Technical Report

Department of Computer Science
and Engineering
University of Minnesota
4-192 EECS Building
200 Union Street SE
Minneapolis, MN 55455-0159 USA

TR 10-017

Common Component Analysis for Multiple Covariance Matrices

Huahua Wang, Arindam Banerjee, and Daniel Boley

August 04, 2010

Common Component Analysis for Multiple Covariance Matrices

Huahua Wang
Dept of Computer Science & Engg
University of Minnesota, Twin Cities
huwang@cs.umn.edu

Arindam Banerjee
Dept of Computer Science & Engg
University of Minnesota, Twin Cities
banerjee@cs.umn.edu

Daniel Boley
Dept of Computer Science & Engg
University of Minnesota, Twin Cities
boley@cs.umn.edu

Abstract

We consider the problem of finding a suitable common low-dimensional subspace for accurately representing a given set of covariance matrices. When the set contains only one covariance matrix, the subspace is given by Principal Component Analysis (PCA). For multiple covariance matrices, we term the problem *Common Component Analysis* (CCA). While CCA can be posed as a tensor decomposition problem, standard approaches to tensor decomposition have two critical issues: (i) Tensor decomposition methods are iterative and rely on the initialization. A bad initialization may lead to poor local optima; (ii) For a given level of approximation error, one does not know how to choose a suitable low dimensionality. In this paper, we present a detailed analysis of CCA which yields an effective initialization and iterative algorithms for the problem. The proposed methodology has provable approximation guarantees w.r.t. the global optimum, and also allows one to choose the dimensionality for a given level of approximation error. We also establish conditions under which the methodology will obtain the global optimum. We illustrate the effectiveness of the proposed method through extensive experiments on synthetic data as well as two real stock market datasets, where major financial events can be visualized in low dimensions.

1 Introduction

In recent years, simultaneous analysis of multiple high-dimensional covariance matrices is becoming increasingly important in diverse application domains ranging from finance to climate and environmental sciences [30, 31, 32, 11, 34]. The traditional approach for finding accurate low dimensional approximation to high dimensional covariance matrices is Principal Component Analysis (PCA) [14, 4]. In particular, PCA finds an orthogonal projection of a single covariance matrix to a low-dimensional space while preserving as much of the "energy" or variance as possible. The problem can be solved by an eigenvalue decomposition (EVD) of the single covariance matrix under consideration.

Given multiple covariance matrices, we consider the problem of finding a suitable common low-dimensional subspace for accurately representing all the covariance matrices. We term the problem *Common Component Analysis* (CCA). PCA is not suitable for finding such a subspace for multiple covariance matrices, particularly if the covariance matrices span different subspaces. Examples include stock market data where financial shocks and volatility arise from different sources, and yield stock return covariance matrices in different subspaces. The low-dimensional covariance representation of the high-dimensional covariance matrices can take two possible forms: diagonal or full. Existing models where diagonal low rank matrices are considered, such as PARAFAC/ CANDECOMP [16, 17, 24, 22] and Common Principal Components (CPC)[13, 12], do not allow interactions among low-dimensional components, and essentially assume that underlying factors are uncorrelated. Moreover, multiple matrices can be simultaneously diagonalized if and only if they commute [19], which need not be true in general. Consequently, in this paper, we consider the case where the low dimensional covariance matrices could be full matrices. Such decompositions have been widely studied under different names, such as Tucker2 models [35, 16, 24, 25, 22], Tensor PCA [6], 2DSVD [9], GLRAM

[36], and in tensor decomposition [21, 22, 26, 33]. Variance-correlation [11] and Cholesky decomposition [3, 31] have also been used to simultaneously model multiple covariance matrices in low dimensions.

While CCA can be posed as a tensor decomposition problem, unlike PCA, standard approaches to tensor decomposition have two critical issues: (i) Tensor decomposition methods are iterative and rely on the initialization. A bad initialization may lead to poor local optima; (ii) For a given level of approximation error, one does not know how to choose a suitable low dimensionality. In this paper, we present a detailed analysis of CCA and present algorithms which address these two issues. We start by showing that our problem is equivalent to maximizing (not minimizing) a convex function over a compact but non-convex set. As a result, finding the global maximum in general is difficult. With an analysis using a simpler variant of CCA, we derive lower and upper bounds for the CCA objective for any orthonormal matrix. The bounds naturally lead to corresponding lower and upper bounds for the global maxima of CCA. We also give sufficiency conditions under which global maxima will be achieved. In [9], similar bounds were established for the local maxima of a related problem, but the closeness of the bounds w.r.t. the global maxima was not explicitly investigated. Using our bounds, we propose an initialization for iterative update methods which have a clear approximation guarantee w.r.t. the global maxima. Related favorable properties of suitable initialization has been observed in [36, 9], particularly for rank-1 approximation [27, 20]. Our analysis shows that instead of starting with a given low dimension, one can start with a approximation error bound, and choose a sufficient dimensionality appropriately for CCA which satisfies the given error bound. Note that such dimensionality selection is not possible for general tensor decomposition problems. We present two iterative update algorithms which start from the prescribed initialization, and monotonically improve the objective function till convergence. One algorithm is based on a standard update used in the tensor decomposition literature [24, 25, 6, 9, 36]. We also propose a novel algorithm based on an auxiliary function [28, 29]. The novel algorithm is substantially more efficient especially for low-dimensional projections, since the update only requires preforming the SVD of a $r \times n$ matrix instead of the EVD of a $n \times n$ matrix used in standard tensor decomposition methods, where n is the dimensionality of observed high-dimensional covariance and r is the dimensionality of latent low-dimensional covariance.

The remainder of this paper is organized as follows. We formulate the Common Component Analysis (CCA) problem in Section 2. In Section 3, we analyze the problem, establish lower and upper bounds for the global maxima, introduce the initialization and its optimality properties, establish sufficient conditions under which global maximum will be achieved, and also discuss the connections to related work. In Section 4, we present two algorithms for CCA given a suitable initialization, which can work with a given dimensionality or given approximation error bound. We report experimental results on synthetic data as well as two stock market datasets to illustrate the performance of the proposed ideas in Section 5, and conclude in Section 6.

Notation: Matrices are denoted by uppercase bold letters $(e.g., \mathbf{X})$. Vectors are denoted by bold lowercase letters $(e.g., \mathbf{x})$. The diagonal entries in a diagonal matrix are generally assumed to be in non-decreasing order. $\mathbb{I}_r$, where r is an integer, denotes an identity matrix of size r. If clear from context, r may be omitted (usually dimension n).

2 Problem Formulation

Assume a set of high dimensional covariance matrices $\mathbf{X}_t \in \mathbb{R}^{n\times n}, 1 \leq t \leq T$. The key hypothesis driving our analysis is that the high-dimensional covariance matrices are indeed a linearly transformed version of a set of low dimensional covariance matrices $\mathbf{Y}_t \in \mathbb{R}^{r\times r}, 1 \leq t \leq T$. While the linear transformation $\mathbf{U} \in \mathbb{R}^{n\times r}$ as well as the low dimensional covariance matrices $\mathbf{Y}_t, 1 \leq t \leq T$, are unknown, $\mathbf{X}_t$ is assumed to be well approximated by $\mathbf{U}\mathbf{Y}_t\mathbf{U}^T$. In particular,

$$\mathbf{X}_t = \mathbf{U}\mathbf{Y}_t\mathbf{U}^T + \mathbf{E}_t \tag{1}$$

where $\mathbf{E}_t$ is the residual matrix. Without loss of generality, $\mathbf{U}$ is assumed to be orthonormal, i.e., $\mathbf{U}^T\mathbf{U} = \mathbb{I}_r$. The goal is to find $\mathbf{U}$ and $\mathbf{Y}_t, 1 \leq t \leq T$ such that the sum of the Frobenius norms of all the residual matrices are minimized. The problem can be formally stated as follows:

$$\min_{\substack{\mathbf{U},\mathbf{Y}_t \\ \mathbf{U}^T\mathbf{U}=\mathbb{I}_r}} \sum_{t=1}^{T} \|\mathbf{X}_t - \mathbf{U}\mathbf{Y}_t\mathbf{U}^T\|_F^2 \,. \tag{2}$$

Since $\mathbf{U}$ determines a common subspace for all the covariance matrices, we call the above formulation *Common Component Analysis* (CCA).

We make a few observations before continuing with our analysis. If there is only one covariance matrix $\mathbf{X}_1$ under consideration, then the model reduces to standard PCA. For a given value of r, the objective $\|\mathbf{X}_1 - \mathbf{U}\mathbf{Y}_1\mathbf{U}^T\|^2$ is minimized when $\mathbf{U}$ consists of the r principal eigenvectors of $\mathbf{X}_1$ and $\mathbf{Y}_1$ is the diagonal matrix of the corresponding eigenvalues. For more than one matrices, the existing literature on tensor decompositions is relevant [24, 26, 22, 21, 6, 9, 36, 16, 17, 35]. If $\mathbf{X}_t$ is not a covariance matrix, i.e., $\mathbf{X}_t \in \mathbb{R}^{m\times n}$, it is modeled as $\mathbf{X}_t = \mathbf{U}\mathbf{Y}_t\mathbf{V}^T + \mathbf{E}_t$, where $\mathbf{U} \in \mathbb{R}^{m\times r}, \mathbf{Y}_t \in \mathbb{R}^{r\times s}, \mathbf{V} \in \mathbb{R}^{s\times n}$. Assuming $r = s$ and restricting $\mathbf{Y}_t$ be to be diagonal leads to PARAFAC/CANDECOMP models [24, 22]. When such restrictions are not imposed, one gets Tucker2 models [24, 22]. Iterative algorithms and data mining applications of such decompositions have been studied in the literature [24, 22, 23, 9, 36]. Unlike most existing settings, in our model each $\mathbf{X}_t$ is a positive semi-definite matrix, and $\mathbf{Y}_t$ is also positive semi-definite. We discuss technical relationships of our analysis to existing models in Section 3.5.

We start the analysis with the following two results:

Lemma 1 *The optimum $\mathbf{Y}_t$ in (2) satisfies $\mathbf{Y}_t = \mathbf{U}^T\mathbf{X}_t\mathbf{U}$. Further, the optimal $\mathbf{U}$ in (2) is the solution to the following problem:*

$$\max_{\mathbf{U}^T\mathbf{U}=\mathbb{I}_r} f(\mathbf{U}) = \max_{\mathbf{U}^T\mathbf{U}} \operatorname{Tr}(\mathbf{U}^T M(\mathbf{U})\mathbf{U}) \,, \tag{3}$$

where

$$M(\mathbf{U}) = \sum_{t=1}^{T} \mathbf{X}_t\mathbf{U}\mathbf{U}^T\mathbf{X}_t \,. \tag{4}$$

Proof: Since $\mathbf{U}^T\mathbf{U} = \mathbb{I}_r$, taking the derivative of objective function in (2) with respect to $\mathbf{Y}_t$ and setting it to zero, we obtain

$$\mathbf{U}^T\mathbf{X}_t\mathbf{U} - \mathbf{Y}_t = 0 \,,$$

proving the first part of the result. Replacing this expression for $\mathbf{Y}_t$ in (2), we obtain

$$\begin{aligned}
&\min_{\mathbf{U}^T\mathbf{U}=\mathbb{I}_r} \sum_{t=1}^{T} \|\mathbf{X}_t - \mathbf{U}\mathbf{Y}_t\mathbf{U}^T\|_F^2 \\
&= \min_{\mathbf{U}^T\mathbf{U}=\mathbb{I}_r} \sum_{t=1}^{T} \operatorname{Tr}((\mathbf{X}_t - \mathbf{U}\mathbf{Y}_t\mathbf{U}^T)^T(\mathbf{X}_t - \mathbf{U}\mathbf{Y}_t\mathbf{U}^T)) \\
&= \min_{\mathbf{U}^T\mathbf{U}=\mathbb{I}_r} \sum_{t=1}^{T} \operatorname{Tr}(\mathbf{X}_t^2 - 2\mathbf{X}_t\mathbf{U}\mathbf{Y}_t\mathbf{U}^T + \mathbf{U}\mathbf{Y}_t\mathbf{U}^T\mathbf{U}\mathbf{Y}_t\mathbf{U}^T) \\
&\overset{(a)}{=} \min_{\mathbf{U}^T\mathbf{U}=\mathbb{I}_r} \sum_{t=1}^{T} \operatorname{Tr}(\mathbf{X}_t^2 - 2\mathbf{X}_t\mathbf{U}\mathbf{U}^T\mathbf{X}_t\mathbf{U}\mathbf{U}^T + \mathbf{U}\mathbf{U}^T\mathbf{X}_t\mathbf{U}\mathbf{U}^T\mathbf{X}_t\mathbf{U}\mathbf{U}^T) \\
&\overset{(b)}{=} \min_{\mathbf{U}^T\mathbf{U}=\mathbb{I}_r} \sum_{t=1}^{T} \operatorname{Tr}(\mathbf{X}_t^2 - \mathbf{U}^T\mathbf{X}_t\mathbf{U}\mathbf{U}^T\mathbf{X}_t\mathbf{U}) \\
&= \min_{\mathbf{U}^T\mathbf{U}=\mathbb{I}_r} \operatorname{Tr}\left(\sum_{t=1}^{T} \mathbf{X}_t^2\right) - \operatorname{Tr}\left(\sum_{t=1}^{T} \mathbf{U}^T\mathbf{X}_t\mathbf{U}\mathbf{U}^T\mathbf{X}_t\mathbf{U}\right) \,,
\end{aligned}$$

where (a) holds because $\mathbf{Y}_t = \mathbf{U}^T\mathbf{X}_t\mathbf{V}$, and (b) holds since $\operatorname{Tr}(AB) = \operatorname{Tr}(BA)$ and $\mathbf{U}^T\mathbf{U} = \mathbb{I}_r$. Since $\operatorname{Tr}(\sum_t^N \mathbf{X}_t^2)$ is a constant, problem (2) is equivalent to the following maximization problem

$$\max_{\mathbf{U}^T\mathbf{U}=\mathbb{I}_r} \operatorname{Tr}(\mathbf{U}^T M(\mathbf{U})\mathbf{U})$$

where

$$M(\mathbf{U}) = \sum_{t} \mathbf{X}_t\mathbf{U}\mathbf{U}^T\mathbf{X}_t \,.$$

That completes the proof. ■

Next we show that $f(\mathbf{U})$ in (3) is convex. For this we need a lemma

Lemma 2 $0 \leq \mathrm{Tr}(\mathbf{A} \cdot \mathbf{B}) \leq \mathrm{Tr}(\mathbf{A}) \cdot \mathrm{Tr}(\mathbf{B})$ *for any two symmetric positive semi-definite matrices* $\mathbf{A}, \mathbf{B}$*,*

Proof: Factor $\mathbf{A} = \mathbf{K}\mathbf{K}^T$, $\mathbf{B} = \mathbf{L}\mathbf{L}^T$. Then using the identity $\mathrm{Tr}(\mathbf{X}\mathbf{Y}) = \mathrm{Tr}(\mathbf{Y}\mathbf{X})$ for any $\mathbf{X}, \mathbf{Y}$:

$$\begin{aligned} \mathrm{Tr}(\mathbf{A} \cdot \mathbf{B}) &= \mathrm{Tr}(\mathbf{K}\mathbf{K}^T\mathbf{L}\mathbf{L}^T) = \mathrm{Tr}(\mathbf{L}^T\mathbf{K}\mathbf{K}^T\mathbf{L}) \\ &= \|\mathbf{K}^T\mathbf{L}\|_F^2. \end{aligned}$$

Hence we have

$$0 \leq \|\mathbf{K}^T\mathbf{L}\|_F^2 \leq \|\mathbf{K}^T\|_F^2 \cdot \|\mathbf{L}\|_F^2 = \mathrm{Tr}(\mathbf{A}) \cdot \mathrm{Tr}(\mathbf{B}).$$

■

Lemma 3 *For* $\mathbf{U} \in \mathbb{R}^{n \times r}$ *(not necessarily orthonormal),* $f(\mathbf{U})$ *is a convex function.*

Proof: It suffices to show $f_{\mathbf{X}}(\mathbf{U}) = \mathrm{Tr}[\mathbf{F}_{\mathbf{X}}(\mathbf{U})]$ is a convex function of $\mathbf{U}$ for any single symmetric positive semidefinite matrix $\mathbf{X}$, where $\mathbf{F}_{\mathbf{X}}(\mathbf{U}) = \mathbf{U}^T\mathbf{X}\mathbf{U}\mathbf{U}^T\mathbf{X}\mathbf{U}$.

We show convexity by showing that the second derivative in any particular direction is non-negative. Pick an arbitrary direction $\mathbf{V}$ and compute

$$\mathbf{F}_{\mathbf{X}}(\mathbf{U} + s\mathbf{V}) = \mathbf{F}_{\mathbf{X}}(\mathbf{U}) + s\mathbf{G} + s^2\mathbf{H} + h.o.t., \tag{5}$$

where $h.o.t.$ denotes the high order terms, $\mathbf{G}, \mathbf{H}$ are expressions in $\mathbf{X}, \mathbf{U}, \mathbf{V}$ to be computed. We want to show $\mathrm{Tr}(\mathbf{H}) \geq 0$. Expanding (5) yields the following expression for $\mathbf{H}$:

$$\begin{aligned} \mathbf{H} = &\mathbf{V}^T\mathbf{X}\mathbf{V}\mathbf{U}^T\mathbf{X}\mathbf{U} + \mathbf{U}^T\mathbf{X}\mathbf{U}\mathbf{V}^T\mathbf{X}\mathbf{V} && \text{(a)} \\ &+ \mathbf{V}^T\mathbf{X}\mathbf{U}\mathbf{U}^T\mathbf{X}\mathbf{V} + \mathbf{U}^T\mathbf{X}\mathbf{V}\mathbf{V}^T\mathbf{X}\mathbf{U} && \text{(b)} \\ &+ \mathbf{V}^T\mathbf{X}\mathbf{U}\mathbf{V}^T\mathbf{X}\mathbf{U} + \mathbf{U}^T\mathbf{X}\mathbf{V}\mathbf{U}^T\mathbf{X}\mathbf{V} && \text{(c)} \\ = &\mathbf{V}^T\mathbf{X}\mathbf{V}\mathbf{U}^T\mathbf{X}\mathbf{U} + \mathbf{U}^T\mathbf{X}\mathbf{U}\mathbf{V}^T\mathbf{X}\mathbf{V} && \text{(a)} \\ &+ (\mathbf{V}^T\mathbf{X}\mathbf{U} + \mathbf{U}^T\mathbf{X}\mathbf{V})^2 && \text{(d)=(b)+(c)} \end{aligned}$$

The trace of (a) is non-negative from Lemma 2. The expression (d) is the square of a symmetric matrix, and hence its trace is also non-negative. ■

Unfortunately, the fact that $f(\mathbf{U})$ is convex does not help us in any way. Note that from (3), the problem is one of *maximizing* $f(\mathbf{U})$ instead of minimizing it. Further, the constraint set $\mathbf{U}^T\mathbf{U} = \mathbb{I}_r$ is not convex. As a result the problem in (3) is not convex. In fact, the problem is one of maximizing a convex function over a non-convex feasible set. As a result, there may be several local maxima. In particular, a standard approach of starting from an initial guess, as is commonly employed in alternating least squares, will likely get stuck in local minima. Furthermore, it is difficult to characterize the proximity of such solutions in terms of the function value achieved with respect to the global optimum. In the next two sections, we develop a novel way to initialize $\mathbf{U}$ along with algorithms for iterative updates with guarantees relative to the global optimum.

3 Analysis of Common Component Analysis

In this section, we analyze CCA in terms of a simpler model we call Common Component Analysis 1 (CCA1). We show that CCA1 is a PCA-style problem, and can be solved using eigen-value decomposition. More importantly, the solution to CCA1 leads to lower and upper bounds on the global maximum of CCA, and suggests a good initialization for any iterative algorithm for solving CCA. Instead of a given dimensionality, if one wants to solve CCA for a given approximation error, our analysis shows how one can choose a suitable dimensionality sufficient to satisfy the approximation error bound.

3.1 A Simpler Model: CCA1

Instead of the original problem in (2), we consider a simpler decomposition given by

$$\mathbf{X}_t = \mathbf{U}\mathbf{Y}_t + \mathbf{E}_t \tag{6}$$

where $\mathbf{U} \in \mathbb{R}^{n\times r}$ and $\mathbf{Y}_t \in \mathbb{R}^{r\times n}$. Assuming the residual norms to be small, the problem of finding $\mathbf{U}, Y_t$ can be posed as follows:

$$\min_{\substack{\mathbf{U},\mathbf{Y}_t \\ \mathbf{U}^T\mathbf{U}=\mathbb{I}_r}} \sum_{t=1}^{T} \|\mathbf{X}_t - \mathbf{U}\mathbf{Y}_t\|_F^2 \,. \tag{7}$$

We call the above problem CCA1 since it only considers one-sided projection compared to two-sided projections in CCA. Similar to CCA, the simplified problem CCA1 allows an alternative characterization as follows:

Lemma 4 *The optimal $\mathbf{Y}_t$ in (7) satisfies $\mathbf{Y}_t = \mathbf{U}^T\mathbf{X}_t$. Further, the optimal $\mathbf{U}$ in (6) is the solution to the following problem:*

$$\max_{\mathbf{U}^T\mathbf{U}=\mathbb{I}_r} f_1(\mathbf{U}) = \max_{\mathbf{U}^T\mathbf{U}=\mathbb{I}_r} \operatorname{Tr}(\mathbf{U}^T M(\mathbb{I}_n)\mathbf{U}) \,, \tag{8}$$

where

$$M(\mathbb{I}_n) = \sum_{t=1}^{T} \mathbf{X}_t^2 \,. \tag{9}$$

Proof: Since $\mathbf{U}^T\mathbf{U} = \mathbb{I}_r$, taking derivative of (6) w.r.t. $\mathbf{Y}_t$ and setting to zero yields $\mathbf{U}^T\mathbf{X}_t - \mathbf{Y}_t = 0$, proving the first part. Replacing this expression for $\mathbf{Y}_t$ in (6), we obtain

$$\begin{aligned}
&\min_{\mathbf{U}^T\mathbf{U}=\mathbb{I}_r} \sum_{t=1}^{T} \|\mathbf{X}_t - \mathbf{U}\mathbf{Y}_t\|_F^2 \\
&= \min_{\mathbf{U}^T\mathbf{U}=\mathbb{I}_r} \sum_{t=1}^{T} \|\mathbf{X}_t - \mathbf{U}\mathbf{U}^T\mathbf{X}_t\|_F^2 \\
&= \min_{\mathbf{U}^T\mathbf{U}=\mathbb{I}_r} \sum_{t=1}^{T} \operatorname{Tr}((\mathbb{I}_n - \mathbf{U}\mathbf{U}^T)\mathbf{X}_t\mathbf{X}_t(\mathbb{I}_n - \mathbf{U}\mathbf{U}^T)) \\
&= \min_{\mathbf{U}^T\mathbf{U}=\mathbb{I}_r} \sum_{t=1}^{T} \operatorname{Tr}(\mathbf{X}_t^2 - \mathbf{U}\mathbf{U}^T\mathbf{X}_t^2 + \mathbf{U}\mathbf{U}^T\mathbf{X}_t^2\mathbf{U}\mathbf{U}^T) \\
&\overset{(a)}{=} \min_{\mathbf{U}^T\mathbf{U}=\mathbb{I}_r} \sum_{t}^{T} \operatorname{Tr}(\mathbf{X}_t^2 - \mathbf{U}^T\mathbf{X}_t^2\mathbf{U}) \\
&= \min_{\mathbf{U}^T\mathbf{U}=\mathbb{I}_r} \operatorname{Tr}\left(\sum_{t=1}^{T} X_t^2\right) - \operatorname{Tr}\left(\sum_{t=1}^{T} \mathbf{U}^T\mathbf{X}_t^2\mathbf{U}\right)
\end{aligned}$$

where (a) holds since $\operatorname{Tr}(AB) = \operatorname{Tr}(BA)$ and $\mathbf{U}^T\mathbf{U} = \mathbb{I}_r$. Since $\operatorname{Tr}(\sum_{t=1}^{T} \mathbf{X}_t^2)$ is a constant, problem (6) is equivalent to the following maximization problem

$$\max_{\mathbf{U}^T\mathbf{U}=\mathbb{I}_r} \operatorname{Tr}(\mathbf{U}^T M(\mathbb{I}_n)\mathbf{U})$$

where $M(\mathbb{I}_n) = \sum_{t=1}^{T} \mathbf{X}_t^2$. That completes the proof. ∎

First note that CCA1 as in (8) is a PCA problem on $M(\mathbb{I}_n)$, which can be solved using eigen-value decomposition. Table 1 shows a relative comparison between CCA and CCA1.

Table 1: CCA and CCA1

CCA	CCA1
$\mathbf{X}_t = \mathbf{U}\mathbf{Y}_t\mathbf{U} + \mathbf{E}_t$	$\mathbf{X}_t = \mathbf{U}\mathbf{Y}_t + \mathbf{E}_t$
$M(\mathbf{U}) = \sum_t \mathbf{X}_t\mathbf{U}\mathbf{U}^T\mathbf{X}_t$	$M(\mathbb{I}_\mathbf{n}) = \sum_t \mathbf{X}_t^2$
$f(\mathbf{U}) = \mathrm{Tr}(\mathbf{U}^T M(\mathbf{U})\mathbf{U})$	$f_1(\mathbf{U}) = \mathrm{Tr}(\mathbf{U}^T M(\mathbb{I}_\mathbf{n})\mathbf{U})$

3.2 Lower and Upper Bounds

The solution of CCA1 helps significantly in characterizing the solution to CCA. We focus on developing lower and upper bounds to optimum value of CCA based on the solution of CCA1. Since CCA1 is essentially the PCA problem over $M(\mathbb{I}_n) = \sum_t \mathbf{X}_t^2$, if $\mathbf{U}_0$ denotes the top r eigenvectors of $M(\mathbb{I}_n) = \sum_{t=1}^T \mathbf{X}_t^2$, then $\mathbf{U}_0$ is the solution to (8). Let $f_1^{\max} = f_1(\mathbf{U}_0)$ be the maximum value of $f_1(\mathbf{U})$. Further, let $M_T = \mathrm{Tr}(M(\mathbb{I}_n)) = \mathrm{Tr}\left(\sum_t \mathbf{X}_t^2\right)$. With this notation, we have the following result:

Theorem 1 *Let $M_T = \mathrm{Tr}(\sum_t \mathbf{X}_t^2)$. Then, with $f_1(\mathbf{U})$ and $f(\mathbf{U})$ denoting the objective functions for CCA1 and CCA respectively as in (8) and (3), for any $\mathbf{U}$ with $\mathbf{U}^T\mathbf{U} = \mathbb{I}_r$, we have*

$$\frac{f_1^2(\mathbf{U})}{M_T} \leq f(\mathbf{U}) \leq f_1(\mathbf{U}) \,. \tag{10}$$

Proof: By definition,

$$\begin{aligned} f(\mathbf{U}) &= \mathrm{Tr}(\mathbf{U}^T M(\mathbf{U})\mathbf{U}) \leq \mathrm{Tr}(M(\mathbf{U})) \\ &= \sum_{t=1}^T \mathrm{Tr}(\mathbf{X}_t\mathbf{U}\mathbf{U}^T\mathbf{X}_t) = \sum_{t=1}^T \mathrm{Tr}(\mathbf{U}^T\mathbf{X}_t^2\mathbf{U}) = f_1(\mathbf{U}) \,. \end{aligned}$$

Now, we prove $f(\mathbf{U}) \geq \frac{f_1^2(\mathbf{U})}{M_T}$. Since $\mathbf{X}_t$ is symmetric positive semidefinite, it can be written as $\mathbf{X}_t = \mathbf{X}_t^{\frac{1}{2}}\mathbf{X}_t^{\frac{1}{2}}$. We define the following matrices:

$$\mathbf{A} = \left[\mathbf{X}_1^{\frac{1}{2}}\mathbf{U}\mathbf{U}^T\mathbf{X}_1^{\frac{1}{2}}, \cdots, \mathbf{X}_T^{\frac{1}{2}}\mathbf{U}\mathbf{U}^T\mathbf{X}_T^{\frac{1}{2}}\right] \qquad \mathbf{B} = [\mathbf{X}_1, \cdots, \mathbf{X}_T] \,.$$

The trace of their product is given by

$$\mathrm{Tr}(\mathbf{A}\mathbf{B}^T) = \sum_{t=1}^T \mathrm{Tr}(\mathbf{X}_t^{\frac{1}{2}}\mathbf{U}\mathbf{U}^T\mathbf{X}_t^{\frac{1}{2}}\mathbf{X}_t) = \sum_{t=1}^T \mathrm{Tr}(\mathbf{U}^T\mathbf{X}_t^2\mathbf{U}) = f_1(\mathbf{U}) \,.$$

Now, $f(\mathbf{U})$ is rewritten as

$$\begin{aligned} f(\mathbf{U}) &= \sum_{t=1}^T \mathrm{Tr}(\mathbf{U}^T\mathbf{X}_t\mathbf{U}\mathbf{U}^T\mathbf{X}_t\mathbf{U}) \\ &= \sum_{t=1}^T \mathrm{Tr}(\mathbf{U}^T\mathbf{X}_t^{\frac{1}{2}}\mathbf{X}_t^{\frac{1}{2}}\mathbf{U}\mathbf{U}^T\mathbf{X}_t^{\frac{1}{2}}\mathbf{X}_t^{\frac{1}{2}}\mathbf{U}) \\ &= \sum_{t=1}^T \mathrm{Tr}(\mathbf{X}_t^{\frac{1}{2}}\mathbf{U}\mathbf{U}^T\mathbf{X}_t^{\frac{1}{2}}\mathbf{X}_t^{\frac{1}{2}}\mathbf{U}\mathbf{U}^T\mathbf{X}_t^{\frac{1}{2}}) \\ &= \mathrm{Tr}(\mathbf{A}\mathbf{A}^T) \,, \end{aligned}$$

and M_T is

$$M_T = \mathrm{Tr}(\sum_{t=1}^T \mathbf{X}_t^2) = \mathrm{Tr}(\mathbf{B}\mathbf{B}^T) \,.$$

From the Cauchy-Schwarz inequality, we have

$$f(\mathbf{U})M_T = \mathrm{Tr}(\mathbf{A}\mathbf{A}^T)\,\mathrm{Tr}(\mathbf{B}\mathbf{B}^T) \geq \left[\mathrm{Tr}(\mathbf{A}\mathbf{B}^T)\right]^2 = f_1^2(\mathbf{U})\ .$$

Dividing both sides by M_T completes the proof. ∎

Definition 1 Let p_1 denote the fraction of 'energy' in $\sum_t \mathbf{X}_t^2$ captured by the rank-r PCA solution $\mathbf{U}_0$. In particular,

$$p_1 = \frac{{f_1}^{\max}}{M_T} = \frac{\mathrm{Tr}\left(\mathbf{U}_0^T\left(\sum_t \mathbf{X}_t^2\right)\mathbf{U}_0\right)}{\mathrm{Tr}\left(\sum_t \mathbf{X}_t^2\right)}\ , \tag{11}$$

so that $0 \leq p_1 \leq 1$.

Using this definition and Theorem 1, we have the following result which bounds the value of the global maximum of CCA.

Corollary 1 *Let ${f_1}^{\max}$ and $f^{\max}$ be the global maximum of CCA1 and CCA respectively over $\mathbf{U}^T\mathbf{U} = \mathbb{I}_r$, and p_1 is as defined in $Definition\ 1$. Then, we have*

$$p_1 {f_1}^{\max} \leq f^{\max} \leq {f_1}^{\max} \tag{12}$$

Proof: Let $\mathbf{U}_0$ be the solution of CCA1, so that ${f_1}^{\max} = f_1(\mathbf{U}_0)$ and $p_1 = {f_1}^{\max}/M_T$. According to Theorem 1, we have

$$f(\mathbf{U}_0) \geq \frac{f_1^2(\mathbf{U}_0)}{M_T} = \frac{f_1^{\max}}{M_T} f_1^{\max} = p_1 {f_1}_{\max}\ .$$

Hence, for the global maximum of CCA, we have

$$f^{\max} \geq f(\mathbf{U}_0) \geq p_1 {f_1}_{\max}$$

Further, since $f_1(\mathbf{U})$ is an upper bound of $f(\mathbf{U})$, we have $f^{\max} \leq {f_1}^{\max}$. That completes the proof. ∎

Recall that the solution to CCA1 is $\mathbf{U}_0$, the top-r eigenvectors of $\sum_t \mathbf{X}_t^2$. Thus, it is easy to compute $f_1^{\max} = f_1(\mathbf{U}_0)$ and $p_1 = f_1^{\max}/M_T$. From Theorem 1, it follows that $p_1 f_1^{\max} \leq f(\mathbf{U}_0) \leq f_1^{\max}$. Now if we do iterative updates for $f(\mathbf{U})$ which start with initialization $\mathbf{U}_0$ and converges to $\mathbf{U}_0^*$ (see Section 4), we have

$$p_1 f_1^{\max} \leq f(\mathbf{U}_0) \leq f(\mathbf{U}_0^*) \leq f^{\max} \leq f_1^{\max}\ . \tag{13}$$

From (13), we note that if p_1 is close to 1, then $f(\mathbf{U}_0^*)$ will be close to the global maximum $f^{\max}$. The relative error of $f(\mathbf{U}_0^*)$ w.r.t. the global maximum is

$$\frac{f^{\max} - f(\mathbf{U}_0^*)}{f^{\max}} \leq \frac{{f_1}^{\max} - f(\mathbf{U}_0^*)}{{f_1}^{\max}} \tag{14}$$

Even before $f(\mathbf{U}_0^*)$ is found, there still exists an upper bound formalized in the result below:

Corollary 2 *Let $\mathbf{U}_0$ be the r principal eigenvectors of $M(\mathbb{I}_n) = \sum_t \mathbf{X}_t^2$, and $f(\mathbf{U}_0^*)$ be the solution to CCA with the initialization $\mathbf{U}_0$. Then, the relative error of $f(\mathbf{U}_0^*)$ with respect to $f^{\max}$ satisfies*

$$\frac{f^{\max} - f(\mathbf{U}_0^*)}{f^{\max}} \leq 1 - p_1 \tag{15}$$

Proof: Consider the inequality $f(\mathbf{U}_0^*) \geq p_1 {f_1}_{\max}$. Dividing both sides by $f^{\max}$ we get

$$\frac{f(\mathbf{U}_0^*)}{f^{\max}} \geq p_1 \frac{{f_1}^{\max}}{f^{\max}} \overset{(a)}{\geq} p_1\ ,$$

where (a) follows since ${f_1}^{\max} \geq f^{\max}$. Consequently

$$\left| \frac{f^{\max} - f(\mathbf{U}_0^*)}{f^{\max}} \right| = 1 - \frac{f(\mathbf{U}_0^*)}{f^{\max}} \leq 1 - p_1 \ .$$

■

Note that initialization itself satisfies the above bound, so that $f(\mathbf{U}_0) \geq p_1 f^{\max}$. In other words, $\mathbf{U}_0$ forms a good initialization assuming p_1 is large. In particular, if $p_1 = 1$, then $\mathbf{U}_0$ achieves the global maximum for $f(\mathbf{U})$. Since $\mathbf{U}_0$ gives a good initialization with guarantees, our algorithm will start with $\mathbf{U}_0$ and do iterative updates to hopefully reach an even better solution. In particular, if p_1 is large and $\mathbf{U}_0$ is in the basin of attraction of the global maxima, the iterative updates will be able to reach the global maxima.

3.3 Approximate Relative Error and Rank

In certain applications, one may have to pick a suitable rank r to preserve certain fraction of the observed covariance structure. The goal is to keep the rank r minimum while explaining a given fraction of the observed covariance, or, equivalently, having the error in approximating the observed covariance go below a given threshold. In PCA, since its solution based on EVD has a nested structure, there is a simple way to obtain a suitable rank r. In particular, one can keep incrementally adding rank till the error goes below the desired threshold. The rank r solution includes the rank $(r-1)$ solution and an additional dimension. Further, obtaining the best rank-r solution from the best rank-$(r-1)$ solution is computationally simple. However, such nested approximation structure is not present in CCA and more generally in case of tensor decompositions. Thus, the best rank $(r-1)$ solution to CCA does not provide any help in computing the best rank r solution. Thus, if the rank $(r-1)$ solution does not satisfy a given threshold in approximation error, the computation has to be entirely redone to check if the rank r solution is sufficient to meet the given approximation error. In this section, we show that such elaborate calculations can be avoided by using the bounds relative to the CCA1 problem.

We start with defining *Approximate Relative Error* (ARE) as a measure of how good the approximation obtained by CCA is. For any $\mathbf{U}$, we have

$$ARE(\mathbf{U}) = \frac{\sum_{t=1}^{T} \|\mathbf{X}_t - \mathbf{U}\mathbf{Y}_t\mathbf{U}^T\|_F^2}{\sum_{t=1}^{T} \|\mathbf{X}_t\|_F^2} \ . \tag{16}$$

We define the cumulative percentage of energy captured by the solution to CCA as follows:

Definition 2 Let $M_T = \mathrm{Tr}(M(\mathbb{I}_n))$, and let $f(\mathbf{U}_0^*)$ be the maximum of CCA obtained by an iterative algorithm with initialization $\mathbf{U}_0$ (see Section 4). The cumulative percent of energy p captured by $\mathbf{U}_0^*$ is defined as

$$p = \frac{f(\mathbf{U}_0^*)}{M_T} \ , \tag{17}$$

so that $0 \leq p \leq 1$.

For our problem, p defines how much energy over all the covariances is preserved by their corresponding latent covariances. Dividing by M_T on both sides of inequality (13) and plugging in $p_1 = {f_1}^{\max}/M_T$, the lower and upper bounds of p are

$$p_1^2 \leq p \leq p_1 \tag{18}$$

Recall that p_1 is defined in the PCA setting. In CCA1, given a p_1, the corresponding rank r is easy to obtain. Using the bounds for p, one can also develop a simple way of obtaining a suitable rank-r for CCA. To do this, we first establish a relationship between p and approximate relative error $ARE(\mathbf{U}_0^*)$.

Proposition 1 *Let $\mathbf{U}_0^*$ be the solution of CCA. Then $ARE(\mathbf{U}_0^*) = 1 - p$.*

Proof: From the proof of Lemma 1, we have

$$\sum_{t=1}^{T} \|\mathbf{X}_t - \mathbf{U}\mathbf{Y}_t\mathbf{U}^T\|_F^2 = \mathrm{Tr}\left(\sum_{t=1}^{T} \mathbf{X}_t^2\right) - \mathrm{Tr}\left(\sum_{t=1}^{T} \mathbf{U}^T\mathbf{X}_t\mathbf{U}\mathbf{U}^T\mathbf{X}_t\mathbf{U}\right)$$
$$= M_T - f(\mathbf{U}) \,.$$

Let $\mathbf{U}_0^*$ be the solution of CCA. Then

$$ARE(\mathbf{U}_0^*) = \frac{M_T - f(\mathbf{U}_0^*)}{M_T} = 1 - p \,.$$

■

Plugging $ARE(\mathbf{U}_0^*)$ into inequality (18), it is easy to derive the following lower and upper bounds for $ARE(\mathbf{U}_0^*)$:

$$1 - p_1 \leq ARE(\mathbf{U}_0^*) \leq 1 - p_1^2 \,. \tag{19}$$

Given an upper bound δ for $ARE(\mathbf{U}_0^*)$, we now show how to obtain a suitable rank r for $\mathbf{U}_0^*$ in CCA. Since $ARE(\mathbf{U}_0^*) \leq 1 - p_1^2$, it sufficient to ensure $1 - p_1^2 \leq \delta \Rightarrow p_1 \geq \sqrt{1-\delta}$. Since p_1 corresponds to $\mathbf{U}_0$ in a PCA setting, one can easily obtain a rank-r $\mathbf{U}_0$ such that $p_1 \geq \sqrt{1-\delta}$. Initializing the iterations for CCA with $\mathbf{U}_0$ will lead to $\mathbf{U}_0^*$ which satisfies $ARE(\mathbf{U}_0^*) \leq \delta$. Note that since the construction is based on a bound, there may be a lower rank $\mathbf{U}_0^*$ which satisfies the constraint.

3.4 Conditions for Global Maximum

We now analyze a condition under which a global maximum of CCA is achieved. The particular case under consideration is when equality holds in (13), i.e., $f(\mathbf{U}_0^*) = f_1^{\max}$, where $\mathbf{U}_0^*$ is the maximum found in Algorithm 1, implying $f(\mathbf{U}_0^*) = f^{\max}$.

We need the following result for the analysis.

Lemma 5 *For any symmetric positive semi-definite matrices* $\mathbf{A}$, $\mathbf{A}_1$, $\mathbf{A}_2$, ... *and vector* $\mathbf{v}$,

(a) $\mathbf{v}^T\mathbf{A}\mathbf{v} = 0$ *iff* $\mathbf{v}^T\mathbf{A}^2\mathbf{v} = 0$;
(b) $\mathbf{v}^T\left(\sum_k \mathbf{A}_k\right)\mathbf{v} = 0$ *iff* $\mathbf{v}^T\mathbf{A}_k\mathbf{v} = 0$ *for every* k;
(c) $\mathrm{colspan}\left(\sum_k \mathbf{A}_k\right) = \mathrm{colspan}\left(\sum_k \mathbf{A}_k^2\right)$
(d) $\mathrm{rank}\left(\sum_k \mathbf{A}_k\right) = \mathrm{rank}\left(\sum_k \mathbf{A}_k^2\right)$

Proof: Let $\mathbf{A} = \mathbf{Q}\mathbf{D}\mathbf{Q}^T$ be the eigendecomposition of $\mathbf{A}$, with $\mathbf{D} = \mathrm{diag}(\mathbf{D}_1, 0)$, where $\mathbf{D}_1$ is a diagonal matrix with strictly positive diagonal elements, and $\mathbf{Q} = [\mathbf{Q}_1, \mathbf{Q}_2]$ is partitioned conformally. Then $\mathbf{A}^2 = \mathbf{Q}\mathbf{D}^2\mathbf{Q}^T$ has the same set of eigenvectors and same nullspace as $\mathbf{A}$. Since $\mathbf{A}$ is positive semi-definite, $\mathbf{v}^T\mathbf{A}\mathbf{v} = 0$ iff $\mathbf{v} \perp \mathbf{Q}_1$, and (a) follows.

To prove (b), note that the term $\mathbf{v}^T\mathbf{A}_k\mathbf{v}$ is never negative, and the left hand side is just the sum of all these terms for all k. A sum of non-negative numbers can be zero iff the numbers themselves are zero. This implies that $\left(\sum_k \mathbf{A}_k\right)\mathbf{v} = 0$ if and only if $\left(\sum_k \mathbf{A}_k^2\right)\mathbf{v} = 0$ for any vector $\mathbf{v}$, which in turn implies that the nullspace of the left hand side of (c) must match the nullspace of the right hand side of (c), proving (c) and (d). ■

Using the above results, we now prove the following theorem.

Theorem 2 *Let* $\mathbf{U}$ *be the* r *principal eigenvectors of* $M(\mathbb{I}_n)$ *associated with nonzero eigenvalues, then* $\mathrm{rank}(M(\mathbf{U}_0)) \geq r$.

Proof: Let s be the rank of $M(\mathbb{I}_n)$ so that $s \geq r$. Then

$$s = \operatorname{rank}\left(\sum_t \mathbf{X}_t^2\right) = \operatorname{rank}\left(\sum_t \mathbf{X}_t\right).$$

For an arbitrary $n \times r$ matrix $\mathbf{U}$ with orthonormal columns,

$$\begin{aligned} p &= \operatorname{rank}\left[\mathbf{U}^T\left(\textstyle\sum_t \mathbf{X}_t\mathbf{U}\mathbf{U}^T\mathbf{X}_t\right)\mathbf{U}\right] \\ &= \operatorname{rank}\left[\textstyle\sum_t \left(\mathbf{U}^T\mathbf{X}_t\mathbf{U}\right)^2\right] \\ &= \operatorname{rank}\left[\textstyle\sum_t \left(\mathbf{U}^T\mathbf{X}_t\mathbf{U}\right)\right] \\ &= \operatorname{rank}\left[\mathbf{U}^T\textstyle\sum_t \left(\mathbf{X}_t\right)\mathbf{U}\right] \\ &\leq r, \end{aligned}$$

with equality if and only if the column space of $\mathbf{U}$ is contained within the column space of $(\sum_t \mathbf{X}_t)$, the latter column space having dimension s. Examples of such a $\mathbf{U}$ include the orthonormal matrix of the eigenvectors corresponding to the leading r eigenvalues of $(\sum_t \mathbf{X}_t)$, or of $\left(\sum_t \mathbf{X}_t^2\right)$.

Using the fact that $\operatorname{rank}(\mathbf{U}^T\mathbf{A}\mathbf{U}) \leq \operatorname{rank}(\mathbf{A})$, it follows that

$$\operatorname{rank}(M(\mathbf{U})) \geq \operatorname{rank}\left[\mathbf{U}^T\left(\sum_t \mathbf{X}_t\mathbf{U}\mathbf{U}^T\mathbf{X}_t\right)\mathbf{U}\right] = p.$$

If $\mathbf{U} \in \operatorname{colsp}(M(\mathbb{I}_n))$, then $p = r$ and the result follows. ∎

Let $\mathbf{U}_0$ be the initialization in Algorithm 1 consisting of the r principal eigenvectors of $M(\mathbb{I})$, and let $\mathbf{U}_0^*$ be the final solution. Based on Theorem 2, we now show that $\operatorname{rank}(M(\mathbf{U}_0)) = r$ is the necessary and sufficient condition that $f(\mathbf{U}_0^*) = f_1^{\max}$, thereby implying that $\mathbf{U}_0^*$ achieves the global optimum. Moreover, in this situation, the solution achieving the global maximum is the initialization $\mathbf{U}_0$ itself.

Theorem 3 *Let $\mathbf{U}_0$ be the solution to CCA1, i.e., the r principal eigenvectors of $M(\mathbb{I})$, and let $\mathbf{U}_0^*$ be the maximum found in Algorithm 1 with initialization $\mathbf{U}_0$. Then, $\operatorname{rank}(M(\mathbf{U}_0)) = r$ is the necessary and sufficient condition that $f(\mathbf{U}_0^*) = f_1^{\max}$. Moreover, $\mathbf{U}_0$ is the solution achieving the global maximum for CCA.*

Proof: Let $\mathbf{U}_0$ be the solution of CCA1, $f_1^{\max} = f_1(\mathbf{U}_0)$. Provided that $\operatorname{rank}(M(\mathbf{U}_0)) = r$, the EVD of $M(\mathbf{U}_0)$ is given by

$$M(\mathbf{U}_0) = \mathbf{U}_1\mathbf{D}_1\mathbf{U}_1^T,$$

where $\mathbf{U}_1$ are the r principal eigenvectors of $M(\mathbf{U}_0)$ associated with the nonzero eigenvalue matrix $\mathbf{D}_1$. According to Theorem 4,

$$f(\mathbf{U}_1) \geq \operatorname{Tr}(\mathbf{U}_1^T M(\mathbf{U}_0)\mathbf{U}_1) = \operatorname{Tr}(M(\mathbf{U}_0)) = f_1(\mathbf{U}_0) = f_1^{\max}$$

On the other hand, we have

$$f_1^{\max} \geq f_1(\mathbf{U}_1) \geq f(\mathbf{U}_1) = f_1^{\max}$$

Therefore, $f_1(\mathbf{U}_1) = f_1^{\max} = f_1(\mathbf{U}_0)$, i.e., $\mathbf{U}_1$ and $\mathbf{U}_0$ spans the same subspace. We can conclude that $f(\mathbf{U}_0) = f(\mathbf{U}_1) = f_1^{\max}$. Since $f^{\max} \leq f_1^{\max}$, $f(\mathbf{U}_0) = f_1^{\max}$ clearly implies $f(\mathbf{U}_0) = f(\mathbf{U}_0^*) = f^{\max}$. $\mathbf{U}_0$ is the solution achieving the global maximum.

We now prove the converse, i.e., $f(\mathbf{U}_0^*) = f_1^{\max} \Rightarrow \operatorname{rank}(M(\mathbf{U}_0)) = r$. Since $f(\mathbf{U}_0^*) = f_1^{\max}$ holds, and since f_1 is the upper bound of f, we have

$$f_1^{\max} = f(\mathbf{U}_0^*) \leq f_1(\mathbf{U}_0^*) \leq f_1^{\max} = f_1(\mathbf{U}_0)$$

So $f_1(\mathbf{U}_0^*) = f_1(\mathbf{U}_0)$, implying $\mathbf{U}_0^*$ and $\mathbf{U}_0$ spans the same subspace. Then

$$f^{\max} \geq f(\mathbf{U}_0^*) = f(\mathbf{U}_0) = f_1^{\max} \geq f^{\max},$$

implying $\mathbf{U}_0$ achieves the global maximum of CCA.

Recall that $\mathbf{U}_0$ are the principal eigenvectors of $M(\mathbf{U}_0)$ corresponding to nonzero eigenvalues, and $\mathrm{rank}(M(\mathbf{U}_0)) \geq r$ according to Theorem 2. If $\mathrm{rank}(M(\mathbf{U}_0)) > r$, there are more than r nonzero eigenvalues. Let $\mathbf{U}_1$ be the r principal eigenvectors of $M(\mathbf{U}_0)$. Then

$$\mathrm{Tr}(M(\mathbf{U}_0)) > \mathrm{Tr}(\mathbf{U}_1^T M(\mathbf{U}_0)\mathbf{U}_1) \geq \mathrm{Tr}(\mathbf{U}_0^T M(\mathbf{U}_0)\mathbf{U}_0) = f(\mathbf{U}_0)\ ,$$

since $\mathbf{U}_1$ are the principal eigenvectors. However, $f_1^{\max} = f_1(\mathbf{U}_0) = \mathrm{Tr}(M(\mathbf{U}_0))$. Consequently, $f(\mathbf{U}_0) < f_1^{\max}$, which contradicts the fact that $f(\mathbf{U}_0) = f_1^{\max}$. Thus, $\mathrm{rank}(M(\mathbf{U}_0)) = r$. ■

A special case of the result is when $\mathrm{rank}(M(\mathbb{I})) = r$. When $\mathrm{rank}(M(\mathbb{I})) = r$, $\mathrm{rank}(M(\mathbf{U}_0)) \leq \mathrm{rank}(M(\mathbb{I})) = r$. According to *Theorem* 2, $\mathrm{rank}(M(\mathbf{U}_0)) \geq r$, implying $\mathrm{rank}(M(\mathbf{U}_0)) = r$. Thus $\mathbf{U}_0$ achieves the global maximum. In this case, since all the eigenvectors are kept, the fraction of energy $p_1 = 1$. The global optimality then follows straightforwardly from the bounds discussed in Section 3.

3.5 Connections to Related Work

Given a set of rectangular matrices $\mathbf{X}_t \in \mathbb{R}^{m\times n}, 1 \leq t \leq T$, the Tucker2 model [35, 16, 24, 22], 2DSVD [9], GLRAM [36], etc., aim to find common components $\mathbf{U} \in \mathbb{R}^{m\times r}$ and $\mathbf{V} \in \mathbb{R}^{n\times s}$ such that

$$\mathbf{X}_t = \mathbf{U}\mathbf{Y}_t\mathbf{V}^T + \mathbf{E}_t \tag{20}$$

where $\mathbf{Y}_t \in^{r\times s}$, $\mathbf{U}$ and $\mathbf{V}$ are orthonormal matrices, and $\mathbf{E}_t$ the residual. $\mathbf{U}$ and $\mathbf{V}$ can be obtained by performing EVD iteratively on matrices $M_1(\mathbf{V}) = \sum_t \mathbf{X}_t\mathbf{V}\mathbf{V}^T\mathbf{X}_t^T \in \mathbb{R}^{m\times m}$ and $M_2(\mathbf{U}) = \sum_t \mathbf{X}_t^T\mathbf{U}\mathbf{U}^T\mathbf{X}_t \in \mathbb{R}^{n\times n}$ respectively. Such methods often use the same initialization as in CCA, since it has been observed empirically that such an initialization usually leads to the good solutions [36, 9], particularly in the rank-1 approximation experiments [27, 20]. When a locally optimal solution is found, say $(\mathbf{U}^*, \mathbf{V}^*)$, Ding et al. [9] established lower and upper bounds for the local optimum based on the eigenvalues of $M_1(\mathbf{V}^*)$ and $M_2(\mathbf{U}^*)$. Since the global optimum is still unknown, their bounds do not tell how close the local optimum is to the global optimum, and hence one does not get approximation guarantees w.r.t. the global optima as we have for CCA.

In (20), if $r = s$ and $\mathbf{Y}_t$ is diagonal, it becomes the PARAFAC / CANDECOMP model with orthonormal constraints [16, 17, 24, 22], abbreviated as PARAFAC in the rest of this paper. Since the off-diagonal elements are zero in $\mathbf{Y}_t$, the PARAFAC model does not allow interactions among components in $\mathbf{U}$ and $\mathbf{V}$. If the PARAFAC model is applied to the covariance matrices in our case, $\mathbf{U}$ and $\mathbf{V}$ are the same. Then the PARAFAC has the same formula as the CCA except that $\mathbf{Y}_t$ is a full matrix in the CCA but is a diagonal matrix in the PARAFAC. However, if covariance matrices are simultaneously diagonalizable [19], i.e., $\mathbf{X}_t = \mathbf{U}\mathbf{Y}_t\mathbf{U}^T$ and $\mathbf{Y}_t$ is diagonal, it turns out that $\mathbf{Y}_t$ is the low dimension covariance matrix in the CCA, as shown in Proposition 2.

Proposition 2 *If covariance matrices are simultaneously diagonalizable, the low dimensional covariance matrix $\mathbf{Y}_t$ in the CCA is diagonal.*

Proof: Suppose covariance matrices are simultaneously diagonalizable, $\mathbf{X}_t = \mathbf{U}\mathbf{Y}_t\mathbf{U}^T$, where $\mathbf{U}$ is the leading r eigenvectors corresponding to the non-zero eigenvalues which are the diagonal entries in $\mathbf{Y}_t$. $M(\mathbb{I}) = \sum_t \mathbf{X}_t^2 = \sum_t \mathbf{U}\mathbf{Y}_t^2\mathbf{U}^T$, thus $\mathbf{U}$ is the solution of CCA1. Plugging U and $\mathbf{X}_t = \mathbf{U}\mathbf{Y}_t\mathbf{U}^T$ into $M(\mathbf{U})$,

$$M(\mathbf{U}) = \sum_t \mathbf{U}\mathbf{Y}_t\mathbf{U}^T\mathbf{U}\mathbf{U}^T\mathbf{U}\mathbf{Y}_t\mathbf{U}^T = \sum_t \mathbf{U}\mathbf{Y}_t^2\mathbf{U}^T$$

Since $rank(M(\mathbf{U})) = r$, U is the solution of the CCA, thus the diagonal matrix $\mathbf{Y}_t$ is the lower dimension covariance matrices in the CCA. ■

4 Algorithm

In this section, we present algorithms for solving CCA for a given dimensionality or a given approximation error bound. For a given dimensionality, we present two algorithms which iteratively improve a given initial solution. For a given approximation error bound, we show how to determine a sufficient dimensionality, and subsequently use any of the iterative algorithms for solving the problem.

4.1 CCA For A Given Dimensionality

Iterative EVD based CCA: For a given dimensionality, EVD can be used to solve for $\mathbf{U}$ in CCA1 as in (8). However, CCA in (3) has four $\mathbf{U}$s which cannot be found using the same approach, since it does not correspond to an EVD problem. Instead, we perform EVD iteratively by fixing two of the inner $\mathbf{U}$ to the current iterate $\mathbf{U}_k$, thereby reducing the problem into an EVD problem. Recall that CCA involves maximizing $f(\mathbf{U}) = \mathrm{Tr}(\mathbf{U}^T M(\mathbf{U})\mathbf{U})$ where $M(\mathbf{U}) = \sum_{t=1}^{T} \mathbf{X}_t \mathbf{U}\mathbf{U}^T \mathbf{X}_t$ is of size $n \times n$. If $\mathbf{U}_k$ is the current iterate, then we compute $M(\mathbf{U}_k)$ and solve the following surrogate problem to obtain $\mathbf{U}_{k+1}$:

$$\max_{\mathbf{U}^T\mathbf{U}=\mathbb{I}_r} \mathrm{Tr}(\mathbf{U}^T M(\mathbf{U}_k)\mathbf{U}) \,. \tag{21}$$

Clearly, $\mathbf{U}_{k+1}$ can be obtained by applying rank-r EVD on $M(\mathbf{U}_k)$. The idea behind such an update has been explored in the literature on tensor decomposition [24, 25, 9, 36]. As the following result shows, such as update will improve the objective function, i.e., $f(\mathbf{U}_{k+1}) \geq f(\mathbf{U}_k)$.

Theorem 4 *Let $\mathbf{U}_{k+1}$ be the r principal eigenvectors of $M(\mathbf{U}_k)$, then $f(\mathbf{U}_{k+1}) \geq \mathrm{Tr}(\mathbf{U}_{k+1}^T M(\mathbf{U}_k)\mathbf{U}_{k+1}) \geq f(\mathbf{U}_k)$. The equality holds when $\mathbf{U}_{k+1}$ and $\mathbf{U}_k$ spans the same subspace.*

Proof: We define the matrix $\mathbf{A}_k = \mathbf{A}(\mathbf{U}_k)$ as follows

$$\mathbf{A}_k = \left[\mathbf{X}_1^{\frac{1}{2}}\mathbf{U}_k\mathbf{U}_k^T\mathbf{X}_1^{\frac{1}{2}}, \cdots, \mathbf{X}_T^{\frac{1}{2}}\mathbf{U}_k\mathbf{U}_k^T\mathbf{X}_T^{\frac{1}{2}}\right]$$

By definition, we have

$$\begin{aligned}
\mathrm{Tr}(\mathbf{A}_k\mathbf{A}_k^T) &= \mathrm{Tr}\left(\sum_{t=1}^{T} \mathbf{X}_t^{\frac{1}{2}}\mathbf{U}_k\mathbf{U}_k^T\mathbf{X}_t^{\frac{1}{2}}\mathbf{X}_t^{\frac{1}{2}}\mathbf{U}_k\mathbf{U}_k^T\mathbf{X}_t^{\frac{1}{2}}\right) \\
&= \mathrm{Tr}\left(\sum_{t=1}^{T} \mathbf{U}_k^T\mathbf{X}_t\mathbf{U}_k\mathbf{U}_k^T\mathbf{X}_t\mathbf{U}_k\right) = f(\mathbf{U}_k)
\end{aligned}$$

Let $\mathbf{U}_{k+1}$ be the r principal eigenvectors of $M(\mathbf{U}_k)$. By a similar analysis, $f(\mathbf{U}_{k+1}) = \mathrm{Tr}(\mathbf{A}_{k+1}\mathbf{A}_{k+1}^T)$. Now note that

$$\begin{aligned}
\mathrm{Tr}(\mathbf{A}_k\mathbf{A}_{k+1}^T) &= \mathrm{Tr}\left(\sum_{t} \mathbf{X}_t^{\frac{1}{2}}\mathbf{U}_k\mathbf{U}_k^T\mathbf{X}_t^{\frac{1}{2}}\mathbf{X}_t^{\frac{1}{2}}\mathbf{U}_{k+1}\mathbf{U}_{k+1}^T\mathbf{X}_t^{\frac{1}{2}}\right) \\
&= \mathrm{Tr}\left(\mathbf{U}_{k+1}^T\sum_{t} \mathbf{X}_t\mathbf{U}_k\mathbf{U}_k^T\mathbf{X}_t\mathbf{U}_{k+1}\right) \\
&= \mathrm{Tr}(\mathbf{U}_{k+1}^T M(\mathbf{U}_k)\mathbf{U}_{k+1})
\end{aligned}$$

Given that $\mathbf{U}_{k+1}$ is the r principal eigenvectors of $M(\mathbf{U}_k)$, then

$$\mathrm{Tr}(\mathbf{U}_{k+1}^T M(\mathbf{U}_k)\mathbf{U}_{k+1}) \geq \mathrm{Tr}(\mathbf{U}_k^T M(\mathbf{U}_k)\mathbf{U}_k) = f(\mathbf{U}_k) \,, \tag{22}$$

where the equality holds iff $\mathbf{U}_{k+1}$ and $\mathbf{U}_k$ span the same subspace.

Algorithm 1 Iterative EVD (IEVD) Algorithm for CCA

1: Input: $\mathbf{X}_t, 1 \leq t \leq T$, initialization $\mathbf{U}_0 \in \mathbb{R}^{n \times r}$
2: Output: $\mathbf{U}, \mathbf{Y}_t, 1 \leq t \leq T$
3: **repeat**
4: Perform EVD on $M(\mathbf{U}_k) = \sum_t \mathbf{X}_t \mathbf{U}_k \mathbf{U}_k^T \mathbf{X}_t$
5: Choose the leading r eigenvectors $\mathbf{U}_{k+1}$
6: Compute $\mathbf{Y}_t = \mathbf{U}_{k+1}^T \mathbf{X}_t \mathbf{U}_{k+1}$
7: **until** $\left| \frac{f(\mathbf{U}_{k+1}) - f(\mathbf{U}_k)}{f(\mathbf{U}_k)} \right| \leq \varepsilon$

Then, we have

$$\begin{aligned} f(\mathbf{U}_k) f(\mathbf{U}_{k+1}) &= \mathrm{Tr}(\mathbf{A}_k \mathbf{A}_k^T)\, \mathrm{Tr}(\mathbf{A}_{k+1} \mathbf{A}_{k+1}^T) \\ &\overset{(a)}{\geq} \left[\mathrm{Tr}(\mathbf{A}_k \mathbf{A}_{k+1}^T)\right]^2 \\ &= \left[\mathrm{Tr}(\mathbf{U}_{k+1}^T M(\mathbf{U}_k) \mathbf{U}_{k+1})\right]^2 , \end{aligned}$$

where (a) follows from Lemma 2, and the equality holds when $\mathbf{U}_{k+1}$ and $\mathbf{U}_k$ span the same subspace. Consider (22), and $f(\mathbf{U}_k)$, $f(\mathbf{U}_{k+1})$ are nonnegative, there is

$$f(\mathbf{U}_{k+1}) \geq \mathrm{Tr}(\mathbf{U}_{k+1}^T M(\mathbf{U}_k) \mathbf{U}_{k+1}) \geq f(\mathbf{U}_k) \tag{23}$$

The equality holds when $\mathbf{U}_k$ and $\mathbf{U}_{k+1}$ spans the same subspace. ■

Algorithm 1 presents the corresponding algorithm for a given dimensionality r as input. The objective function increases every step until a certain stopping criterion is satisfied. If $\mathbf{U}_0^*$ is the final solution, from the analysis of Section 3, we know that $f(\mathbf{U}_0^*) \geq p_1 f^{\max}$, and the approximate relative error satisfies $1 - p_1 \leq ARE(\mathbf{U}_0^*) \leq 1 - p_1^2$.

Auxiliary Function based CCA: In I-EVD, the update has to repeatedly calculate the EVD of a $n \times n$ matrix. If n is large, the update becomes a bottleneck. In this section, we present an efficient update which only calculates the SVD of a $r \times n$ matrix. To introduce the new update, we first define an auxiliary function $g(\mathbf{U}, \mathbf{V})$ as follows

$$g(\mathbf{U}, \mathbf{V}) = \mathrm{Tr}\left(\sum_t^T (\mathbf{U}^T \mathbf{X}_t \mathbf{U})(\mathbf{V}^T \mathbf{X}_t \mathbf{V}) \right) . \tag{24}$$

where $\mathbf{U}^T \mathbf{U} = \mathbb{I}_r$ and $\mathbf{V}^T \mathbf{V} = \mathbb{I}_r$. Clearly, $g(\mathbf{U}, \mathbf{U}) = f(\mathbf{U})$.

Given $\mathbf{U}_k$, if we can find a $\mathbf{U}_{k+1}$ satisfying $g(\mathbf{U}_k, \mathbf{U}_{k+1}) \geq g(\mathbf{U}_k, \mathbf{U}_k)$, the auxiliary function increases. Theorem 5 shows that $\mathbf{U}_{k+1}$ can be obtained by performing the SVD on the $r \times n$ matrix $\sum_t^T \mathbf{Y}_t^k \mathbf{V}^T \mathbf{X}_t$, where $\mathbf{Y}_t^k = \mathbf{U}_k^T \mathbf{X}_t \mathbf{U}_k$. Such a $\mathbf{U}_{k+1}$ increases $f(\mathbf{U})$.

To prove that Theorem 5, we need the following results.

Lemma 6 $\mathrm{Tr}(\mathbf{A^T B}) = vec(\mathbf{A})^T vec(\mathbf{B})$.

Proof: Let $\mathbf{a}_i, \mathbf{b}_i$ be the columns of $\mathbf{A}, \mathbf{B}$ respectively, then $\mathrm{Tr}(\mathbf{A^T B}) = \sum_i \mathbf{a}_i \mathbf{b}_i = vec(\mathbf{A})^T vec(\mathbf{B})$. ■

Lemma 7 (Wiki[1]) $vec(\mathbf{AXB}) = (\mathbf{B}^T \otimes \mathbf{A}) vec(\mathbf{X})$.

Theorem 5 *Let $\mathbf{U}_{k+1} = \mathbf{Q}\mathbf{P}^T$, where $\mathbf{P}$ and $\mathbf{Q}$ are the left and right r singular vectors of $\sum_t^T \mathbf{Y}_t^k \mathbf{U}_k^T \mathbf{X}_t$, where $\mathbf{Y}_t^k = \mathbf{U}_k^T \mathbf{X}_t \mathbf{U}_k$, then*

$$f(\mathbf{U}_k) \leq g(\mathbf{U}_k, \mathbf{U}_{k+1}) \leq f(\mathbf{U}_{k+1})$$

The equality holds when $\mathbf{U}_k$ and $\mathbf{U}_{k+1}$ span the same subspace.

Proof: Denote the SVD $\sum_t^T \mathbf{Y}_t^k \mathbf{U}_k^T \mathbf{X}_t = \mathbf{P}\mathbf{D}\mathbf{Q}^T$, then

$$g(\mathbf{U}_k, \mathbf{U}_k) = \mathrm{Tr}(\sum_t \mathbf{Y}_t^k \mathbf{U}_k^T \mathbf{X}_t \mathbf{U}_k) = \mathrm{Tr}(\mathbf{P}\mathbf{D}\mathbf{Q}^T \mathbf{U}_k) \overset{(a)}{\leq} \mathrm{Tr}(\mathbf{D})$$

(a) holds because $\mathbf{P}, \mathbf{Q}$ and $\mathbf{U}_k$ are orthonormal matices, and the SVD attains the global maximum. The equality holds when $\mathbf{U}_k$ spans the same subspace as $\mathbf{Q}\mathbf{P}^T$.

Let $\mathbf{U}_{k+1} = \mathbf{Q}\mathbf{P}^T$, there is

$$\mathrm{Tr}(\sum_t \mathbf{Y}_t^k \mathbf{U}_k^T \mathbf{X}_t \mathbf{U}_{k+1}) = \mathrm{Tr}(\mathbf{P}\mathbf{D}\mathbf{P}^T) = \mathrm{Tr}(\mathbf{D}) \geq g(\mathbf{U}_k, \mathbf{U}_k)$$

On the other hand, there is

$$\begin{aligned}
\mathrm{Tr}(\sum_t \mathbf{Y}_t^k \mathbf{U}_k^T \mathbf{X}_t \mathbf{U}_k) &\overset{(a)}{=} vec(\sum_t \mathbf{X}_t \mathbf{U}_k \mathbf{Y}_t^k)^T vec(\mathbf{U}_k) \\
&\overset{(b)}{=} vec(\mathbf{U}_k)^T \left(\sum_t \mathbf{Y}_t^k \otimes \mathbf{X}_t \right) vec(\mathbf{U}_k)
\end{aligned}$$

(a) follows from lemma 6 and (b) from lemma 7. Since $\mathbf{Y}_t^k$ and $\mathbf{X}_t$ are positive semi-definite, $\sum_t \mathbf{Y}_t^k \otimes \mathbf{X}_t$ is positive semi-definite. According to the Cauchy-Schwarz inequality,

$$\begin{aligned}
g(\mathbf{U}_k, \mathbf{U}_k) g(\mathbf{U}_k, \mathbf{U}_{k+1}) &\geq \left[vec(\mathbf{U}_k)^T \sum_t \mathbf{Y}_t^k \otimes \mathbf{X}_t vec(\mathbf{U}_{k+1}) \right]^2 \\
&= \left[\mathrm{Tr}(\sum_t \mathbf{Y}_t^k \mathbf{U}_k^T \mathbf{X}_t \mathbf{U}_{k+1}) \right]^2
\end{aligned}$$

The equality holds when $\mathbf{U}_k$ and $\mathbf{U}_{k+1}$ spans the same subspace. Thus

$$g(\mathbf{U}_k, \mathbf{U}_{k+1}) \geq \mathrm{Tr}(\sum_t \mathbf{Y}_t^k \mathbf{U}_k^T \mathbf{X}_t \mathbf{U}_{k+1}) \geq g(\mathbf{U}_k, \mathbf{U}_k)$$

Now we prove the second inequality. Following the Cauchy-Schwarz inequality,

$$\begin{aligned}
& g(\mathbf{U}_k, \mathbf{U}_k) g(\mathbf{U}_{k+1}, \mathbf{U}_{k+1}) \\
&= \mathrm{Tr}(\sum_t (\mathbf{U}_k^T \mathbf{X}_t \mathbf{U}_k)^2) \, \mathrm{Tr}(\sum_t (\mathbf{U}_{k+1}^T \mathbf{X}_t \mathbf{U}_{k+1})^2) \\
&\overset{(a)}{\geq} \left[\mathrm{Tr} \left(\sum_t (\mathbf{U}_k^T \mathbf{X}_t \mathbf{U}_k)(\mathbf{U}_{k+1}^T \mathbf{X}_t \mathbf{U}_{k+1}) \right) \right]^2 \\
&= g^2(\mathbf{U}_k, \mathbf{U}_{k+1}) \, ,
\end{aligned}$$

The equality holds in (a) when $\mathbf{U}_k$ and $\mathbf{U}_{k+1}$ spans the same subspace.

Since $g(\mathbf{U}_k, \mathbf{U}_k) \leq g(\mathbf{U}_k, \mathbf{U}_{k+1})$, then

$$g(\mathbf{U}_k, \mathbf{U}_{k+1}) \leq g(\mathbf{U}_{k+1}, \mathbf{U}_{k+1}) = f(\mathbf{U}_{k+1})$$

■

With theorem 5, an algorithm based on the auxiliary function is proposed, which is presented in Algorithm 2. The solution of Algorithm 2 satisfies the bounds established in Section 3.

Algorithm 2 Auxiliary Function (AF) Algorithm for CCA

Input: $\mathbf{X}_t, 1 \leq t \leq T$, initialization $\mathbf{U}_0 \in \mathbb{R}^{n \times r}$
Output: $\mathbf{U}, \mathbf{Y}_t, 1 \leq t \leq T$
Compute $\mathbf{Y}_t^0 = \mathbf{U}_0^T \mathbf{X}_t \mathbf{U}_0$
repeat
  Perform the SVD on matrix $\sum_t \mathbf{Y}_t^k \mathbf{U}_k^T \mathbf{X}_t = \mathbf{PDQ}^T$
  Compute $\mathbf{U}_{k+1} = \mathbf{QP}^T$
  Compute $\mathbf{Y}_t^{k+1} = \mathbf{U}_{k+1}^T \mathbf{X}_t \mathbf{U}_{k+1}$
until $\left| \frac{g(\mathbf{U}_{k+1}, \mathbf{U}_{k+1}) - g(\mathbf{U}_k, \mathbf{U}_k)}{g(\mathbf{U}_k, \mathbf{U}_k)} \right| \leq \varepsilon$

4.2 CCA For A Given Approximation Error

We consider a setting where instead of the dimensionality r, an upper bound δ on the approximate relative error (ARE) is given. In such a setting, one can choose a sufficient dimensionality r and a corresponding initialization $\mathbf{U}_0$ based on our analysis in Section 3, and use any of the algorithms in Section 4.1 to obtain a $\mathbf{U}$ which guarantees the error bound. In particular, it is sufficient to choose the dimensionality r of the initialization $\mathbf{U}_0$ such that the fraction of energy captured in CCA1 given by $p_1 = \frac{\mathrm{Tr}(\mathbf{U}_0^T M(\mathbb{I}) \mathbf{U}_0)}{\mathrm{Tr}(M(\mathbb{I}))}$ satisfies $p_1 \geq \sqrt{1-\delta}$, as discussed in Section 3. Since $M(\mathbb{I})$ is fixed, and CCA1 is an EVD problem, choosing a suitable dimensionality r such that $p_1 \geq \sqrt{1-\delta}$ is straightforward since EVD has a nested structure. If such a $\mathbf{U}_0$ is used to initialize the algorithms in Section 4.1, the final solution $\mathbf{U}_0^*$ will satisfy $f(\mathbf{U}_0^*) \geq \sqrt{1-\delta} f^{max}$ and $ARE(\mathbf{U}_0^*) \leq \delta$, which is the prescribed bound on the approximation error.

5 Experimental Results

In this section, the performance of CCA is evaluated on both artificial datasets and two real-world stock market datasets respectively spanning 21 years from 1990-2010, and 14 years from 1971-1984. Evaluation is done in terms of the *Approximate Relative Error* (ARE) (16) for all datasets, and also the ability to track volatility in low-dimensions for the stock market datasets. The performance of CCA is compared with PARAFAC with orthonormal constraints, PCA, and Random Projection (RP) [8, 1]. While CCA and PARAFAC are computed on the entire set of covariance matrices, PCA is computed based on the single aggregated covariance. For RP, $\mathbf{U}$ was generated as follows: (i) Each entry of $\mathbf{U}$ is generated via an i.i.d. normal distribution; and (ii) $\mathbf{U}$ is normalized via Gram-Schmidt orthogonalization [15] and normalization.

5.1 Artificial Data

Artificial data was generated following the model in (1). In particular, $\mathbf{Y}_t$ and $\mathbf{U}$ were generated first, then $\mathbf{X}_t$ was calculated by adding noise to $\mathbf{U}\mathbf{Y}_t\mathbf{U}^T$. $\mathbf{Y}_t$ was generated as the covariance matrix of a set of randomly generated samples. The samples were generated from the following four Gaussian distributions with mean

$$m1 = [0, 0], m2 = [5, 0], m3 = [0, 5], m4 = [5, 5]$$

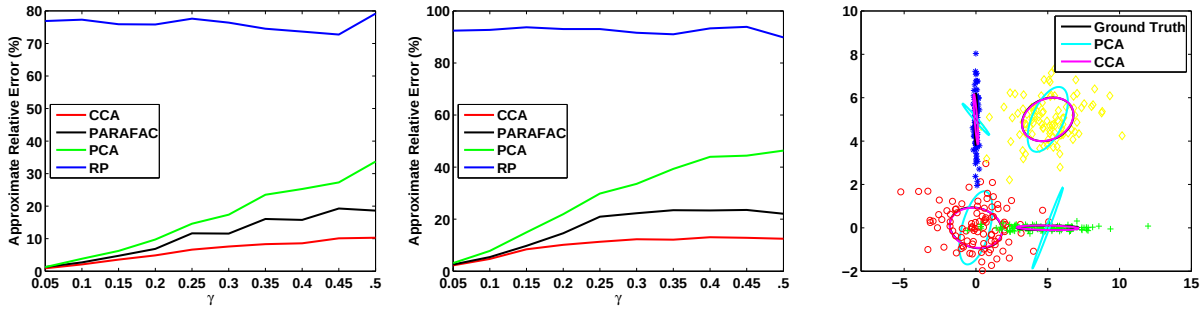
and covariance

$$[\mathbf{\Sigma}_1 | \mathbf{\Sigma}_2 | \mathbf{\Sigma}_3 | \mathbf{\Sigma}_4] = \left[\begin{array}{cc|cc|cc|cc} 4 & 0 & 4 & 0 & 0.01 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0.01 & 0 & 2 & 0 & 2 \end{array}\right],$$

Instead of using a fixed $\mathbf{U}$, it was mildly perturbed as follows:

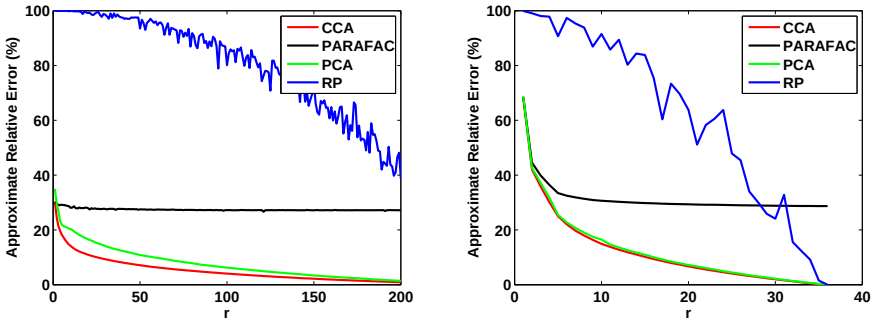
$$\mathbf{U}_{t+1} \leftarrow QR(\mathbf{U}_t + \gamma E_t) \tag{25}$$

where γ is a small constant, $E_t \in \mathbb{R}^{n \times r}$ where $E_{ij} \sim N(0, 1)$, and $\mathbf{U}_{t+1}$ is obtained from the QR factorization of $(\mathbf{U}_t + \gamma E_t)$. In (25), $\mathbf{U}_1$ is randomly generated, $r = 2$, and we consider two values of the high-dimensionality $n = 5, 10$. The experiment was repeated 50 times, and the final results reported are the average over the 50 runs.



(a) $n = 5$ (b) $n = 10$ (c) Latent Covariances in 2D

Figure 1: (a)-(b) Approximation Relative Error (ARE) on artificial data in different dimensions r and increasing noise level γ. CCA outperforms PARAFAC, PCA and RP, especially with high noise levels. (c) 2D latent covariances. CCA tracks the true covariance better than PCA.



(a) S&P500 (b) NYSE

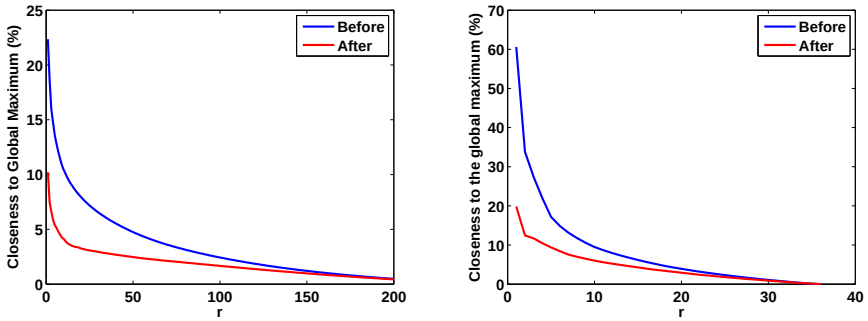
Figure 2: Approximation Relative Error (ARE) on S&P 500 and NYSE in different dimensions r. CCA outperforms PARAFAC, PCA and RP.

Results: Figure 1 (a)-(b) shows the comparative performance of CCA, PARAFAC, PCA, and RP in terms of ARE (lower is better) across different noise levels γ for fixed low-dimensionality $r = 2$. As seen in the figure, CCA outperforms PARAFAC and PCA, and significantly outperforms RP. The improvements of CCA over other methods is more pronounced for high noise levels (high γ). For low noise levels, CCA and PARAFAC are competitive since all the covariance matrices are nearly diagonal. Because of the structure of the covariance matrices, i.e., nearly diagonal but different, PARAFAC outperforms PCA which maximizes the total covariance instead.

Figure 1(c) shows the shape of 2-dimensional covariances when $n = 10, \gamma = 0.1$. The ground truth are estimated covariances from the samples plotted in black. The latent covariances for CCA and PCA shown respectively in magenta and cyan, were calculated based on the leading 2 components. A visual comparison readily shows that CCA is able to recover the ground truth, while PCA seems to find a subspace which maximize the total covariance but not suitable for separate covariances. Figure 1 is a canonical example of a situation where CCA will always outperform PCA.

5.2 Stocks Data

We considered two real world stock market datasets. The first dataset, S&P500, is based on daily closing prices of all 263 stocks in the current S&P500 index which has been in the S&P index from 1990 to 2010. The second dataset, NYSE, is a widely used dataset of daily closing prices consisting of 36 stocks at daily resolution spanning from 1971 to 1984 [18, 2, 7].



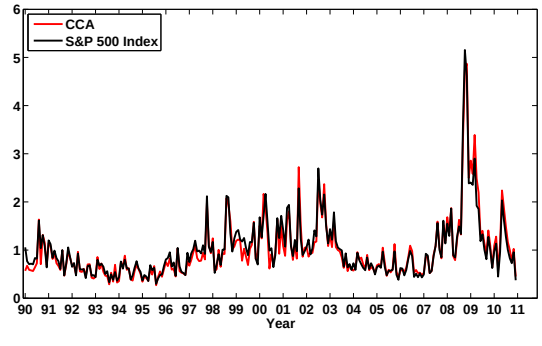
(a) S&P500 (b) NYSE

Figure 3: The upper bound of the relative error of the CCA results w.r.t. the global maximum on S&P 500 and NYSE in different dimension r. The upper bounds before and after the CCA runs are in blue and red respectively. The CCA results are very close to the global maximum.

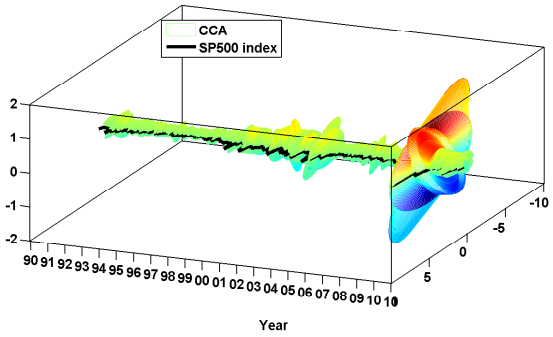
Methology: For the experiments, the covariance of the daily log-return was considered for both datasets, where return$_t$ = $\log \frac{x_t}{x_{t-1}} \times 100\%$, x_t is the daily closing stock price. For each dataset, we construct the monthly average of the daily covariances, and each average monthly covariance was considered as an observed covariance matrix $\mathbf{X}_t$. For S&P500, there are $21 \times 12 = 252$ observed covariance matrices $\mathbf{X}_t \in \mathbb{R}^{263\times 263}$. For NYSE, there are $14 \times 12 = 168$ covariance matrices $\mathbf{X}_t \in \mathbb{R}^{36\times 36}$.

ARE: The performance of the four methods evaluated in terms of ARE on S&P500 and NYSE are shown in Figure 2 . On both datasets, CCA outperforms PARAFAC and PCA, and significantly outperforms RP. Interestingly, PARAFAC does not improve with increasing r (dimensionality) possibly because the covariances cannot be simultaneously diagonalized. PCA performs much better than PARAFAC, which is in direct contrast with the observed results for the artificial dataset. Note that CCA performs the best on both types of data, which illustrates the flexibility of the model. PCA is competitive with CCA on NYSE, but worse on S&P500 except for high-dimensions when PCA tends to catch up. There are two possible explanations: NYSE is a low-d dataset with with $n = 36$, whereas S&P500 is relatively high-d with $n = 263$; and the stock market has been more volatile in the 1990-2010 range (S&P500) as compared to the 1971-1984 range (NYSE). Figure 3 shows the upper bound of the relative error of CCA results w.r.t. the global maximum. The upper bound $1 - p1$ in Corollary 2 is plotted in blue, and it is known before the CCA runs. When r is small, the upcoming maximum is far from the global maximum, but it definitely approaches the global maximum as r increases. After the local maximum $f(\mathbf{U}_0^*)$ is found, the upper bound $\frac{f_1^{\max} - f(\mathbf{U}_0^*)}{f_1^{\max}}$ in (14) is plotted in red. It shows that the upper bound has been greatly improved, and the local maximum actually is very close to the global maximum.

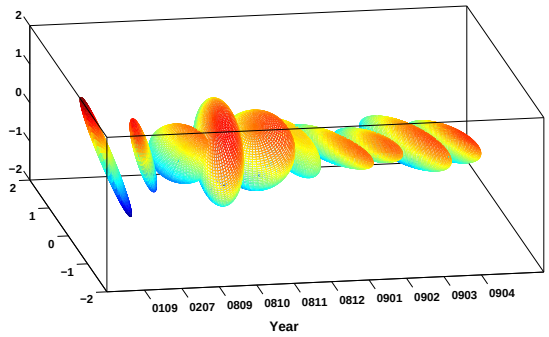
Volatility: In Figures 4 and 5 we plot the latent covariance matrices (level sets) obtained from CCA in dimensions $r = 1, 2, 3$ for S&P500 and NYSE, and compare them to the volatilities [5, 11, 10] of their proxies. The proxy of S&P 500 dataset is the S&P500 index, while the proxy of NYSE is the average of 36 stocks. The reason we expect $\mathbf{Y}_t$ to track volatility well is as follows: For n stocks, the trace of the covariance $\mathbf{X}_t$ is equal to $n\sigma^2$, where σ is the volatility (standard deviation) of their proxies. If $\mathbf{Y}_t$ approximates $\mathbf{X}_t$ well, the trace of $\sqrt{\mathbf{Y}_t/n}$ should approximate σ. In both datasets, for 1D, $\sqrt{\mathbf{Y}_t/n}$ tracks the volatility almost exactly. For 2D, $\sqrt{\mathbf{Y}_t/n}$ are ellipses which change shape/size over time, and the volatility (black curve) is always on the circumference of the ellipses. It is interesting to note that the latent covariances for S&P 500 (Figure 4) seem to capture the two major financial meltdowns, viz the dot-com bubble around 2001 and the major financial crisis around 2008, even in such a low dimension. The crisis in 2008 looks significantly worse, and the ellipses in the 2D plot have different shapes possibly indicating different market segments being more adversely affected. Similarly, the latent covariances for NYSE (Figure 5) capture the stock market crash around 1973-1974 resulting from the collapse of the Bretton Woods system along with the 'Nixon Shock' and the devaluation of the US dollar. We also show the largest 10 3D latent covariances. As seen in Fig. 4(c), the largest 10 3D latent covariances of S&P 500 correspond to month 200109(911 terrorist attacks), 200207(stock market downturn) and 8 months after the Lehman Brothers' bankruptcy in September 2008. The largest 10 3D covariances of NYSE basically correspond to the collapse of the Bretton Woods system, as plotted in Fig. 5(c). Such interpretable results show the potential of CCA in high-dimensional real world problems.



(a) 1D Latent Covariances S&P 500

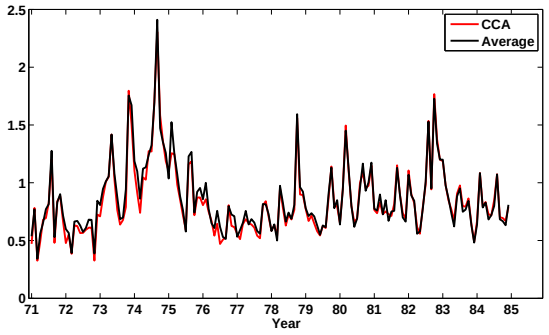


(b) 2D Latent Covariances S&P 500

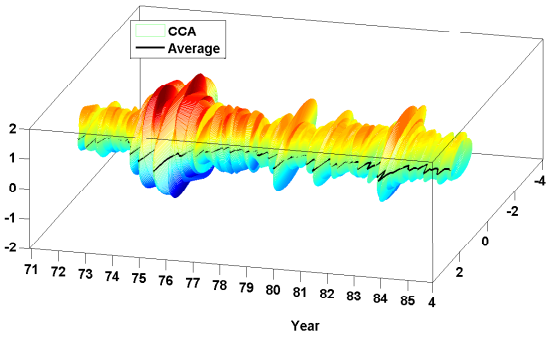


(c) The largest 10 3D Latent Covariances S&P 500

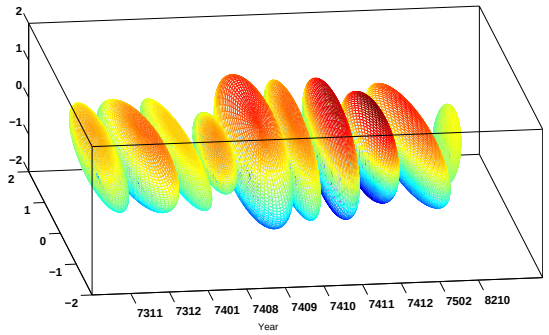
Figure 4: Latent Covariances over time for S&P500 from 1990 to 2010. The two financial meltdowns in 2001 and 2008 are prominently captured in the latent low dimensional space. (Best viewed in color)



(a) 1D Latent Covariances NYSE



(b) 2D Latent Covariances NYSE



(c) The largest 10 3D Latent Covariances NYSE

Figure 5: Latent Covariances over time for NYSE from 1970 to 1984. The stock market crash of 1974 is captured in the latent low dimensional space. (Be viewed in color)

Choose r given ARE: We also evaluated the efficacy of choosing dimensionality r given an ARE upper bound. The results on S&P 500 and NYSE dataset are shown in Table 2 and 3. The first row is the given ARE upper bound δ, the second row shows the sufficient r computed as in Section 4.2 and the corresponding ARE, and the third row shows the smallest r which would have satisfied the bound and the corresponding ARE. The chosen r satisfies the bound, but can be conservative at times especially when ARE decreases rapidly with increasing r, as in the cases of $\sigma = 20$ and 10 on S&P 500.

$\delta(\%)$	30	20	10	5
Chosen r (ARE)	3(21.50)	10(14.18)	45(7.58)	97(4.20)
Smallest r (ARE)	2(24.67)	4(19.70)	26(9.88)	81(5.00)

Table 2: Choosing r given an ARE upper bound on S&P 500.

$\delta(\%)$	30	20	10	5
Chosen r (ARE)	6(22.06)	9(16.36)	18(7.94)	25(4.07)
Smallest r (ARE)	5(24.99)	7(19.71)	16(9.37)	24(4.54)

Table 3: Choosing r given an ARE upper bound on NYSE.

Running Time: Figure 6 compares the running times (in seconds) of Algorithm 1 and Algorithm 2 on the S&P 500 dataset. The experiments were run in Matlab 7.1 on an Intel P8600 2.4GHz PC with Windows Vista OS and 2G memory. When r is small, i.e., low dimensional projections, Algorithm 2 is much faster than Algorithm 1. As r increases, Algorithm 2 possibly spends more time on the SVD step, and probably require more steps to converge, so its superiority in running time decreases.

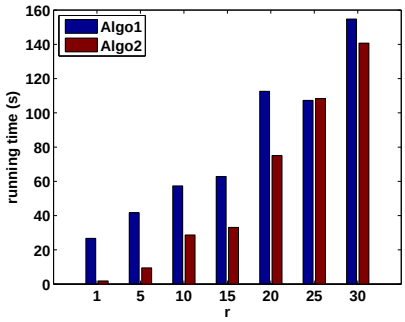


Figure 6: Running times of Algorithm 1 and Algorithm 2 on the S&P 500 dataset for different dimensionality. The auxiliary function based method (Algorithm 2) is distinctly faster for low-d projections.

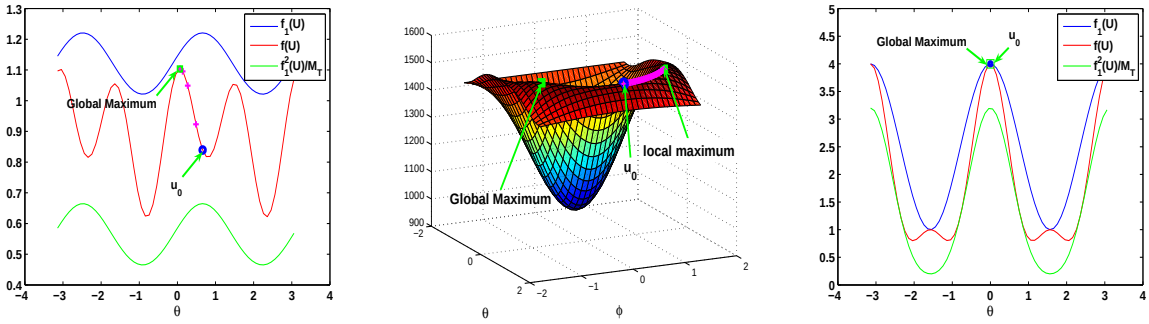
5.3 Additional Numerical Simulations

We study the CCA (Algorithm 1) on low dimensional problems to get additional insights into workings of the proposed ideas, including cases where the approach can and cannot find the global maxima of $f(\mathbf{U})$. It is important to recall that while $f(\mathbf{U})$ is a convex function for unconstrained $\mathbf{U}$, the model requires maximizing $f(\mathbf{U})$ on the domain of $\mathbf{U}$ determined by $\mathbf{U}^T\mathbf{U} = \mathbb{I}_r$, and the problem may thus have multiple local maxima.

We illustrate different scenarios for using Algorithm 1 to solve CCA in Figure 7. In Figure 7(a), we consider 3 time steps for a 2-dimensional covariance matrix, with

$$\mathbf{X} = [\mathbf{X}_1|\mathbf{X}_2|\mathbf{X}_3] = \left[\begin{array}{cc|cc|cc} 1 & 0 & 0 & 0 & 0.22 & 0.22 \\ 0 & 0.25 & 0 & 1 & 0.22 & 0.22 \end{array}\right].$$

The vector $\mathbf{u}$ is parameterized as $\mathbf{u} = [\sin(\theta), \cos(\theta)]$, and the x-axis denotes θ. Note that $f(\mathbf{u})$ is convex in $\mathbf{u}$ but not that θ, which explains the nonconvex plot of the objective (in red). Further, the domain of θ in $[-\pi, \pi]$, and the



(a) Global maximum found with the proposed initialization

(b) The global maximum is not found with the proposed initialization

(c) The global maximum is the initialization

Figure 7: Optimizing $f(\mathbf{U})$ in CCA based on CCA1 initialization and iterative updates. Objective $f(\mathbf{U})$ for CCA is shown in red; the lower and upper bounds based on $f_1(\mathbf{U})$ for CCA1 is shown in green and blue respectively. Three scenarios: (a) Iterations converge to the global maxima, (b) Iterations converge to a local maxima, and (c) Initialization is the global maxima.

function is periodic beyond that domain. Algorithm 1 is used to find the best rank-1 approximation $\mathbf{u}$. In particular, the initialization $\mathbf{u}_0$ is the optimal solution of $f_1(\mathbf{u})$, which is denoted by a small blue circle $\circ$. The searching trajectory is denoted by magenta +, and the optimal solution of $f(\mathbf{U})$ by a green $\square$. The upper and lower bounds are plotted in blue and green respectively. For this scenario, with the proposed initialization, the global maximum can be found, as illustrated in Figure 7(a). However, the initialization does not always lead to the global maximum as shown in Figure 7(b). In Figure 7(b), we consider

$$\mathbf{X}_1 = \begin{bmatrix} 29.7995 & 2.5707 & 1.7377 \\ 2.5707 & 30.1445 & -0.0292 \\ 1.7377 & -0.0292 & 24.1799 \end{bmatrix},$$

$$\mathbf{X}_2 = \begin{bmatrix} 21.8515 & -2.2068 & 2.0377 \\ -2.2068 & 22.8371 & 0.0490 \\ 2.0377 & 0.0490 & 21.1336 \end{bmatrix},$$

$$\mathbf{X}_3 = \begin{bmatrix} 8.5273 & -2.5322 & 1.1011 \\ -2.5322 & 9.6724 & -0.9796 \\ 1.1011 & -0.9796 & 6.4754 \end{bmatrix},$$

and the vector $\mathbf{u}$ is parameterized as $\mathbf{u} = [\sin(\theta), \cos(\theta)\sin(\phi), \cos(\theta)\cos(\phi)]$. In Figure 7(b), θ and ϕ are the x-axis and y-axis respectively, and $f(\mathbf{u})$ is shown in the z-axis. For this scenario, the final solution is a good local maxima but is not the global maxima, which is also marked in the figure. Finally, Figure 7(c) shows a case where the initialization itself achieves the global maximum of CCA. In Figure 7(c), we consider

$$\mathbf{X} = [\mathbf{X}_1 | \mathbf{X}_2] = \left[\begin{array}{cc|cc} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{array}\right],$$

and $\mathbf{u}$ is parameterized as in Figure 7(a). For this scenario, if $\mathbf{u}_0$ denotes the initialization obtained from CCA1, we see that $f_1^{\max} = f_1(\mathbf{u}_0) = f(\mathbf{u}_0)$, implying $f(\mathbf{u}_0) = f^{\max}$.

6 Conclusions

In this paper, we have introduced a framework called CCA for simultaneously modeling multiple covariance matrices in low dimensions. While the framework has similarities with existing approaches to tensor decomposition, we present a novel and unique analysis of the CCA in terms of a more tractable PCA framework called CCA1, which provides the lower and upper bounds for the global maximum for the CCA. The bounds also lead to an effective initialization scheme so that the results of the CCA has clear approximation guarantees w.r.t. the global maximum. We also discuss

non-trivial conditions under which a global maximum will be achieved. Two algorithms, a standard tensor decomposition algorithm and an efficient auxiliary function based algorithm, are presented. They can work with either a fixed dimensionality or a approximate relative error. We illustrate the effectiveness of the approach on synthetic data as well as two real world stock market datasets.

While the CCA can be considered as a special case of classical tensor decomposition methods, the analysis presented in the paper relates the two issues encountered in the general case. Such an analysis can potentially be extended to more general settings considered in the tensor decomposition literature, and will be considered in the future work. In the analysis, all covariance matrices were assumed to be available. In real life domains such as finance and climate sciences, the observed covariance matrices become available over time. We plan to investigate extensions of the CCA framework to the online setting where the observed matrices become available over time.

Acknowledgment

This research was supported by NSF grants IIS-0916750, IIS-0812183, IIS-0534286, NSF CAREER grant IIS-0953274, and NASA grant NNX08AC36A.

References

[1] D. Achlioptas. Database-friendly random projections. In *ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 274–281, 2001.

[2] A. Agarwal, E. Hazan, S. Kale, and R. E. Schapire. Algorithms for portfolio management based on the newton method. In *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, 2006.

[3] T. Anderson. *An Introduction to Multivariate Statistics, 3rd ed.* 2003.

[4] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2007.

[5] T. Bollerslev, J. Russell, and M. Watson. *Volatility and Time Series Econometrics: Essays in Honor of Robert Engle*. 2010.

[6] D. Cai, X. He, and J. Han. Subspace learning based on tensor analysis. In *Technical Report UIUCDCS-R-2005-2572*, 2005.

[7] T. M. Cover. Universal portfolios. *Mathematical Finance*, 1(1):1–29, 1991.

[8] S. Dasgupta. Experiments with random projection. In *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*, pages 143–151, 2000.

[9] C. Ding and J. Ye. Two-dimensional singular value decomposition (2dsvd) for 2d maps and images. In *Proceedings of the 5th SIAM International Conference on Data Mining (SDM)*, pages 32–43, 2005.

[10] R. Engle. Autoregressive conditional heteroskedasticity with estimates of the variance of uk inflation. *Econometrica*, 50:987–1008, 1982.

[11] R. Engle. Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroscedasticity models. *Journal of Business and Economic Statistics*, 20:339–350, 2002.

[12] B. Flury. *Common Principal Components and Related Multivarate Models*. John Wiley, 1988.

[13] B. N. Flury. Common principal components in k groups. *Journal of American Statistical Association*, 79(388):892–898, 1984.

[14] K. Fukunaga. *Introduction to Statistical Pattern Recognition, 2nd edition*. Academic Press, 1990.

[15] G. H. Golub and C. V. Loan. *Matrix Computations,3rd ed.* Johns Hopkins University Press, 1996.

[16] R. A. Harshman. Foundations of the parafac procedure: Models and conditions for an .explanatory. multimodal factor analysis. *UCLA Working Papers in Phonetics*, 16:1–84, 1970.

[17] R. A. Harshman. Parafac. tutorial and applications. *Chemometrics and Intelligent Laboratory Systems*, 38(2):149–171, 1997.

[18] D. Helmbold, R. Schapire, Y. Singer, and M. Warmuth. Online portfolio setection using multiplicative weights. *Mathematical Finance*, 8(4):325–347, 1998.

[19] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 1985.

[20] E. Kofidis, Phillip, and A. Regalia. On the best rank-1 approximation of higher-order supersymmetric tensors. *SIAM Journal on Matrix Analysis and Applications*, 23(3):863–884, 2000.

[21] T. G. Kolda. Orthogonal tensor decompositions. *SIAM Journal on Matrix Analysis and Applications*, 23(1):243–255, 2001.

[22] T. G. Kolda and B. W. Bader. Tensor decompositions and applications. *SIAM Review*, 51(3):455–500, 2009.

[23] T. G. Kolda, B. W. Bader, and J. P. Kenny. Higher- order web link analysis using multilinear algebra. In *ICDM*, pages 242–249, 2005.

[24] P. M. Kroonenberg. *Applied Multiway Data Analysis*. Wiley, 2008.

[25] P. M. Kroonenberg and J. de Leeuw. Principal component analysis of three-mode data by means of alternating least squares algorithms. *Psychometrika*, 45(1):69–97, 1980.

[26] L. D. Lathauwer, B. D. Moor, J. Vandewalle, and J. V. A multilinear singular value decomposition. *SIAM Journal on Matrix Analysis and Applications*, 21(4):1253–1278, 2000.

[27] L. D. Lathauwer, B. D. Moor, J. Vandewalle, and J. V. On the best rank-1 and rank-(r1,r2,. . .,rn) approximation of higher-order tensors. *SIAM Journal on Matrix Analysis and Applications*, 21(4):1324 – 1342, 2000.

[28] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *NIPS01*, pages 556–562, 2001.

[29] J. D. Leeuw and G. Michaildis. Majorization methods in statistics. *Biostatistics*, 9(3):432–441, 2008.

[30] J. A. Patz, D. Campdell-Lendrum, T. Holloway, and J. A. Foley. Impact of regional climate change on human health. *Nature*, 438:310–317, 2005.

[31] M. Pourahmadi, M. J. Daniels, and T. Park. Simultaneous modelling of the cholesky decomposition of several covariance matrices. *J. Multivar. Anal*, 98:568–587, 2006.

[32] J. T. Scruggs and P. Glabadanidis. Risk premia and the dynamic covariance between stock and bond returns. *Journal of finance and quantitative analysis*, 38(2):295–316, 2003.

[33] J. Sun, D. Tao, and C. Faloutsos. Beyond streams and graphs: Dynamic tensor analysis. In *in Knowledge Discovery and Data Mining*, pages 374–383, 2006.

[34] S. Tadjudin and D. A. Landgrebe. Covariance estimation with limited training samples. *IEEE Transactions on Geoscience and Remote Sensing*, 37(4), July 1999.

[35] L. R. Tucker. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31:279–311, 1966.

[36] J. Ye. Generalized low rank approximations of matrices. *Machine Learning Journal*, 61:167–191, 2005.