by* 0 *or* $\pm\sqrt{1+(\lambda_{(i)}/\mu)}$*, i.e., any point W* = [$d_{(1)}\phi_{(1)}$ $d_{(2)}\phi_{(2)}$ …$d_{(p)}\phi_{(p)}$]*, where* $d_{(i)}$=0 *or* $\pm\sqrt{1+(\lambda_{(i)}/\mu)}$ *is an equilibrium point of the ODE* (5.28).

Proof. We need to find a $W \in \Re^{n\mathrm{X}p}$ such that

$$AW - \mu W \mathrm{UT}\left(W^T W - I_p\right) = 0. \tag{5.29}$$

The trivial solution is W=0. We next assume that $W \neq 0$. We multiply (5.29) on the left by W^T, and define $G = W^T A W + (\mu/2) W^T W$, and $H = W^T W$. From (5.29), we obtain:

$$G = H\,\mathrm{UT}(H). \tag{5.30}$$

Since G is symmetric, HUT(H) is also symmetric. Since W is assumed to be nonzero, H has positive diagonal elements. From (5.30), we obtain that H and G are diagonal, i.e., both $W^T W$ and $W^T A W$ are diagonal. Then, W is of the form $W = \Phi P D$, where $P \in \Re^{n\mathrm{X}n}$ is an arbitrary permutation matrix, and $D = [D_1 \,|\, 0]^T \in \Re^{n\mathrm{X}p}$ where $D_1 \in \Re^{p\mathrm{X}p}$ is a diagonal matrix with elements d_i for i=1,…,p. Substituting $W = \Phi P D$ in (5.29), we obtain $\Lambda_1 D_1 = \mu D_1 (D_1^2 - I_p)$, where Λ_1 is a pXp partition of the permuted eigenvector matrix Λ of A. We obtain:

$$D_1 = 0 \text{ or } D_1^2 = I_p + (\Lambda_1/\mu). \tag{5.31}$$

From the second equation in (5.31), we obtain $d_i^2 = 1 + (\lambda_{(i)}/\mu)$. Combining all results, we conclude that $W = \Phi P D$, where d_i = 0 or $\pm\sqrt{1+(\lambda_{(i)}/\mu)}$. ■

Theorem 5.4. *Let A5.1 and A5.3 hold. Then, the points* $W^*=[d_1\phi_1\ d_2\phi_2\ \ldots\ d_p\phi_p]$*, where* $d_i = \pm\sqrt{1+(\lambda_i/\mu)}$*, are the strict global minimum points of the objective function* (5.20) *and stable equilibrium points of the ODE* (5.28). *In addition, the points* $W = [d_{(1)}\phi_{(1)}\ d_{(2)}\phi_{(2)}\ \ldots\ d_{(p)}\phi_{(p)}]$*, where* $d_{(i)} = 0$ *or* $\phi_{(i)} \neq \phi_i$ *for* $i\in\{1,2,\ldots,p\}$ *are unstable equilibrium points of the ODE* (5.28).

Proof. From (5.20), the energy function $E(W)$ for the PF Deflation adaptive algorithm is:

$$E(W)=\sum_{i=1}^{p}E_i(\mathbf{w}_i)\text{, where } E_i(\mathbf{w}_i)=-\mathbf{w}_i^TA\mathbf{w}_i+\mu\left(\sum_{j=1}^{i-1}\left(\mathbf{w}_j^T\mathbf{w}_i\right)^2+\frac{1}{2}\left(\mathbf{w}_i^T\mathbf{w}_i-1\right)^2\right).$$

From Theorem 5.3, $\mathrm{w}_i=d_{(i)}\phi_{(i)}$, $d_{(i)}=0$, $\pm\sqrt{1+(\lambda_{(i)}/\mu)}$, is the i^{th} column of W for $i=1,\ldots,p$. We first prove that $d_{(i)}=0$ is an unstable equilibrium point of $E(W)$. We perturb w_i by $\delta\phi_i$, for $\delta>0$. Then

$$E(\mathbf{w}_i=\delta\boldsymbol{\phi}_i)-E(\mathbf{w}_i=0)=-\delta^2\lambda_i+(\mu/2)(\delta^2-1)^2<0$$
$$\text{for } \left(\lambda+\mu-\sqrt{\lambda(\lambda+2\mu)}\right)/\mu<\delta^2<\left(\lambda+\mu+\sqrt{\lambda(\lambda+2\mu)}\right)/\mu.$$

We next prove that $\mathrm{w}_i\neq d_i\phi_i$, $d_i=\pm\sqrt{1+(\lambda_i/\mu)}$, $(i\leq p)$ is an unstable equilibrium point of $E(W)$. If $\mathrm{w}_i\neq d_i\phi_i$ then there exists a pair of columns $\mathrm{w}_r=d_{(r)}\phi_{(r)}$, and $\mathrm{w}_s=d_{(s)}\phi_{(s)}$, such that $r<s\leq p$ and $\lambda_{(r)}<\lambda_{(s)}$. We perturb w_r by $\delta\phi_{(s)}$, i.e., $\mathbf{w}_r=d_{(r)}(\boldsymbol{\phi}_{(r)}+\delta\boldsymbol{\phi}_{(s)})/\sqrt{(1+\delta^2)}$. Then

$$E\left(\mathbf{w}_r=d_{(r)}(\boldsymbol{\phi}_{(r)}+\delta\boldsymbol{\phi}_{(s)})/\sqrt{(1+\delta^2)}\right)-E\left(\mathbf{w}_r=d_{(r)}\boldsymbol{\phi}_{(r)}\right)=-\delta^2d_{(r)}^2(\lambda_{(s)}-\lambda_{(r)})/(1+\delta^2)<0.$$

We next prove that $\mathrm{w}_i= d_i\phi_i$, $d_i=\pm\sqrt{1+(\lambda_i/\mu)}$, $(i\leq p)$ is a stable equilibrium point of $E(W)$. We perturb w_i by $\delta\phi_s$, i.e., $\mathrm{w}_i=d_i(\boldsymbol{\phi}_i+\delta\boldsymbol{\phi}_s)/\sqrt{(1+\delta^2)}$ for $s>i$. Then

$$E\left(\mathbf{w}_i=d_i(\boldsymbol{\phi}_i+\delta\boldsymbol{\phi}_s)/\sqrt{(1+\delta^2)}\right)-E\left(\mathbf{w}_i=d_i\boldsymbol{\phi}_i\right)=\delta^2d_i^2(\lambda_i-\lambda_s)/(1+\delta^2)>0,$$

i.e., the energy function increases. Thus, $W=[d_1\phi_1\ \ldots\ d_p\phi_p]$, $d_i=\pm\sqrt{1+(\lambda_i/\mu)}$, $(i\leq p)$ are the stable equilibrium points of $E(W)$. ■

PF Weighted Algorithm

The objective function for PF Weighted PCA algorithm is:

$$J(\mathbf{w}_k^i;A_k)=-c_i\mathbf{w}_k^{i\,T}A_k\mathbf{w}_k^i+\mu\left(\sum_{j=1,\,j\neq i}^{p}c_j\left(\mathbf{w}_k^{j\,T}\mathbf{w}_k^i\right)^2+\frac{c_i}{2}\left(\mathbf{w}_k^{i\,T}\mathbf{w}_k^i-1\right)^2\right), \tag{5.32}$$

where $c_1>c_2>\ldots>c_p>0$, $\mu>0$, and $i=1,\ldots,p$. The PF Weighted adaptive gradient descent algorithm for PCA is:

$$W_{k+1}=W_k+\eta_k\left(A_kW_kC-\mu W_kC\left(W_k^TW_k-I_p\right)\right), \tag{5.33}$$

where $C=diag(c_1,\ldots,c_p)$. The ODE associated with (5.33) is:

$$\frac{d}{dt}W(t)=AWC-\mu WC\left(W^TW-I_p\right). \tag{5.34}$$

Theorem 5.5. Let A5.1 and A5.3 hold. Then, all the equilibrium points of the ODE (5.34) are up to an arbitrary permutation of the eigenvectors of A weighted by 0 or $\pm\sqrt{1+(\lambda_{(i)}/\mu)}$, i.e., any point W = $[d_{(1)}\boldsymbol{\phi}_{(1)}\ d_{(2)}\boldsymbol{\phi}_{(2)}\ldots d_{(p)}\boldsymbol{\phi}_{(p)}]$, where $d_{(i)}$=0 or $\pm\sqrt{1+(\lambda_{(i)}/\mu)}$, is an equilibrium point of the ODE (5.34).

Proof. We need to find a $W\in\Re^{nXp}$ such that

$$AWC-\mu WC\left(W^TW-I_p\right)=0\,. \tag{5.35}$$

The trivial solution is W=0. We next assume that $W\neq 0$. We multiply (5.35) on the left by W^T, and define $G=W^TAW+(\mu/2)W^TW$, and $H=W^TW$. From (5.35), we obtain:

$$GC=\mu HCH. \tag{5.36}$$

Since H, G, and HCH are symmetric matrices, we conclude that $GC=CG$. Since C is diagonal with distinct diagonal elements, G is diagonal. Let $GC=CG=D\in\Re^{pXp}$ be a diagonal matrix. Then, from (5.36), we get:

$$D=\mu HCH. \tag{5.37}$$

Since W is assumed to be nonzero, the diagonal elements of $H=W^TW$ are positive. We conclude from (5.37) that H is diagonal. The rest of the proof is similar to Theorem 5.3 above.

■

Theorem 5.6. Let A5.1 and A5.3 hold. Then, the points $W^*=[d_1\boldsymbol{\phi}_1\ d_2\boldsymbol{\phi}_2\ \ldots\ d_p\boldsymbol{\phi}_p]$, where $d_i=\pm\sqrt{1+(\lambda_i/\mu)}$, are stable equilibrium points of the ODE (5.34) and strict global minimum points of the objective function in (5.32). In addition, the points $W=[d_{(1)}\boldsymbol{\phi}_{(1)}\ d_{(2)}\boldsymbol{\phi}_{(2)}\ \ldots\ d_{(p)}\boldsymbol{\phi}_{(p)}]$, where $d_{(i)}=0$ or $\boldsymbol{\phi}_{(i)}\neq\boldsymbol{\phi}_i$ for $i\in\{1,2,\ldots,p\}$ are unstable equilibrium points of the ODE (5.34).

Proof. From (5.32), the energy function E(W) for the PF Weighted adaptive algorithm is:

$$E(W)=\sum_{i=1}^{p}E_i(\mathbf{w}_i)\text{, where }E_i(\mathbf{w}_i)=-c_i\mathbf{w}_i^TA\mathbf{w}_i+\mu\left(\sum_{j=1,j\neq i}^{p}c_j\left(\mathbf{w}_j^T\mathbf{w}_i\right)^2+\frac{c_i}{2}\left(\mathbf{w}_i^T\mathbf{w}_i-1\right)^2\right).$$

From Theorem 5.5, $\mathbf{w}_i=d_{(i)}\boldsymbol{\phi}_{(i)}$, $d_{(i)}$=0 or $\pm\sqrt{1+(\lambda_{(i)}/\mu)}$, is the ith column of W for i=1,…,p. We first prove that $d_{(i)}$=0 is an unstable equilibrium point of $E(W)$. We perturb $\mathbf{w}_i$ by $\delta\boldsymbol{\phi}_i$. Then

$$E\left(\mathbf{w}_i=\delta\boldsymbol{\phi}_i\right)-E\left(\mathbf{w}_i=0\right)=-\delta^2c_i\lambda_i+(\mu/2)c_i(\delta^2-1)^2<0$$

$$\text{for }\left(\lambda+\mu-\sqrt{\lambda(\lambda+2\mu)}\right)/\mu<\delta^2<\left(\lambda+\mu+\sqrt{\lambda(\lambda+2\mu)}\right)/\mu\,.$$

The rest of the proof is similar to Theorem 5.4 above. ■

5.7. AL2 Algorithm

AL2 Homogeneous Algorithm

The AL2 objective function can be derived from AL1 Homogeneous objective function (5.27) by replacing α,β_1,β_2,…,β_p from (5.28) into (5.27) as:

$$J(\mathbf{w}_k^i;A_k) = -\mathbf{w}_k^{i\,T} A_k \mathbf{w}_k^i + \left(\mathbf{w}_k^{i\,T} A_k \mathbf{w}_k^i\right)\left(\mathbf{w}_k^{i\,T} \mathbf{w}_k^i - 1\right) + 2\sum_{j=1,\, j\neq i}^{p} \mathbf{w}_k^{i\,T} A_k \mathbf{w}_k^j \mathbf{w}_k^{j\,T} \mathbf{w}_k^i +$$
$$\mu\left(\sum_{j=1,\, j\neq i}^{p} \left(\mathbf{w}_k^{j\,T} \mathbf{w}_k^i\right)^2 + \frac{1}{2}\left(\mathbf{w}_k^{i\,T} \mathbf{w}_k^i - 1\right)^2\right), \qquad (5.38)$$

for i=1,…,p, and μ>0. As seen with the XU objective function, (5.38) also has the constraints $\mathbf{w}_k^{i\,T} \mathbf{w}_k^i = \delta_{ij}$ built into it. The AL2 Homogeneous adaptive gradient descent algorithm for PCA is:

$$W_{k+1} = W_k + \eta_k\left(2A_kW_k - W_kW_k^TA_kW_k - A_kW_kW_k^TW_k - \mu W_k\left(W_k^TW_k - I_p\right)\right), \qquad (5.39)$$

where I_p is a pXp identity matrix. The ODE associated with (5.39) is:

$$\frac{d}{dt}W(t) = 2AW - WW^TAW - AWW^TW - \mu W(W^TW - I_p). \qquad (5.40)$$

We look for the stable equilibrium points of the ODE, and the convergence properties of its initial states close to the stable points.

Theorem 5.7. For the ordinary differential equation (5.40), let A5.1 and A5.3 hold. Then W=Φ PDU are equilibrium points of (5.46), where D=[D_1|0]^T$\in\Re^{nXp}$ with $D_1\in\Re^{pXp}$ is diagonal with elements d_i=+1, –1 or 0, $P\in\Re^{nXn}$ is an arbitrary permutation matrix, and $U\in\Re^{pXp}$ is an arbitrary rotation matrix, i.e., U^TU=UU^T=I_p.

Proof. We need to find a $W\in\Re^{nXp}$ such that

$$2AW - WW^TAW - AWW^TW - \mu W(W^TW - I_p) = 0. \qquad (5.41)$$

The trivial solution is W=0. We next assume that W≠0. Let W=QDV be the singular value decomposition of W, where $Q\in\Re^{nXn}$, and $V\in\Re^{pXp}$ are orthonormal, and $D\in\Re^{nXp}$ is diagonal. Replacing QDV for W in (5.41) and defining B=Q^TAQ, we get from (5.41):

$$2BD - DD^T BD - BDD^T D - \mu D(D^T D - I_p) = 0. \tag{5.42}$$

Let D=[D_1|0]^T where D_1 is a pXp diagonal matrix with diagonal elements d_i for i=1,…,p. Let $B = \begin{bmatrix} B_1 & B_2 \\ B_2^T & B_3 \end{bmatrix}$ be a partition of B where $B_1 \in \Re^{pXp}$. From (5.42) we get:

$$\left(2B_1 - D_1^2 B_1 - B_1 D_1^2 - \mu\left(D_1^2 - I_p\right)\right) D_1 = 0\,. \tag{5.43}$$

We conclude:

$$D_1 = 0 \text{ or } 2B_1 - D_1^2 B_1 - B_1 D_1^2 - \mu(D_1^2 - I_p) = 0\,. \tag{5.44}$$

From the second equation in (5.44), we obtain:

$$(2b_{ii} + \mu)(d_i^2 - 1) = 0 \text{ for i=j and } b_{ij}(d_i^2 + d_j^2 - 2) = 0 \text{ for } i \neq j, \tag{5.45}$$

where B_1=[b_{ij}] for i,j=1,…,p. Since μ>0, we satisfy both equations with d_i=±1. Thus, W=QDV, where d_i=±1 for i=1,…,p. Clearly, W^TW=I_p. Then, from (5.41), we obtain:

$$AW = WW^TAW. \tag{5.46}$$

Let W^TAW=U^TLU be the eigen-decomposition of W^TAW, where $U \in \Re^{pXp}$ is orthonormal, and $L \in \Re^{pXp}$ is diagonal. From (5.46), we get A(WU^T) = (WU^T)L, where WU^T is orthonormal. Then, the columns of WU^T consist of the p orthonormal eigenvectors of A, i.e., WU^T = ΦPD, where $P \in \Re^{nXn}$ is an arbitrary permutation matrix, and D=[D$_1$|0]$^T \in \Re^{\mathrm{nXp}}$ where D$_1 \in \Re^{\mathrm{pXp}}$ is a diagonal matrix with elements d_i=±1 for i=1,…,p. Combining all results, we conclude that W = ΦPDU, where U is orthonormal, and d$_i$ = 0 or ±1. ■

Theorem 5.8. Let A5.1 and A5.3 hold. Then W=ΦDU, where D=[D_1|0]$^T \in \Re^{nXp}$, D$_1$= diag(d_1,…,d_p)$\in \Re^{pXp}$, d_i=±1 for i=1,…,p, and $U \in \Re^{pXp}$ is an arbitrary rotation matrix, are stable equilibrium points of the ODE (5.40) and strict global minimum points of the objective function (5.38). In addition, W=ΦPDU, where d$_i$=0 for i≤p or P≠I, are unstable equilibrium points of the ODE (5.40).

Proof. From (5.38), the energy function E(W) for the AL2 Homogeneous adaptive algorithm is:

$$E(W) = -2tr\left(W^T AW\right) + tr\left(W^T AWW^T W\right) + \frac{\mu}{2} tr\left(\left(W^T W - I_p\right)^2\right). \tag{5.47}$$

From Theorem 5.7, W=ΦPDU=ΨU, where Ψ=ΦPD. Then from (5.47):

$$E(W) = \sum_{i=1}^{p} E_i(\mathbf{w}_i)\text{, where}$$

$$E_i(\mathbf{w}_i) = -2\boldsymbol{\psi}_i^T A\boldsymbol{\psi}_i + \boldsymbol{\psi}_i^T A\boldsymbol{\psi}_i\boldsymbol{\psi}_i^T\boldsymbol{\psi}_i + \sum_{j=1,\neq i}^{p} \boldsymbol{\psi}_j^T A\boldsymbol{\psi}_i\boldsymbol{\psi}_j^T\boldsymbol{\psi}_i + \mu\left(\sum_{j=1,\neq i}^{p}\left(\boldsymbol{\psi}_j^T\boldsymbol{\psi}_i\right)^2 + \left(\boldsymbol{\psi}_i^T\boldsymbol{\psi}_i - 1\right)^2\right).$$

Here $\boldsymbol{\psi}_i$=$d_{(i)}\boldsymbol{\phi}_{(i)}$ is the i^{th} column of Ψ for i=1,…,p, and $d_{(i)}$=0 or ±1. We first prove that $d_{(i)}$=0 is an unstable equilibrium point of $E(W)$. We perturb $\boldsymbol{\psi}_i$ by $\delta\boldsymbol{\phi}_i$. Then

$$E(\boldsymbol{\psi}_i = \delta\boldsymbol{\phi}_i) - E(\boldsymbol{\psi}_i = 0) = -2\delta^2\lambda_i + \delta^4\lambda_i + \mu(\delta^2 - 1)^2 < 0$$

$$\text{for } \left(\lambda + \mu - \sqrt{\lambda(\lambda+\mu)}\right)\big/(\lambda+\mu) < \delta^2 < \left(\lambda + \mu + \sqrt{\lambda(\lambda+\mu)}\right)\big/(\lambda+\mu).$$

We next prove that $\boldsymbol{\psi}_i$=±$\boldsymbol{\phi}_r$,, (r>p) is an unstable equilibrium point of E(W). We perturb $\boldsymbol{\psi}_i$ by $\delta\boldsymbol{\phi}_p$, i.e., $\boldsymbol{\psi}_i = \pm(\boldsymbol{\phi}_r + \delta\boldsymbol{\phi}_p)/\sqrt{(1+\delta^2)}$. Then

$$E\left(\pm(\boldsymbol{\phi}_r + \delta\boldsymbol{\phi}_p)/\sqrt{(1+\delta^2)}\right) - E(\pm\boldsymbol{\phi}_r) = -\delta^2(\lambda_p - \lambda_r)/(1+\delta^2) < 0.$$

We next prove that $\boldsymbol{\psi}_i$=±$\boldsymbol{\phi}_r$, (r≤p) is a stable equilibrium point of E(W). We perturb $\boldsymbol{\psi}_i$ by $\delta\boldsymbol{\phi}_s$, i.e., $\boldsymbol{\psi}_i = \pm(\boldsymbol{\phi}_r + \delta\boldsymbol{\phi}_s)/\sqrt{(1+\delta^2)}$ for s>p. Then

$$E\left(\pm(\boldsymbol{\phi}_r + \delta\boldsymbol{\phi}_s)/\sqrt{(1+\delta^2)}\right) - E(\pm\boldsymbol{\phi}_r) = \delta^2(\lambda_r - \lambda_s)/(1+\delta^2) > 0.$$

Thus, the columns of Ψ consisting of the first p orthonormal eigenvectors of A are the stable minimum points of $E(W)$. ■

AL2 Deflation Algorithm

The objective function for AL2 Deflation PCA algorithm is:

$$J(\mathbf{w}_k^i; A_k) = -\mathbf{w}_k^{i\,T} A_k \mathbf{w}_k^i + \left(\mathbf{w}_k^{i\,T} A_k \mathbf{w}_k^i\right)\left(\mathbf{w}_k^{i\,T}\mathbf{w}_k^i - 1\right) + 2\sum_{j=1}^{i-1} \mathbf{w}_k^{i\,T} A_k \mathbf{w}_k^j \mathbf{w}_k^{j\,T}\mathbf{w}_k^i +$$

$$\mu\left(\sum_{j=1}^{i-1}\left(\mathbf{w}_k^{j\,T}\mathbf{w}_k^i\right)^2 + \frac{1}{2}\left(\mathbf{w}_k^{i\,T}\mathbf{w}_k^i - 1\right)^2\right), \tag{5.48}$$

for i=1,…,p and μ > 0. Taking the gradient of (5.48) with respect to $\mathbf{w}_k^i$ we obtain the AL2 Deflation adaptive gradient descent algorithm for PCA as:

$$W_{k+1} = W_k + \eta_k\left(2A_kW_k - W_k\mathrm{UT}\left(W_k^T A_k W_k\right) - A_kW_k\mathrm{UT}\left(W_k^T W_k\right) - \mu W_k\mathrm{UT}\left(W_k^T W_k - I_p\right)\right), \tag{5.49}$$

where μ > 0, and UT[·] sets all elements below the diagonal of its matrix argument to zero. The ODE associated with (5.49) is:

$$\frac{d}{dt}W(t) = 2AW - W\mathrm{UT}\left(W^T AW\right) - AW\mathrm{UT}\left(W^T W\right) - \mu W\mathrm{UT}\left(W^T W - I_p\right). \tag{5.50}$$

Theorem 5.9. *Let A5.1 and A5.3 hold. Then, all the equilibrium points of the ODE* (5.50) *are up to an arbitrary permutation of the eigenvectors of A weighted by* 0, +1 *or* –1, *i.e., any point W* = [$d_{(1)}\boldsymbol{\phi}_{(1)}$ $d_{(2)}\boldsymbol{\phi}_{(2)}$ … $d_{(p)}\boldsymbol{\phi}_{(p)}$]*, where* $d_{(i)}$ = 0, +1 *or* –1, *is an equilibrium point of the ODE* (5.50).

Proof. We need to find a $W\in\Re^{nXp}$ such that

$$2AW - W\mathrm{UT}\left(W^T AW\right) - AW\mathrm{UT}\left(W^T W\right) - \mu W\mathrm{UT}\left(W^T W - I_p\right) = 0 . \qquad (5.51)$$

The trivial solution is W=0. We next assume that $W\neq0$. We multiply (5.57) on the left by W^T, and define $G=W^TAW+(\mu/2)W^TW$, and $H=W^TW$. From (5.51), we obtain:

$$2G = H\ \mathrm{UT}(G) + G\ \mathrm{UT}(H). \qquad (5.52)$$

Since A is positive definite by Assumption A5.1, and since W is assumed to be nonzero, both G and H have positive diagonal elements. From (5.52), we conclude that $H=I_p$, and G is diagonal. Thus, $W^TW=I_p$, and W^TAW is diagonal. Then, the columns of W consist of the p orthonormal eigenvectors of A, i.e., $W=\Phi PD$, where $P\in\Re^{nXn}$ is an arbitrary permutation matrix, and $D=[D_1\,|\,0]^T\in\Re^{nXp}$ where $D_1\in\Re^{pXp}$ is a diagonal matrix with elements $d_i=\pm1$ for $i=1,\ldots,p$. Combining all results, we conclude that $W = \Phi PD$, where d_i = 0 or ±1. ■

Theorem 5.10. *Let A5.1 and A5.3 hold. Then, the points* $W^* = [\pm\phi_1\ \pm\phi_2 \ldots \pm\phi_p]$ *are the strict global minimum points of the objective function* (5.54) *and stable equilibrium points of the ODE* (5.52). *In addition, the points* $W = [d_{(1)}\phi_{(1)}\ d_{(2)}\phi_{(2)} \ldots d_{(p)}\phi_{(p)}]$, *where* $d_{(i)} = 0$ *or* $\phi_{(i)} \neq \phi_i$ *for* $i\in\{1,2,\ldots,p\}$ *are unstable equilibrium points of the ODE* (5.52).

Proof. From (5.38), the energy function $E(W)$ for the AL2 Deflation adaptive algorithm is:

$$E(W) = \sum_{i=1}^{p} E_i(\mathbf{w}_i)\text{, where}$$

$$E_i(\mathbf{w}_i) = -2\mathbf{w}_i^T A\mathbf{w}_i + \mathbf{w}_i^T A\mathbf{w}_i\mathbf{w}_i^T\mathbf{w}_i + 2\sum_{j=1}^{i-1}\mathbf{w}_j^T A\mathbf{w}_i\mathbf{w}_j^T\mathbf{w}_i + \mu\left(\sum_{j=1}^{i-1}\left(\mathbf{w}_j^T\mathbf{w}_i\right)^2 + \frac{1}{2}\left(\mathbf{w}_i^T\mathbf{w}_i - 1\right)^2\right).$$

From Theorem 5.9, $\mathbf{w}_i = d_{(i)}\phi_{(i)}$, $d_{(i)}$=0, ±1, is the i^{th} column of W for $i=1,\ldots,p$. The rest of the proof is same as Theorem 5.8 with ψ_i substituted by $\mathbf{w}_i$. We conclude that the columns of W consisting of the first p orthonormal eigenvectors of A as $W=[\pm\phi_1 \ldots \pm\phi_p]$ are stable minimum points of $E(W)$.

■

AL2 Weighted Algorithm

The objective function for AL2 Weighted PCA algorithm is:

$$J(\mathbf{w}_k^i;A_k) = -c_i\mathbf{w}_k^{i\,T}A_k\mathbf{w}_k^i + c_i\left(\mathbf{w}_k^{i\,T}A_k\mathbf{w}_k^i\right)\left(\mathbf{w}_k^{i\,T}\mathbf{w}_k^i - 1\right) + 2\sum_{j=1,\,j\neq i}^{p} c_j\mathbf{w}_k^{i\,T}A_k\mathbf{w}_k^j\mathbf{w}_k^{j\,T}\mathbf{w}_k^i +$$

$$\mu\left(\sum_{j=1,\,j\neq i}^{p} c_j\left(\mathbf{w}_k^{j\,T}\mathbf{w}_k^i\right)^2 + \frac{c_i}{2}\left(\mathbf{w}_k^{i\,T}\mathbf{w}_k^i - 1\right)^2\right), \tag{5.53}$$

where *i*=1,…,*p*, μ>0, $c_1 > c_2 > \ldots > c_p > 0$. The AL2 Weighted adaptive gradient descent algorithm is:

$$W_{k+1} = W_k + \eta_k\left(2A_kW_kC - W_kCW_k^TA_kW_k - A_kW_kCW_k^TW_k - \mu W_kC\left(W_k^TW_k - I_p\right)\right), \tag{5.54}$$

where $C=diag(c_1,\ldots,c_p)$. The ODE associated with (5.54) is:

$$\frac{d}{dt}W(t) = 2AWC - WCW^TAW - AWCW^TW - \mu WC\left(W^TW - I_p\right). \tag{5.55}$$

Theorem 5.11. *Let A5.1 and A5.3 hold. Then, all the equilibrium points of the ODE* (5.55) *are up to an arbitrary permutation of the eigenvectors of A weighted by* 0, +1 *or* –1, *i.e., any point W* $[d_{(1)}\phi_{(1)}\ d_{(2)}\phi_{(2)} \ldots d_{(p)}\phi_{(p)}]$, *where* $d_{(i)}$ = 0, +1 *or* –1, *is an equilibrium point of the ODE* (5.55).

Proof. In order to satisfy the first order conditions (see Section 2.10) for the existence of the equilibrium points of the joint objective functions $J(\mathbf{w}_i;A)$ (in (5.53)) for *i*=1,…,*p*, we need to find a $W \in \Re^{n\mathrm{X}p}$ such that

$$2AWC - WCW^TAW - AWCW^TW - \mu WC\left(W^TW - I_p\right) = 0\,. \tag{5.56}$$

The trivial solution is *W*=0. We next assume that *W*≠0. We multiply (5.62) on the left by W^T, and define $G=W^TAW+(\mu/2)W^TW$, and $H=W^TW$. From (5.56), we obtain:

$$2GC = HCG + GCH. \tag{5.57}$$

Since *H*, *G*, and *HCG*+*GCH* are symmetric matrices, we conclude that *GC*=*CG*. Since *C* is diagonal with distinct diagonal elements, *G* is diagonal. Let $GC=CG=D\in\Re^{p\mathrm{X}p}$ be a diagonal matrix. Then, from (5.57), we get:

$$2D=HD+DH. \tag{5.58}$$

Since *W* is assumed to be nonzero, the diagonal elements of $H=W^TW$ are positive. This implies that $H=I_p$. The rest of the proof is similar to Theorem 5.9 above. ■

Theorem 5.12. *Let A5.1 and A5.3 hold. Then, the points* W^* = [$\pm\phi_1$ $\pm\phi_2$ … $\pm\phi_p$] *are stable equilibrium points of the ODE* (5.55) *and strict global minimum points of the objective function in* (5.53). *In addition, the points* W = [$d_{(1)}\phi_{(1)}$ $d_{(2)}\phi_{(2)}$ … $d_{(p)}\phi_{(p)}$], *where* $d_{(i)}$ = 0 *or* $\phi_{(i)} \neq \phi_i$ *for* $i\in\{1,2,\ldots,p\}$ *are unstable equilibrium points of the ODE* (5.55).

Proof. From (5.53), the energy function $E(W)$ for the PF Weighted adaptive algorithm is:

$$E(W)=\sum_{i=1}^{p}E_i(\mathbf{w}_i)\text{, where}$$

$$E_i(\mathbf{w}_i)=-2c_i\mathbf{w}_i^T A\mathbf{w}_i+c_i\mathbf{w}_i^T A\mathbf{w}_i\mathbf{w}_i^T\mathbf{w}_i+2\sum_{j=1,\neq i}^{p}c_j\mathbf{w}_j^T A\mathbf{w}_i\mathbf{w}_j^T\mathbf{w}_i+\mu\left(\sum_{j=1,\neq i}^{p}c_j\left(\mathbf{w}_j^T\mathbf{w}_i\right)^2+\frac{c_i}{2}\left(\mathbf{w}_i^T\mathbf{w}_i-1\right)^2\right)$$

Here $\mathbf{w}_i=d_{(i)}\phi_{(i)}$, $d_{(i)}$=0 or ±1, is the i^{th} column of W for i=1,…,p. The rest of the proof is similar to Theorem 5.10 above. ■

6. Proofs for Chapter 6

No proofs.

7. Proofs for Chapter 7

7.2 Algorithms and Objective Functions

Proofs of Convergence and Assumptions

As discussed before, we use Stochastic Approximation theory [Ljung77,92; Benveniste *et al.* 90] for the convergence proofs. However, due to the repetitive nature of the proofs and since some of these algorithms have been proven by other practitioners, we state the results, and prove the convergence for the XU and AL1 algorithms. We require the following assumptions:

Assumption (A7.1). Each A_k and B_k is bounded with probability one (w.p.1), symmetric, real and nonnegative definite, and one of the following conditions hold: (1) $\lim_{k\to\infty}E[A_k]$

= *A* and $B_k \to B$ w.p.1, or (2) $A_k \to A$ w.p.1 and $\lim_{k\to\infty} E[B_k] = B$, or (3) $A_k \to A$ and $B_k \to B$ w.p.1, where *A* and *B* are positive definite.

Assumption (A7.2). Same as A2.2 in Section 2.4.1.

Assumption (A7.3). The $p \le n$ largest eigenvalues of *A* with respect to *B* are each of unit multiplicity.

In the following discussions, we denote $\Phi=[\phi_1 \; \dots \; \phi_n] \in \Re^{n \mathrm{X} n}$ as the orthonormal generalized eigenvector matrix of *A* with respect to *B*, and $\Lambda= diag(\lambda_1,\dots,\lambda_n)$ as the generalized eigenvalue matrix, such that $\lambda_1>\lambda_2>\dots>\lambda_p>\lambda_{p+1}\ge\dots\ge\lambda_n>0$. We use the subscript (*i*) to denote the *i*th permutation of the indices $\{1,2,\dots,n\}$.

7.4 XU GEVD Algorithms

Xu Homogeneous Algorithm

The objective function for XU Homogeneous adaptive GEVD algorithm is:

$$J(\mathbf{w}_k^i; A_k, B_k) = -2\mathbf{w}_k^{i\,T} A_k \mathbf{w}_k^i + \left(\mathbf{w}_k^{i\,T} A_k \mathbf{w}_k^i\right)\left(\mathbf{w}_k^{i\,T} B_k \mathbf{w}_k^i\right) + 2\sum_{j=1,\, j\neq i}^{p} \mathbf{w}_k^{i\,T} A_k \mathbf{w}_k^j \mathbf{w}_k^{j\,T} B_k \mathbf{w}_k^i \;, \qquad (7.14)$$

for *i*=1,…,*p* ($p \le n$). From the gradient of (7.14) with respect to $\mathbf{w}_k^i$, we obtain the XU Homogeneous adaptive gradient descent algorithm as:

$$\mathbf{w}_{k+1}^i = \mathbf{w}_k^i + \eta_k \left(2A_k \mathbf{w}_k^i - \sum_{j=1}^{p} A_k \mathbf{w}_k^j \mathbf{w}_k^{j\,T} B_k \mathbf{w}_k^i - \sum_{j=1}^{p} B_k \mathbf{w}_k^j \mathbf{w}_k^{j\,T} A_k \mathbf{w}_k^i \right) \qquad (7.15)$$

for *i*=1,…,*p*, whose matrix form is:

$$W_{k+1} = W_k + \eta_k \left(2A_k W_k - A_k W_k W_k^T B_k W_k - B_k W_k W_k^T A_k W_k\right). \qquad (7.16)$$

The ODE associated with (7.16) is:

$$\frac{d}{dt} W(t) = 2AW - AWW^T BW - BWW^T AW. \qquad (7.17)$$

We look for the stable equilibrium points of the ODE, and the convergence properties of its initial states close to the stable points.

Theorem 7.1. *For the ordinary differential equation* (7.17)*, let A7.1 and A7.3 hold. Then W*=Φ*PDU are equilibrium points of* (7.17)*, where D*=$[D_1\,|\,0]^T \in \Re^{n \mathrm{X} p}$ *with* $D_1 \in \Re^{p \mathrm{X} p}$ *is diagonal with elements*

d_i=+1, –1 or 0, $P\in\Re^{n\text{X}n}$ is an arbitrary permutation matrix, and $U\in\Re^{p\text{X}p}$ is an arbitrary rotation matrix, i.e., $U^TU=UU^T=I_p$.

Proof. We need to find a $W\in\Re^{n\text{X}p}$ such that

$$2AW - AWW^TBW - BWW^TAW = 0. \tag{7.18}$$

The trivial solution is W=0. We next assume that $W\neq 0$. Let $W=B^{-1/2}QDV$ be the singular value decomposition of W, where $Q\in\Re^{n\text{X}n}$, and $V\in\Re^{p\text{X}p}$ are orthonormal, and $D\in\Re^{n\text{X}p}$ is diagonal. Replacing QDV for W in (7.28) and defining $M=Q^TB^{-1/2}AB^{-1/2}Q$, we get from (7.28):

$$2MD - MDD^TD - DD^TMD = 0. \tag{7.19}$$

Let $D=[D_1\,|\,0]^T$ where D_1 is a pXp diagonal matrix with diagonal elements d_i for i=1,…,p. Let $M=\begin{bmatrix} M_1 & M_2 \\ M_2^T & M_3 \end{bmatrix}$ be a partition of M where $M_1\in\Re^{p\text{X}p}$. From (7.19) we get:

$$\left(2M_1 - M_1D_1^2 - D_1^2M_1\right)D_1 = 0. \tag{7.20}$$

We conclude:

$$D_1 = 0 \text{ or } 2M_1 - M_1D_1^2 - D_1^2M_1 = 0. \tag{7.21}$$

From the second equation in (7.21), we obtain:

$$m_{ii}(d_i^2 - 1) = 0 \text{ for } i=j\text{, and } m_{ij}(d_i^2 + d_j^2 - 2) = 0 \text{ for } i\neq j, \tag{7.22}$$

where $M_1=[m_{ij}]$ for i,j=1,…,p. We satisfy both equations with $d_i=\pm 1$. Thus, $W=B^{-1/2}QDV$, where $d_i=\pm 1$ for i=1,…,p. Clearly, $W^TBW=I_p$. Then, from (7.18), we obtain:

$$AW = BWW^TAW. \tag{7.23}$$

Let $W^TAW=U^TLU$ be the eigen-decomposition of W^TAW, where $U\in\Re^{p\text{X}p}$ is orthonormal, and $L\in\Re^{p\text{X}p}$ is diagonal. From (7.23), we get $AWU^T=BWU^TL$, where WU^T is orthonormal with respect to B. Then, the columns of WU^T consist of the p orthonormal eigenvectors of A with respect to B, i.e., $WU^T = \Phi PD$, where $P\in\Re^{n\text{X}n}$ is an arbitrary permutation matrix, and $D=[D_1\,|\,0]^T\in\Re^{n\text{X}p}$ where $D_1\in\Re^{p\text{X}p}$ is a diagonal matrix with elements $d_i=\pm 1$ for i=1,…,p. Combining all results, we conclude that $W = \Phi PDU$, where U is orthonormal, and d_i = 0 or ±1.

■

Theorem 7.2. *Let A7.1 and A7.3 hold. Then $W=\Phi DU$, where $D=[D_1\,|\,0]^T\in\Re^{n\text{X}p}$, D_1= diag(d_1,…,d_p)$\in\Re^{p\text{X}p}$, $d_i=\pm 1$ for i=1,…,p, and $U\in\Re^{p\text{X}p}$ is an arbitrary rotation matrix, are stable equilibrium points of the ODE* (7.17) *and strict global minimum points of the objective function* (7.14)*. In addition, $W=\Phi PDU$, where d_i=0 for $i\leq p$ or $P\neq I$, are unstable equilibrium points of the ODE* (7.27).

Proof. From (7.14), the energy function $E(W)$ for the XU Homogeneous adaptive algorithm is:

$$E(W) = -2tr\left(W^T AW\right) + tr\left(W^T AWW^T BW\right). \tag{7.24}$$

From Theorem 7.1, W=ΦPDU=ΨU, where Ψ=ΦPD. Then from (7.14):

$$E(W) = \sum_{i=1}^{p} E_i(\mathbf{w}_i) \text{, where}$$

$$E_i(\mathbf{w}_i) = -2\Psi_i^T A \;\; \Psi_i + \Psi_i^T A\Psi_i \Psi_i^T B\Psi_i + \sum_{j=1,\neq i}^{p} \Psi_j^T A \;\; \Psi_i \Psi_j^T B\Psi_i \,.$$

Here $\boldsymbol{\psi}_i$=$d_{(i)}\boldsymbol{\phi}_{(i)}$ is the i^{th} column of Ψ for i=1,…,p, and $d_{(i)}$=0 or ±1. We first prove that $d_{(i)}$=0 is an unstable equilibrium point of $E(W)$. We perturb $\boldsymbol{\psi}_i$ by $\delta\boldsymbol{\phi}_i$. Then

$$E\left(\Psi_i = \delta_i\right) - E\left(\Psi_i = 0\right) = -2\delta^2\lambda_i + \delta^4\lambda_i < 0 \text{ for } 0 < \delta < \sqrt{2}\,.$$

We next prove that $\boldsymbol{\psi}_i$=±$\boldsymbol{\phi}_r$, (r>p) is an unstable equilibrium point of $E(W)$. We perturb $\boldsymbol{\psi}_i$ by $\delta\boldsymbol{\phi}_p$, i.e., $\boldsymbol{\psi}_i = \pm(\boldsymbol{\phi}_r + \delta\boldsymbol{\phi}_p)/\sqrt{(1+\delta^2)}$. Then

$$E\left(\pm(\boldsymbol{\phi}_r + \delta\boldsymbol{\phi}_p)/\sqrt{(1+\delta^2)}\right) - E\left(\pm\boldsymbol{\phi}_r\right) = -\delta^2(\lambda_p - \lambda_r)/(1+\delta^2) < 0\,.$$

We next prove that $\boldsymbol{\psi}_i$=±$\boldsymbol{\phi}_r$, (r≤p) is a stable equilibrium point of $E(W)$. We perturb $\boldsymbol{\psi}_i$ by $\delta\boldsymbol{\phi}_s$, i.e., $\boldsymbol{\psi}_i = \pm(\boldsymbol{\phi}_r + \delta\boldsymbol{\phi}_s)/\sqrt{(1+\delta^2)}$ for s>p. Then

$$E\left(\pm(\boldsymbol{\phi}_r + \delta\boldsymbol{\phi}_s)/\sqrt{(1+\delta^2)}\right) - E\left(\pm\boldsymbol{\phi}_r\right) = \delta^2(\lambda_r - \lambda_s)/(1+\delta^2) > 0\,.$$

Thus, the columns of Ψ consisting of the first p orthonormal eigenvectors of A with respect to B are the stable minimum points of $E(W)$. ■

Xu Weighted Algorithm

The objective function for XU Weighted adaptive GEVD algorithm is:

$$J(\mathbf{w}_k^i; A_k, B_k) = -2c_i \mathbf{w}_k^{i\,T} A_k \mathbf{w}_k^i + c_i\left(\mathbf{w}_k^{i\,T} A_k \mathbf{w}_k^i\right)\left(\mathbf{w}_k^{i\,T} B_k \mathbf{w}_k^i\right) + 2\sum_{j=1,\, j\neq i}^{p} c_j \mathbf{w}_k^{i\,T} A_k \mathbf{w}_k^j \mathbf{w}_k^{j\,T} B_k \mathbf{w}_k^i \tag{7.20}$$

for i=1,…,p (p≤n), where c_1,…,c_p are small positive numbers satisfying (7.12). The adaptive algorithm is:

$$W_{k+1} = W_k + \eta_k\left(2A_k W_k C - B_k W_k C W_k^T A_k W_k - A_k W_k C W_k^T B_k W_k\right), \tag{7.21}$$

where C = $diag$(c_1,…,c_p). The ODE associated with (7.21) is:

$$\frac{d}{dt}W(t) = 2AWC - BWCW^T AW - AWCW^T BW\,. \tag{7.22}$$

Theorem 7.3. Let A7.1 and A7.3 hold. Then, all the equilibrium points of the ODE (7.29) are up to an arbitrary permutation of the eigenvectors of A weighted by 0, +1 or –1, i.e., any point $W = [d_{(1)}\phi_{(1)}\ d_{(2)}\phi_{(2)}\ \ldots\ d_{(p)}\phi_{(p)}]$, where $d_{(i)}$ = 0, +1 or –1, is an equilibrium point of the ODE (7.29).

Proof. In order to satisfy the first order conditions (see Section 2.10) for the existence of the equilibrium points of the joint objective functions $J(w_i;A,B)$ (in (7.27)) for i=1,…,p, we need to find a $W \in \Re^{n\mathrm{X}p}$ such that

$$2AWC - BWCW^T AW - AWCW^T BW = 0. \tag{7.30}$$

The trivial solution is W=0. We next assume that $W \neq 0$. We multiply (7.40) on the left by W^T, and define $G=W^TAW$, and $H=W^TBW$. From (7.30), we obtain:

$$2GC = HCG + GCH. \tag{7.31}$$

Since H, G, and C are symmetric matrices, we conclude that GC is symmetric. Since C is diagonal with distinct diagonal elements, G is also diagonal. Let $GC=CG=D \in \Re^{p\mathrm{X}p}$ be a diagonal matrix. Then, from (7.31), we get:

$$2D=HD+DH. \tag{7.32}$$

Since W is assumed to be nonzero, the diagonal elements of $H=W^TBW$ are positive. This implies that $H=I_p$. Thus, $W^TBW=I_p$, and W^TAW is diagonal. Then, the columns of W consist of the p orthonormal eigenvectors of A with respect to B, i.e., $W=\Phi PD$, where $P \in \Re^{n\mathrm{X}n}$ is an arbitrary permutation matrix, and $D=[D_1 | 0]^T \in \Re^{n\mathrm{X}p}$ where $D_1 \in \Re^{p\mathrm{X}p}$ is a diagonal matrix with elements $d_i = \pm 1$ for i=1,…,p. Combining all results, we conclude that $W = \Phi PD$, where d_i = 0 or ±1. ■

Theorem 7.4. Let A7.1 and A7.3 hold. Then, the points $W^* = [\pm\phi_1\ \pm\phi_2\ \ldots\ \pm\phi_p]$ are stable equilibrium points of the ODE (7.29) and strict global minimum points of the objective function in (7.27). In addition, the points W = $[d_{(1)}\phi_{(1)}\ d_{(2)}\phi_{(2)}\ \ldots\ d_{(p)}\phi_{(p)}]$, where $d_{(i)}$ = 0 or $\phi_{(i)} \neq \phi_i$ for $i \in \{1,2,\ldots,p\}$ are unstable equilibrium points of the ODE (7.29).

Proof. From (7.27), the energy function $E(W)$ for the XU Weighted adaptive algorithm is:

$$E(W) = \sum_{i=1}^{p} E_i(\mathbf{w}_i)\text{, where}$$

$$E_i(\mathbf{w}_i) = -2c_i \mathbf{w}_k^{i\,T} A_k \mathbf{w}_k^i + c_i\left(\mathbf{w}_k^{i\,T} A_k \mathbf{w}_k^i\right)\left(\mathbf{w}_k^{i\,T} B_k \mathbf{w}_k^i\right) + 2\sum_{j=1,\, j\neq i}^{p} c_j \mathbf{w}_k^{i\,T} A_k \mathbf{w}_k^j \mathbf{w}_k^{j\,T} B_k \mathbf{w}_k^i\ .$$

Here $w_i = d_{(i)}\phi_{(i)}$, $d_{(i)}$=0 or ±1, is the i^{th} column of W for i=1,…,p. The rest of the proof is similar to Theorem 7.2 above. ■

RQ Deflation Algorithm

The objective function for RQ Deflation GEVD algorithm is:

$$J(\mathbf{w}_k^i; A_k, B_k) = -\frac{\mathbf{w}_k^{i\,T} A_k \mathbf{w}_k^i}{\mathbf{w}_k^{i\,T} B_k \mathbf{w}_k^i} + \alpha\left(\mathbf{w}_k^{i\,T} B_k \mathbf{w}_k^i - 1\right) + 2\sum_{j=1}^{i} \beta_j \mathbf{w}_k^{j\,T} B_k \mathbf{w}_k^i \ , \tag{7.56}$$

for *i*=1,…,*p*, where $(\alpha, \beta_1, \beta_2, \ldots, \beta_p)$ are Lagrange multipliers. By solving $(\alpha, \beta_1, \beta_2, \ldots, \beta_k)$ and replacing them in the gradient of (7.56), we obtain the adaptive algorithm:

$$W_{k+1} = W_k + \eta_k \left(A_k W_k - B_k W_k \mathrm{UT}\left(W_k^T A_k W_k\right)\right)\mathrm{DIAG}(W_k^T B_k W_k)^{-1} . \tag{7.57}$$

The ODE associated with (7.57) is:

$$\frac{d}{dt} W(t) = \left(AW - BW\mathrm{UT}\left(W^T AW\right)\right)\left(W^T BW\right)^{-1} . \tag{7.58}$$

Theorem 7.5. Let A7.1 and A7.3 hold. Then, all the equilibrium points of the ODE (7.68) are up to an arbitrary permutation of the eigenvectors of *A* weighted by 0, +1 or –1, i.e., any point *W* = $[d_{(1)}\phi_{(1)}\ d_{(2)}\phi_{(2)}\ \ldots\ d_{(p)}\phi_{(p)}]$, where $d_{(i)}$ = 0, +1 or –1, is an equilibrium point of the ODE (7.68).

Proof. We need to find a $W \in \Re^{nXp}$ such that

$$\left(AW - BW\mathrm{UT}\left(W^T AW\right)\right)\left(W^T BW\right)^{-1} = 0 . \tag{7.69}$$

The trivial solution is *W*=0. We next assume that W≠0. We multiply (7.69) on the left by W^T, and define $G=W^TAW$, and $H=W^TBW$. From (7.79), we obtain:

$$G = H\ UT(G). \tag{7.70}$$

Since *G* is symmetric, *HUT*(*G*) is also symmetric. Since *W* is assumed to be nonzero, both *G* and *H* have positive diagonal elements. From (7.70), we conclude that $H=I_p$, and *G* is diagonal. Thus, $W^TBW=I_p$, and W^TAW is diagonal. The rest of the proof is similar to Theorem 7.3. We conclude that $W = \Phi PD$, where d_i = 0 or ±1.

■

Theorem 7.6. *Let A7.1 and A7.3 hold. Then, the points* $W^* = [\pm\phi_1\ \pm\phi_2\ \ldots\ \pm\phi_p]$ *are the strict global minimum points of the objective function* (7.66) *and stable equilibrium points of the ODE* (7.68). *In addition, the points* $W = [d_{(1)}\phi_{(1)}\ d_{(2)}\phi_{(2)}\ \ldots\ d_{(p)}\phi_{(p)}]$, *where* $d_{(i)} = 0$ *or* $\phi_{(i)} \neq \phi_i$ *for* $i \in \{1,2,\ldots,p\}$ *are unstable equilibrium points of the ODE* (7.68).

Proof. From (7.66), the energy function *E*(*W*) for the RQ Deflation adaptive algorithm is:

$$E(W) = \sum_{i=1}^{p} E_i(\mathbf{w}_i) \text{, where}$$

$$E_i(\mathbf{w}_i) = -\frac{\mathbf{w}_k^{i\,T} A_k \mathbf{w}_k^i}{\mathbf{w}_k^{i\,T} B_k \mathbf{w}_k^i} + \alpha\left(\mathbf{w}_k^{i\,T} B_k \mathbf{w}_k^i - 1\right) + 2\sum_{j=1}^{i} \beta_j \mathbf{w}_k^{j\,T} B_k \mathbf{w}_k^i \ .$$

From Theorem 7.5, $\mathbf{w}_i = d_{(i)}\phi_{(i)}$, $d_{(i)}=0, \pm 1$, is the i^{th} column of W for i=1,…,p. Rest of the proof is same as Theorem 7.4 with ψ_i substituted by $\mathbf{w}_i$. We conclude that the columns of W consisting of the first p orthonormal eigenvectors of A since $W=[\pm\phi_1 \ldots \pm\phi_p]$ are stable minimum points of $E(W)$. ■

8. Proofs for Chapter 8

No proofs.