

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

Modeling Time Varying Covariance Matrices in Low Dimensions

Huahua Wang, Arindam Banerjee, and Daniel Boley

August 04, 2010

Modeling Time Varying Covariance Matrices in Low Dimensions

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

In several modern applications, one considers high dimensional data with time varying covariance matrices. While methods such as Principal Component Analysis (PCA) are well suited for dimensionality reduction of static data, such methods were not designed to find a suitable subspace which can account for the variability in the covariance structure. In this paper, we present a new model which finds a suitable low-dimensional subspace which captures the variations of high-dimensional covariance matrices. While the problem can be posed as one of tensor decomposition, standard approaches to tensor decomposition rely on suitable initialization and may obtain poor local minima on convergence. By analyzing the structure of the problem, we establish lower and upper bounds of the global maximum in terms of a simpler problem. We use the bounds to propose an initialization and an iterative update algorithm with provable approximation guarantees with respect to the global maximum. We also establish conditions under which the method will obtain the global maxima. We illustrate the effectiveness of the proposed method through experiments on synthetic data as well as two real stock market datasets each spanning 14 years where the method finds major financial events in low dimensions.

1 Introduction

In recent years, the availability of high-dimensional temporal data ranging from finance to climate and environmental sciences is making the study of time varying covariance matrices increasingly important [21, 20, 19]. Traditionally, analysis of covariance matrices has been shown to be particularly challenging when the data is high dimensional and/or the sample size is small (see e.g., [6, 21]). The high dimensionality or small sample size often leads to covariance matrices which are not invertible. In addition, many processes depends on latent factors lying in a space with dimension much lower than the observation space. Consequently, latent covariance is not only more tractable but also more expressible than observation covariance. The most famous method for analyzing covariance matrices in low dimensional spaces is Principal Component Analysis (PCA) [7]. In PCA, a single observation covariance is first built for the entire dataset under consideration. PCA attempts to find a lower dimensional representation (latent covariance) which is close to the observation covariance in the least squares sense.

When data varies over time, the standard assumption that covariance is stationary does not apply in several domains, e.g., finance, climate sciences, etc. The covariance matrices are themselves time varying, and there is an increasing need for methods which systematically study such time varying covariance matrices. While the observed covariance matrices over time are still high-dimensional, a natural question to investigate can be posed as follows: Is there is a low-dimensional space where the time varying covariances can be suitably captured? In one can find such a space, then it is sufficient to track the dynamics of the time varying covariance matrices in the low-dimensional space resulting in both computational feasibility and interpretability of results.

Based on such a motivation, we investigate the problem of modeling time varying covariance matrices in low dimensions. We introduce a new model for finding a suitable low dimension which approximates the observed high-dimensional covariance matrices with corresponding low-dimensional covariance matrices in a well-defined least squares sense. The model can start with a pre-defined low dimension or a pre-defined approximate relative error w.r.t. the observed covariances. For a given low-dimension r , the goal is to find a r -dimensional subspace such that the projection of the observed covariance matrices into this subspace is as close as possible to the observed covariances in a least square sense. The problem can be posed as one of tensor decomposition and, in principle, recent advances in multiway data analysis [14] or tensor decomposition [12, 16, 13, 10] can be suitably leveraged. However, unlike PCA, tensor decomposition methods are not guaranteed to give globally optimal solutions. Most existing approaches start with an initialization, often chosen at random, and iteratively improves it to reach a local optimum. The quality of the final solution depends crucially on the initialization. With a careful analysis of the problem under consideration, we establish lower and upper bounds for the global maximum of our problem based on a related eigen-value decomposition problem. The analysis leads an effective initialization for our problem. Based on existing ideas in tensor decomposition, we then propose an iterative algorithm for improving the objective till convergence. The final solution will have clear approximation guarantees w.r.t. the global maximum. Instead of a low-dimension r , if a approximate relative error is specified, existing tensor decomposition algorithms do not have an effective way to choose a low dimensionality which guarantees the given approximate relative error. The reason is simply because tensor decompositions do not enjoy the incremental residual error reduction property which PCA has. By utilizing the bounds on the problem under consideration, we propose an effective way to choose a dimensionality which is guaranteed to satisfy a given approximate relative error guarantee. We also derive conditions under which the proposed algorithm is guaranteed to find the global maximum. We illustrate the effectiveness of the proposed method through experiments on synthetic data as well as two real stock market datasets each spanning 14 years where the method finds major financial events in low dimensions.

To summarize, four contributions of our work are highlighted here:

1. We propose a new model, called time varying covariance approximation 2 (TVCA2), for low dimensional approximation of time varying covariance matrices. While PCA tries to find a single latent covariance for the entire period, TVCA2 attempts to find a single subspace but a sequence of different latent covariances corresponding to different time points.
2. The computational problem for learning TVCA2 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 TVCA2, we derive lower and upper bounds for the global maximum of TVCA2.
3. Using the bounds, we propose an initialization and an iterative algorithm which is guaranteed to converge with approximation guarantees w.r.t. the global maximum. We also give sufficiency conditions under which global maximum will be achieved.
4. Instead of starting with a given low dimension, we show that the model can start with a target relative approximation error w.r.t. the global maximum and choose the dimensionality appropriately.

The remainder of this paper is organized as follows. Section 2 reviews some related work, including PCA and tensor decomposition. In section 3, to suitably model time varying high-dimensional covariance matrices, we introduce Time Varying Covariance Approximation 2 (TVCA2). In section 4, we analyze the problem and establish lower and upper bounds for the global maximum in terms of a simpler variant of the problem, called Time Varying Covariance Approximation 1 (TVCA1). Section 5 presents two algorithms for learning the model approximately based on the bounds give a low-dimensionality and given an approximate relative error respectively; we also establish conditions under which global maximum will be achieved. We report experimental results on synthetic data as well as two stock market datasets each spanning 14 years to illustrate the performance of the proposed ideas in Section 6, and conclude in Section 7.

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 Related Work

Given a set of observations, Principal Component Analysis (PCA) first constructs a total observation covariance \mathbf{X} and then aims to find the so-called principal components $\mathbf{U} \in \mathbb{R}^{n \times r}$ ($r \leq n$) so that the observation covariance $\mathbf{X} \in \mathbb{R}^{n \times n}$ is well preserved in a lower dimensional subspace in the least square sense. Mathematically, PCA solves the following maximization problem

$$\max_{\mathbf{U}^T \mathbf{U} = \mathbf{I}_r} f_0(\mathbf{U}) = \text{Tr}(\mathbf{U}^T \mathbf{X} \mathbf{U}) \quad (1)$$

which can be solved using the eigenvalue decomposition (EVD) of \mathbf{X} . The global maximum f_0^{\max} is attained when \mathbf{U} is the matrix of leading eigenvectors of \mathbf{X} , the eigenvalue matrix $\mathbf{U}^T \mathbf{X} \mathbf{U}$ is a diagonal matrix and its diagonal entries are the leading eigenvalues of \mathbf{X} . Any orthonormal matrix \mathbf{U} whose column space matches the span of the leading eigenvectors of \mathbf{X} will do, but using the eigenvectors themselves simplifies the analysis. In the EVD, if eigenvalues are distinct, eigenvectors are unique up to sign. Therefore, the solution of problem (1) is essentially unique in the EVD sense for a given chosen rank r . Letting latent covariance $\mathbf{Y} = \mathbf{U}^T \mathbf{X} \mathbf{U} \in \mathbb{R}^{r \times r}$ be the diagonal matrix of eigenvalues, the observation covariance can be well approximated by its latent covariance such that

$$\mathbf{X} = \mathbf{U} \mathbf{Y} \mathbf{U}^T + \mathbf{E} \quad (2)$$

where \mathbf{E} is the residual.

In realizing that data is essentially high dimensional, there are numerous methods proposed for higher-order tensor decomposition [14, 16, 12, 13, 18, 17]. In Kroonenberg, several classical tensor decomposition methods, such as variants of Tucker model and Parafac Model, are discussed systematically. In [16], the authors extended the classical matrix SVD to the higher-order tensor SVD, called HOSVD. In [10], Inoue et. al. shows the equivalence among HOSVD, 2DSVD, Tensor PCA and etc.. Kolda and Bader [12] gives a through review about tensor decomposition. However, all these methods lead to a local maximum. It is still unclear how good the local maximum is or how to achieve the global maximum.

3 Problem Formulation

In real word, it is highly possible that covariance matrices of interest change over time. Consider the covariance over 500 stocks in the S&P500 index. If one computes the 500×500 covariance matrix \mathbf{X}_t over each month t , it is likely that \mathbf{X}_t will change from month to month. One possible way to explain the fluctuations in \mathbf{X}_t is by treating them as random perturbations of an otherwise stationary covariance matrix \mathbf{X} . In other words, the environment is assumed to be stationary, and a simple PCA over the accumulated data should serve well to explain the observed covariance in low dimensions. However, if one looks at real historical market data, it is difficult to justify stationarity of the covariance matrices. A similar argument holds for several other domains with dynamic data. This motivates our proposed model which allows the covariance matrices to change over time, but assumes that change in covariance can mostly be explained by changes in a set of much lower dimensional covariance matrices \mathbf{Y}_t .

Assume a set of time varying 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,

$$\mathbf{X}_t = \mathbf{U} \mathbf{Y}_t \mathbf{U}^T + \mathbf{E}_t \quad (3)$$

where \mathbf{E}_t is the residual matrix. Without loss of generality, \mathbf{U} is assumed to be orthonormal, i.e., $\mathbf{U}^T \mathbf{U} = \mathbf{I}_r$. Consider that problem (3) is a mode-2 decomposition as Tucker2 model [14], our problem is called Time Varying Covariance Approximation 2 (TVCA2) in this paper.

Since the residual matrices at every time step is expected to small Frobenius norm, the problem is posed as follows:

$$\min_{\substack{\mathbf{U}, \mathbf{Y}_t \\ \mathbf{U}^T \mathbf{U} = \mathbf{I}_r}} \sum_{t=1}^T \|\mathbf{X}_t - \mathbf{U} \mathbf{Y}_t \mathbf{U}^T\|_F^2. \quad (4)$$

We start the analysis with the following result.

Lemma 1 *The optimum \mathbf{Y}_t in (4) satisfies $\mathbf{Y}_t = \mathbf{U}^T \mathbf{X}_t \mathbf{U}$. Further, the optimal \mathbf{U} in (4) is the solution to the following problem:*

$$\max_{\mathbf{U}^T \mathbf{U} = \mathbb{I}_r} f_2(\mathbf{U}) = \max_{\mathbf{U}^T \mathbf{U}} \text{Tr}(\mathbf{U}^T M(\mathbf{U}) \mathbf{U}), \quad (5)$$

where

$$M(\mathbf{U}) = \sum_{t=1}^T \mathbf{X}_t \mathbf{U} \mathbf{U}^T \mathbf{X}_t. \quad (6)$$

Proof: Since $\mathbf{U}^T \mathbf{U} = \mathbb{I}_r$, taking the derivative of objective function in (4) 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 (4), we obtain

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

where (a) holds because $\mathbf{Y}_t = \mathbf{U}^T \mathbf{X}_t \mathbf{U}$, and (b) holds since $\text{Tr}(AB) = \text{Tr}(BA)$ and $\mathbf{U}^T \mathbf{U} = \mathbb{I}_r$. Since $\text{Tr}(\sum_{t=1}^T \mathbf{X}_t^2)$ is a constant, problem (4) is equivalent to the following maximization problem

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

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 $f_2(\mathbf{U})$ in (5) is convex. For this we need a lemma

Lemma 2 $0 \leq \text{Tr}(\mathbf{A} \cdot \mathbf{B}) \leq \text{Tr}(\mathbf{A}) \cdot \text{Tr}(\mathbf{B})$ 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 $\text{Tr}(\mathbf{X} \mathbf{Y}) = \text{Tr}(\mathbf{Y} \mathbf{X})$ for any \mathbf{X}, \mathbf{Y} :

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

Hence we have

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

Lemma 3 For $\mathbf{U} \in \mathbb{R}^{n \times r}$ (not necessarily orthonormal), $f_2(\mathbf{U})$ is a convex function.

Proof: It suffices to show $f_{\mathbf{X}}(\mathbf{U}) = \text{Tr}[\mathbf{F}_{\mathbf{X}}(\mathbf{U})]$ 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

$$\mathbf{F}_{\mathbf{X}}(\mathbf{U} + s\mathbf{V}) = \mathbf{F}_{\mathbf{X}}(\mathbf{U}) + s\mathbf{G} + s^2\mathbf{H} + h.o.t., \quad (7)$$

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 $\text{Tr}(\mathbf{H}) \geq 0$. Expanding (7) yields the following expression for \mathbf{H} :

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

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. \blacksquare

Unfortunately, the fact that $f_2(\mathbf{U})$ is convex does not help us in any way. Note that from (5), the problem is one of *maximizing* $f_2(\mathbf{U})$ 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 (5) 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 an algorithm for iterative updates with guarantees relative to the global optimum.

4 Analysis of Time Varying Covariance Approximation 2 (TVCA2)

In this section, we analyze TVCA2 in terms of a simpler model called Time Varying Covariance Approximation 1 (TVCA1). We show that TVCA1 can be solved using a suitable eigen-value decomposition (EVD). More importantly, the solution to TVCA1 leads to lower and upper bounds of the global maximum of TVCA2, and suggests a good initialization for any iterative algorithm for solving TVCA2. Further, the analysis shows how one can start with a give upper bound on the approximate relative error (ARE) for TVCA2 and obtain a \mathbf{U} of suitable dimensionality to satisfy the bound.

4.1 A Simpler Model: TVCA1

Instead of the original problem in (4), we consider a simpler decomposition given by

$$\mathbf{X}_t = \mathbf{U} \mathbf{Y}_t + \mathbf{E}_t \quad (8)$$

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}, \mathbf{Y}_t can be posed as follows:

$$\min_{\substack{\mathbf{U}, \mathbf{Y}_t \\ \mathbf{U}^T \mathbf{U} = \mathbb{I}_r}} \sum_{t=1}^T \|\mathbf{X}_t - \mathbf{U} \mathbf{Y}_t\|_F^2. \quad (9)$$

We call the above problem TVCA1 since it corresponds to a mode-1 decomposition as in Tucker1 model[14]. Note that TVCA2 corresponds to a mode-2 decomposition as in Tucker2 model. As in the original problem, the simplified problem TVCA1 allows an alternative characterization as follows:

Table 1: TVCA2 and TVCA1

TVCA2	TVCA1
$\mathbf{X}_t = \mathbf{U}\mathbf{Y}_t\mathbf{U} + \mathbf{E}_t$	$\mathbf{X}_t = \mathbf{U}\mathbf{Y}_t + \mathbf{E}_t$
$M(\mathbf{U}) = \sum_t \mathbf{X}_t \mathbf{U} \mathbf{U}^T \mathbf{X}_t$	$M(\mathbb{I}_n) = \sum_t \mathbf{X}_t^2$
$f_2(\mathbf{U}) = \text{Tr}(\mathbf{U}^T M(\mathbf{U}) \mathbf{U})$	$f_1(\mathbf{U}) = \text{Tr}(\mathbf{U}^T M(\mathbb{I}_n) \mathbf{U})$

Lemma 4 The optimal \mathbf{Y}_t in (9) satisfies $\mathbf{Y}_t = \mathbf{U}^T \mathbf{X}_t$. Further, the optimal \mathbf{U} in (8) is the solution to the following problem:

$$\max_{\mathbf{U}^T \mathbf{U} = \mathbb{I}_r} f_1(\mathbf{U}) = \max_{\mathbf{U}^T \mathbf{U} = \mathbb{I}_r} \text{Tr}(\mathbf{U}^T M(\mathbb{I}_n) \mathbf{U}), \quad (10)$$

where

$$M(\mathbb{I}_n) = \sum_{t=1}^T \mathbf{X}_t^2. \quad (11)$$

Proof: Since $\mathbf{U}^T \mathbf{U} = \mathbb{I}_r$, taking derivative of (8) 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 (8), we obtain

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

where (a) holds since $\text{Tr}(AB) = \text{Tr}(BA)$ and $\mathbf{U}^T \mathbf{U} = \mathbb{I}_r$. Since $\text{Tr}(\sum_{t=1}^T \mathbf{X}_t^2)$ is a constant, problem (8) is equivalent to the following maximization problem

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

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

First note that TVCA1 as in (10) is exactly the PCA problem, which is much easier to solve than TVCA2. Table 1 shows a relative comparison between TVCA1 and TVCA2.

4.2 Lower and Upper Bounds

The solution of TVCA1 helps significantly in characterizing the solution to TVCA2. We focus on developing lower and upper bounds to optimum value of TVCA2 based on the solution of TVCA1. Since TVCA1 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 (10). Let $f_1^{\max} = f_1(\mathbf{U}_0)$ be the maximum value of $f_1(\mathbf{U})$. Further, let $M_T = \text{Tr}(M(\mathbb{I}_n)) = \text{Tr}(\sum_t \mathbf{X}_t^2)$. With this notation, we have the following result:

Theorem 1 Let $M_T = \text{Tr}(\sum_t \mathbf{X}_t^2)$. Then, with $f_1(\mathbf{U})$ and $f_2(\mathbf{U})$ denoting the objective functions for TVCA1 and TVCA2 respectively as in (10) and (5), for any \mathbf{U} with $\mathbf{U}^T \mathbf{U} = \mathbb{I}_r$, we have

$$\frac{f_1^2(\mathbf{U})}{M_T} \leq f_2(\mathbf{U}) \leq f_1(\mathbf{U}). \quad (12)$$

Proof: By definition,

$$f_2(\mathbf{U}) = \text{Tr}(\mathbf{U}^T M(\mathbf{U}) \mathbf{U}) \leq \text{Tr}(M(\mathbf{U})) = \sum_{t=1}^T \text{Tr}(\mathbf{X}_t \mathbf{U} \mathbf{U}^T \mathbf{X}_t) = \sum_{t=1}^T \text{Tr}(\mathbf{U}^T \mathbf{X}_t^2 \mathbf{U}) = f_1(\mathbf{U}).$$

Now, we prove $f_2(\mathbf{U}) \geq \frac{f_1^2(\mathbf{U})}{M_T}$. 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:

$$\mathbf{A} = [\mathbf{X}_1^{\frac{1}{2}} \mathbf{U} \mathbf{U}^T \mathbf{X}_1^{\frac{1}{2}}, \dots, \mathbf{X}_T^{\frac{1}{2}} \mathbf{U} \mathbf{U}^T \mathbf{X}_T^{\frac{1}{2}}] \quad \mathbf{B} = [\mathbf{X}_1, \dots, \mathbf{X}_T].$$

The trace of their product is given by

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

Now, $f_2(\mathbf{U})$ is rewritten as

$$\begin{aligned} f_2(\mathbf{U}) &= \sum_{t=1}^T \text{Tr}(\mathbf{U}^T \mathbf{X}_t \mathbf{U} \mathbf{U}^T \mathbf{X}_t \mathbf{U}) \\ &= \sum_{t=1}^T \text{Tr}(\mathbf{U}^T \mathbf{X}_t^{\frac{1}{2}} \mathbf{X}_t^{\frac{1}{2}} \mathbf{U} \mathbf{U}^T \mathbf{X}_t^{\frac{1}{2}} \mathbf{X}_t^{\frac{1}{2}} \mathbf{U}) \\ &= \sum_{t=1}^T \text{Tr}(\mathbf{X}_t^{\frac{1}{2}} \mathbf{U} \mathbf{U}^T \mathbf{X}_t^{\frac{1}{2}} \mathbf{X}_t^{\frac{1}{2}} \mathbf{U} \mathbf{U}^T \mathbf{X}_t^{\frac{1}{2}}) \\ &= \text{Tr}(\mathbf{A} \mathbf{A}^T), \end{aligned}$$

and M_T is

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

From the Cauchy-Schwarz inequality, we have

$$f_2(\mathbf{U}) M_T = \text{Tr}(\mathbf{A} \mathbf{A}^T) \text{Tr}(\mathbf{B} \mathbf{B}^T) \geq [\text{Tr}(\mathbf{A} \mathbf{B}^T)]^2 = f_1^2(\mathbf{U}).$$

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,

$$p_1 = \frac{f_1^{\max}}{M_T} = \frac{\text{Tr}(\mathbf{U}_0^T (\sum_t \mathbf{X}_t^2) \mathbf{U}_0)}{\text{Tr}(\sum_t \mathbf{X}_t^2)}, \quad (13)$$

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 TVCA2.

Corollary 1 Let f_1^{\max} and f_2^{\max} be the global maximum of TVCA1 and TVCA2 respectively over $\mathbf{U}^T \mathbf{U} = \mathbb{I}_r$, and p_1 is as defined in Definition 1. Then, we have

$$p_1 f_1^{\max} \leq f_2^{\max} \leq f_1^{\max} \quad (14)$$

Proof: Let \mathbf{U}_0 be the solution of TVCA1, so that $f_1^{\max} = f_1(\mathbf{U}_0)$ and $p_1 = f_1^{\max}/M_T$. According to Theorem 1, we have

$$f_2(\mathbf{U}_0) \geq \frac{f_1^2(\mathbf{U}_0)}{M_T} = \frac{f_1^{\max}}{M_T} f_1^{\max} = p_1 f_1^{\max}.$$

Hence, for the global maximum of TVCA2, we have

$$f_2^{\max} \geq f_2(\mathbf{U}_0) \geq p_1 f_1^{\max}$$

Further, since $f_1(\mathbf{U})$ is an upper bound of $f_2(\mathbf{U})$, we have $f_2^{\max} \leq f_1^{\max}$. That completes the proof. \blacksquare

Recall that the solution to TVCA1 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 $p_1 f_1^{\max} \leq f_2(\mathbf{U}_0) \leq f_1^{\max}$. Now if we do iterative updates for $f_2(\mathbf{U})$ which start with initialization \mathbf{U}_0 and converges to \mathbf{U}_0^* , we have

$$p_1 f_1^{\max} \leq f_2(\mathbf{U}_0) \leq f_2(\mathbf{U}_0^*) \leq f_2^{\max} \leq f_1^{\max}. \quad (15)$$

From (15), we note that if p_1 is close to 1, then $f_2(\mathbf{U}_0^*)$ will be close to the global maximum f_2^{\max} . The property is 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 $f_2(\mathbf{U}_0^*)$ be the solution to TVCA2 with the initialization \mathbf{U}_0 . Then, the relative error of $f_2(\mathbf{U}_0^*)$ with respect to f_2^{\max} satisfies

$$\left| \frac{f_2^{\max} - f_2(\mathbf{U}_0^*)}{f_2^{\max}} \right| \leq 1 - p_1 \quad (16)$$

Proof: Consider the inequality $f_2(\mathbf{U}_0^*) \geq p_1 f_1^{\max}$. Dividing both sides by f_2^{\max} we get

$$\frac{f_2(\mathbf{U}_0^*)}{f_2^{\max}} \geq p_1 \frac{f_1^{\max}}{f_2^{\max}} \stackrel{(a)}{\geq} p_1,$$

where (a) follows since $f_1^{\max} \geq f_2^{\max}$. Consequently

$$\left| \frac{f_2^{\max} - f_2(\mathbf{U}_0^*)}{f_2^{\max}} \right| = 1 - \frac{f_2(\mathbf{U}_0^*)}{f_2^{\max}} \leq 1 - p_1. \quad \blacksquare$$

Note that initialization itself satisfies the above bound, so that $f_2(\mathbf{U}_0) \geq p_1 f_2^{\max}$. 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 $f_2(\mathbf{U})$. 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.

4.3 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 TVCA2 and more generally in case of tensor decompositions. Thus, the best rank $(r - 1)$ solution to TVCA2 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 TVCA1 problem.

We start with defining *Approximate Relative Error* (ARE) as a measure of how good the approximation obtained by TVCA2 is. For any \mathbf{U} , we have

$$ARE = \frac{\sum_{t=1}^T \|\mathbf{X}_t - \mathbf{U}\mathbf{Y}_t\mathbf{U}^T\|_F^2}{\sum_{t=1}^T \|\mathbf{X}_t\|_F^2}, ARE(\mathbf{U}) \quad (17)$$

We define the cumulative percentage of energy captured by the solution to TVCA2 as follows:

Definition 2 Let $M_T = \text{Tr}(M(\mathbb{I}))$, and let $f_2(\mathbf{U}_0^*)$ be the maximum of TVCA2 with initialization \mathbf{U}_0 . The cumulative percent of energy p_2 captured by \mathbf{U}_0^* is defined as the

$$p_2 = \frac{f_2(\mathbf{U}_0^*)}{M_T} \quad (18)$$

where $0 \leq p_2 \leq 1$.

For our problem, p_2 defines how much energy of time varying covariances is preserved by their corresponding latent covariances. Dividing by M_T on both sides of inequality (15) and plugging in $p_1 = f_1^{\max}/M_T$, the lower and upper bounds of p_2 are

$$p_1^2 \leq p_2 \leq p_1 \quad (19)$$

Recall that p_1 is defined in the PCA setting. In TVCA1, given a p_1 , the corresponding rank r is easy to obtain. Using the bounds for p_2 , one can also develop a simple way of obtaining a suitable rank- r for TVCA2. To do this, we first establish a relationship between p_2 and approximate relative error $ARE(\mathbf{U}_0^*)$.

Proposition 1 Let \mathbf{U}_0^* be the solution of TVCA2. Then $ARE(\mathbf{U}_0^*) = 1 - p_2$.

Proof: From the proof of Lemma 1, we have

$$\sum_{t=1}^T \|\mathbf{X}_t - \mathbf{U}\mathbf{Y}_t\mathbf{U}^T\|_F^2 = \text{Tr} \left(\sum_{t=1}^T \mathbf{X}_t^2 \right) - \text{Tr} \left(\sum_{t=1}^T \mathbf{U}^T \mathbf{X}_t \mathbf{U} \mathbf{U}^T \mathbf{X}_t \mathbf{U} \right) = M_T - f_2(\mathbf{U}).$$

Let \mathbf{U}_0^* be the solution of TVCA2. Then

$$ARE(\mathbf{U}_0^*) = \frac{M_T - f_2(\mathbf{U}_0^*)}{M_T} = 1 - p_2. \quad \blacksquare$$

Plugging $ARE(\mathbf{U}_0^*)$ into inequality (19), it is easy to derive the following lower and upper bounds for $ARE(\mathbf{U}_0^*)$:

$$1 - p_1 \leq ARE(\mathbf{U}_0^*) \leq 1 - p_1^2 \quad (20)$$

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 TVCA2. 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 TVCA2 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.

5 Algorithms

In this section, we present two algorithms for solving TVCA2. Algorithm 1 is applicable to the case when a fixed dimensionality r of \mathbf{U} is given. Algorithm 2 is applicable when an upper bound to the approximate relative error (ARE) is given, and the goal is to find \mathbf{U} of suitable dimensionality which satisfies the given ARE bound.

5.1 TVCA2 Algorithm For A Given Dimensionality

If a dimensionality r is given, EVD can be used to solve for U in the PCA problem as in (1) as well as TVCA1 as in (10). However, TVCA2 in (5) has four 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 TVCA2 involves maximizing $f_2(\mathbf{U}) = \text{Tr}(\mathbf{U}^T M(\mathbf{U})\mathbf{U})$ where $M(\mathbf{U}) = \sum_{t=1}^T \mathbf{X}_t \mathbf{U} \mathbf{U}^T \mathbf{X}_t$. 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} :

$$\max_{\mathbf{U}^T \mathbf{U} = \mathbf{I}_r} \text{Tr}(\mathbf{U}^T M(\mathbf{U}_k)\mathbf{U}). \quad (21)$$

Clearly, \mathbf{U}_{k+1} can be obtained by applying rank- r EVD on $M(\mathbf{U}_k)$. As the following result shows, such an update will improve the objective function, i.e., $f_2(\mathbf{U}_{k+1}) \geq f_2(\mathbf{U}_k)$.

Theorem 2 *Let \mathbf{U}_{k+1} be the r principal eigenvectors of $M(\mathbf{U}_k)$, then $f_2(\mathbf{U}_{k+1}) \geq f_2(\mathbf{U}_k)$. 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

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

By definition, we have

$$\begin{aligned} \text{Tr}(\mathbf{A}_k \mathbf{A}_k^T) &= \text{Tr} \left(\sum_{t=1}^T \mathbf{X}_t^{\frac{1}{2}} \mathbf{U}_k \mathbf{U}_k^T \mathbf{X}_t^{\frac{1}{2}} \mathbf{X}_t^{\frac{1}{2}} \mathbf{U}_k \mathbf{U}_k^T \mathbf{X}_t^{\frac{1}{2}} \right) = \text{Tr} \left(\sum_{t=1}^T \mathbf{U}_k^T \mathbf{X}_t \mathbf{U}_k \mathbf{U}_k^T \mathbf{X}_t \mathbf{U}_k \right) \\ &= f_2(\mathbf{U}_k) \end{aligned}$$

Let \mathbf{U}_{k+1} be the r principal eigenvectors of $M(\mathbf{U}_k)$. By a similar analysis, $f_2(\mathbf{U}_{k+1}) = \text{Tr}(\mathbf{A}_{k+1} \mathbf{A}_{k+1}^T)$. Now note that

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

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

$$\text{Tr}(\mathbf{U}_{k+1}^T M(\mathbf{U}_k) \mathbf{U}_{k+1}) \geq \text{Tr}(\mathbf{U}_k^T M(\mathbf{U}_k) \mathbf{U}_k) = f_2(\mathbf{U}_k), \quad (22)$$

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

Then, we have

$$\begin{aligned} f_2(\mathbf{U}_k) f_2(\mathbf{U}_{k+1}) &= \text{Tr}(\mathbf{A}_k \mathbf{A}_k^T) \text{Tr}(\mathbf{A}_{k+1} \mathbf{A}_{k+1}^T) \\ &\stackrel{(a)}{\geq} [\text{Tr}(\mathbf{A}_k \mathbf{A}_{k+1}^T)]^2 \\ &= [\text{Tr}(\mathbf{U}_{k+1}^T M(\mathbf{U}_k) \mathbf{U}_{k+1})]^2 \\ &\stackrel{(b)}{\geq} f_2^2(\mathbf{U}_k), \end{aligned}$$

Algorithm 1 TVCA2 Algorithm for a given dimensionality r

Input: $\mathbf{X}_t, 1 \leq t \leq T$, dimensionality r
Output: $\mathbf{U}, \mathbf{Y}_t, 1 \leq t \leq T$
{Initialization}
Perform EVD on $M(\mathbb{I}) = \sum_t \mathbf{X}_t^2$ and choose the leading r eigenvectors \mathbf{U}_0
{Iteration}
repeat
 Perform EVD on $M(\mathbf{U}_k) = \sum_t \mathbf{X}_t \mathbf{U}_k \mathbf{U}_k^T \mathbf{X}_t$
 Choose the leading r eigenvectors \mathbf{U}_{k+1}
until $\left| \frac{f_2(\mathbf{U}_{k+1}) - f_2(\mathbf{U}_k)}{f_2(\mathbf{U}_k)} \right| \leq \varepsilon$
 Perform the SVD on $\mathbf{U}_{k+1} M(\mathbf{U}_{k+1}) \mathbf{U}_{k+1}$ to get P , then $\mathbf{U} = \mathbf{U}_{k+1} P$
 Compute $\mathbf{Y}_t = \mathbf{U}^T \mathbf{X}_t \mathbf{U}$

where (a) follows from Lemma 2 and (b) follows from (22). If \mathbf{U}_{k+1} and \mathbf{U}_k span the same subspace, the equality in (b) holds, and $\mathbf{A}_k = \mathbf{A}_{k+1}$ by definition. As a consequence, equality in Cauchy-Schwarz inequality (a) also holds.

Since $f_2(\mathbf{U}_k)$ and $f_2(\mathbf{U}_{k+1})$ are nonnegative, therefore

$$f_2(\mathbf{U}_{k+1}) \geq f_2(\mathbf{U}_k)$$

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

By performing the EVD iteratively, the objective function increases every step until a certain stopping criterion is satisfied such that

$$\left| \frac{f_2(\mathbf{U}_{k+1}) - f_2(\mathbf{U}_k)}{f_2(\mathbf{U}_k)} \right| \leq \varepsilon \quad (23)$$

where ε is a small constant. In the ideal case, when \mathbf{U}_k and \mathbf{U}_{k+1} span the same subspace, the stopping criterion value will be 0, but we can still have $\mathbf{U}_{k+1} \neq \mathbf{U}_k$. To ensure $\mathbf{U}_{k+1} = \mathbf{U}_k$, a more strict stopping criterion to consider would be

$$\frac{\|\mathbf{U}_{k+1} - \mathbf{U}_k\|_F}{\|\mathbf{U}_k\|_F} \leq \varepsilon \quad (24)$$

If $\mathbf{U}_{k+1} = \mathbf{U}_k$, then $f_2(\mathbf{U}_{k+1}) = f_2(\mathbf{U}_k)$. However, ensuring (24) through iterative updates is extremely time consuming in practice, so we will use (23) in the experiments.

Let \mathbf{U}_k denote the basis for the subspace obtained after the algorithm converges. The objective function $f_2(\mathbf{U}_k)$ only depends on the subspace determined by \mathbf{U}_k . A different run of the algorithm may lead to a different \mathbf{U}'_k which spans the same space, but is different from \mathbf{U}_k . Such differences can be inconvenient for interpretation and development of subsequent applications based on the solution. To avoid such issues, in the algorithm, we output a unique \mathbf{U} corresponding to each subspace. Thus, if \mathbf{U}_k and \mathbf{U}'_k are different but determine the same subspace, the output \mathbf{U} of the algorithm will be the same in both cases.

For any solution \mathbf{U}_k , let the EVD of $\mathbf{U}_k M(\mathbf{U}_k) \mathbf{U}_k$ be \mathbf{PDP}^T . Let $\mathbf{U} = \mathbf{U}_k \mathbf{P}$, then $M(\mathbf{U}) = M(\mathbf{U}_k)$, so

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

We use \mathbf{U} as the unique basis for the subspace determined by \mathbf{U}_k . The following result shows that \mathbf{U} is the r principal eigenvectors of $M(\mathbf{U})$.

Theorem 3 *Let \mathbf{U} be a basis of the subspace determined by the solution \mathbf{U}_k . If*

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

where \mathbf{D} is a diagonal matrix, then \mathbf{U} are the r principal eigenvectors of $M(\mathbf{U})$ corresponding to r leading nonzero eigenvalue matrix \mathbf{D} .

Proof: For any basis \mathbf{V} of the subspace determined by the solution \mathbf{U}_k , both $A = M(\mathbf{V})$ and $f_2(\mathbf{V})$ remain unchanged. Let the SVD of $\mathbf{V}^T \mathbf{A} \mathbf{V}$ be $\mathbf{P}^T \mathbf{D} \mathbf{P}$ and $\mathbf{U} = \mathbf{V} \mathbf{P}$, so $\mathbf{U}^T \mathbf{A} \mathbf{U} = \mathbf{D}$, where $A = M(\mathbf{U})$. Let $\hat{\mathbf{U}}$ be the r principal eigenvectors of \mathbf{A} associated with nonzero eigenvalues $\hat{\mathbf{D}}$, then $\hat{\mathbf{U}}^T \mathbf{A} \hat{\mathbf{U}} = \hat{\mathbf{D}}$. Since the EVD is unique for the same subspace, $\mathbf{U} = \hat{\mathbf{U}}$ and $\mathbf{D} = \hat{\mathbf{D}}$. ■

Algorithm 1 presents the TVCA2 algorithm for a given dimensionality r as input. From Theorem 3, the iterative updates converge to \mathbf{U}_k and the subspace is represented with basis \mathbf{U} . Note that such a representation is unique in the EVD sense, and thus $\mathbf{Y}_t = \mathbf{U} \mathbf{X}_t \mathbf{U}$ is unique.

The following result shows that Algorithm 1 converges to a local maximum with guarantees on the approximate relative error with respect to the global maximum.

Proposition 2 *Let \mathbf{U}_0 be the initialization, i.e., the r primary eigenvectors of $M(\mathbb{I}) = \sum_t \mathbf{X}_t^2$, and let $p_1 = \frac{\text{Tr}(\mathbf{U}_0^T M(\mathbb{I}) \mathbf{U}_0)}{\text{Tr}(M(\mathbb{I}))}$ be the fraction of energy captured by \mathbf{U}_0 . Then, the solution \mathbf{U}^* obtained by Algorithm 1 has the following properties:*

1. \mathbf{U}^* is a local maximum such that $f_2(\mathbf{U}^*) \geq p_1 f_2^{\max}$, where f_2^{\max} is the global maximum of TVCA2;
2. The approximate relative error (ARE) of \mathbf{U}^* w.r.t. the global maximum satisfies $1 - p_1 \leq \text{ARE}(\mathbf{U}^*) \leq 1 - p_1^2$.

Proof: (1) From Theorem 2, the objective function increases iteratively at each step till subsequent iterates \mathbf{U}_k and \mathbf{U}_{k+1} span the same space. Thus, the final solution is a local maximum. According to (15), $f_2(\mathbf{U}_0) \geq p_1 f_1^{\max}$. Since $f_1^{\max} \geq f_2^{\max}$ and $f_2(\mathbf{U}^*) \geq f_2(\mathbf{U}_0)$, we have $f_2(\mathbf{U}^*) \geq p_1 f_2^{\max}$.

(2) The bounds on ARE follow directly from (20). ■

5.2 TVCA2 Algorithm For A Given ARE Upper Bound

We consider a setting where the dimensionality r is not given, but an upper bound δ on the approximate relative error (ARE) w.r.t. the global maximum is given. Since r is not given, Algorithm 1 cannot be directly used. In *Algorithm 2*, we present an algorithm which takes a given ARE bound δ as input and chooses the dimensionality r of the initialization \mathbf{U}_0 such that the bound will be satisfied. In particular, it is sufficient to choose the dimensionality r of the initialization \mathbf{U}_0 such that the fraction of energy captured in the context of TVCA1 given by $p_1 = \frac{\text{Tr}(\mathbf{U}_0^T M(\mathbb{I}) \mathbf{U}_0)}{\text{Tr}(M(\mathbb{I}))}$ satisfies $p_1 \geq \sqrt{1 - \delta}$, as discussed in Section 4. Since $M(\mathbb{I})$ is fixed, and TVCA1 is an EVD problem, choosing a suitable dimensionality r such that $p_1 \geq \sqrt{1 - \delta}$ is straightforward.

The following result establishes the properties of Algorithm 2.

Proposition 3 *Let \mathbf{U}_0 be the initialization and \mathbf{U}^* be the final solution from Algorithm 2. The final solution has the following properties:*

1. \mathbf{U}^* is a local maximum such that $f_2(\mathbf{U}^*) \geq \sqrt{1 - \delta} f_2^{\max}$, where f_2^{\max} is the global maximum of TVCA2;
2. The approximate relative error ARE of \mathbf{U}^* satisfies the given upper bound on ARE, i.e. $\text{ARE}(\mathbf{U}^*) \leq \delta$.

Proof: (1) We have already shown that the iterative EVD converges to a local maximum. According to (15), $f_2(\mathbf{U}^*) \geq p_1 f_1^{\max}$. Since $p_1 \geq \sqrt{1 - \delta}$ and $f_1^{\max} \geq f_2^{\max}$, we have $f_2(\mathbf{U}^*) \geq \sqrt{1 - \delta} f_2^{\max}$.

(2) From (20) and the fact that $p_1 \geq \sqrt{1 - \delta}$, we have $\text{ARE}(\mathbf{U}^*) \leq \delta$. ■

Remark. The iterative steps in both algorithms were also used in the *Tucker ALS* algorithm [14, 15]. In particular, [15] shows that such an iterative EVD converges to a local maximum. The proposed initialization, called high-order

Algorithm 2 TVCA2 Algorithm for a give ARE upper bound δ

Input: $\mathbf{X}_t, 1 \leq t \leq T$, ARE upper bound δ Output: \mathbf{U}, \mathbf{Y}_t and p_2

{Initialization}

Perform EVD on $M(\mathbb{I}) = \sum_t \mathbf{X}_t^2$ and choose the leading r eigenvectors \mathbf{U}_0 such that $\text{Tr}(\mathbf{U}_0^T M(\mathbb{I}) \mathbf{U}_0) \geq \sqrt{1 - \delta} \text{Tr}(M(\mathbb{I}))$

{Iteration}

repeat Perform EVD on $M(\mathbf{U}_k) = \sum_t \mathbf{X}_t \mathbf{U}_k \mathbf{U}_k^T \mathbf{X}_t$ Choose the leading r eigenvectors \mathbf{U}_{k+1} **until** $\left| \frac{f_2(\mathbf{U}_{k+1}) - f_2(\mathbf{U}_k)}{f_2(\mathbf{U}_k)} \right| \leq \varepsilon$ Perform the SVD on $\mathbf{U}_{k+1} M(\mathbf{U}_{k+1}) \mathbf{U}_{k+1}$ to get P , then $\mathbf{U} = \mathbf{U}_{k+1} P$ Compute $\mathbf{Y}_t = \mathbf{U}^T \mathbf{X}_t \mathbf{U}$

SVD initialization in tensor decomposition, have also been widely used in practice. The fact that such an initialization often leads to a global maximum has also been observed, particularly in the rank-1 approximation experiments [17, 11]. However, to the best of our knowledge, there is no existing theoretical analysis to explain these observations. This paper is the first to establish the lower and upper bounds which explains the good performance of such algorithms. Further, this paper also shows how to choose the dimensionality r according to a given upper bound on the approximate relative error.

5.3 Conditions for Global Maximum

We now analyze a condition under which a global maximum of TVCA2 is achieved. The particular case under consideration is when equality holds in (15), i.e., $f_2(\mathbf{U}_0^*) = f_1^{\max}$, where \mathbf{U}_0^* is the maximum found in Algorithm 1, implying $f_2(\mathbf{U}_0^*) = f_2^{\max}$. We need the following result for the analysis.

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

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

Proof: Let $\mathbf{A} = \mathbf{Q} \mathbf{D} \mathbf{Q}^T$ be the eigendecomposition of \mathbf{A} , with $\mathbf{D} = \text{diag}(\mathbf{D}_1, 0)$, 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 $(\sum_k \mathbf{A}_k) \mathbf{v} = 0$ if and only if $(\sum_k \mathbf{A}_k^2) \mathbf{v} = 0$ 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 4 Let \mathbf{U} be the r principal eigenvectors of $M(\mathbb{I}_n)$ associated with nonzero eigenvalues, then $\text{rank}(M(\mathbf{U}_0)) \geq r$.

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

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

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

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

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 $(\sum_t \mathbf{X}_t^2)$.

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

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

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

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 4, we now show that $\text{rank}(M(\mathbf{U}_0)) = r$ is the necessary and sufficient condition that $f_2(\mathbf{U}_0^*) = f_1^{\max}$, 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 5 *Let \mathbf{U}_0 be the solution to TVCA1, 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, $\text{rank}(M(\mathbf{U}_0)) = r$ is the necessary and sufficient condition that $f_2(\mathbf{U}_0^*) = f_1^{\max}$. Moreover, \mathbf{U}_0 is the solution achieving the global maximum for TVCA2.*

Proof: Let \mathbf{U}_0 be the solution of TVCA1, $f_1^{\max} = f_1(\mathbf{U}_0)$. Provided that $\text{rank}(M(\mathbf{U}_0)) = r$, 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 . Then

$$f_2(\mathbf{U}_1) = \text{Tr}(\mathbf{D}_1) = \text{Tr}(M(\mathbf{U}_0)) = f_1(\mathbf{U}_0) = f_1^{\max}$$

On the other hand, we have

$$f_1^{\max} \geq f_1(\mathbf{U}_1) \geq f_2(\mathbf{U}_1) = f_1^{\max}$$

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 $f_2(\mathbf{U}_0) = f_2(\mathbf{U}_1) = f_1^{\max}$. Since $f_2^{\max} \leq f_1^{\max}$, $f_2(\mathbf{U}_0) = f_1^{\max}$ clearly implies $f_2(\mathbf{U}_0) = f_2^{\max}$. \mathbf{U}_0 is the solution achieving the global maximum.

We now prove the converse, i.e., $f_2(\mathbf{U}_0^*) = f_1^{\max} \Rightarrow \text{rank}(M(\mathbf{U}_0)) = r$. Since $f_2(\mathbf{U}_0^*) = f_1^{\max}$ holds, and since f_1 is the upper bound of f_2 , we have

$$f_1^{\max} = f_2(\mathbf{U}_0^*) \leq f_1(\mathbf{U}_0^*) \leq f_1^{\max} = f_1(\mathbf{U}_0)$$

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

$$f_2^{\max} \geq f_2(\mathbf{U}_0^*) = f_2(\mathbf{U}_0) = f_1^{\max} \geq f_2^{\max},$$

implying \mathbf{U}_0 achieves the global maximum of TVCA2.

Recall that \mathbf{U}_0 are the principal eigenvectors of $M(\mathbf{U}_0)$ corresponding to nonzero eigenvalues, and $\text{rank}(M(\mathbf{U}_0)) \geq r$ according to Theorem 4. If $\text{rank}(M(\mathbf{U}_0)) > r$, there are more than r nonzero eigenvalues. Let \mathbf{U}_1 be the r principal eigenvectors of $M(\mathbf{U}_0)$. Then

$$\text{Tr}(M(\mathbf{U}_0)) > \text{Tr}(\mathbf{U}_1^T M(\mathbf{U}_0) \mathbf{U}_1) \geq \text{Tr}(\mathbf{U}_0^T M(\mathbf{U}_0) \mathbf{U}_0) = f_2(\mathbf{U}_0),$$

since \mathbf{U}_1 are the principal eigenvectors. However, $f_1^{\max} = f_1(\mathbf{U}_0) = \text{Tr}(M(\mathbf{U}_0))$. Consequently, $f_2(\mathbf{U}_0) < f_1^{\max}$, which contradicts the fact that $f_2(\mathbf{U}_0) = f_1^{\max}$. Thus, $\text{rank}(M(\mathbf{U}_0)) = r$. ■

A special case of the result is when $\text{rank}(M(\mathbf{I})) = r$. When $\text{rank}(M(\mathbf{I})) = r$, $\text{rank}(M(\mathbf{U}_0)) \leq \text{rank}(M(\mathbf{I})) = r$. According to Theorem 4, $\text{rank}(M(\mathbf{U}_0)) \geq r$, implying $\text{rank}(M(\mathbf{U}_0)) = r$. 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 4.

6 Experimental Results

In this section, the performance of TVCA2 is evaluated on both artificial datasets and two real-world stock market datasets in terms of the *Approximate Relative Error* (ARE) defined in (17). The performance of TVCA2 is compared with PCA and Random Projection (RP) [5, 1]. PCA is computed based on the aggregated covariance over the entire time period. 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 [8] and normalization. For both artificial and real datasets, we perform experiments on the training data, from where \mathbf{U} is estimated, and on test data, where the estimated \mathbf{U} is used to evaluate approximation performance. Since there is no learning involved in RP, it was evaluated only on the training data. We also report some results comparing the lower and upper bounds with the true performance. We also report results on the stock market datasets based on Algorithm 2, where a dimensionality is selected based on a prescribed threshold on the fraction of covariance explained by the low dimensional representation. Finally, we take a closer look at some specific simple examples in low dimensions where the objective functions for TVCA2 and its lower and upper bounds based on TVCA1 can be plotted and the path of the iterative solution can be traced. We give examples of scenarios where the iterations converge to the global maxima as well as local maxima.

6.1 Artificial Data

Artificial data was generated following the model in (3). 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 a Gaussian distribution with mean is $\mu = [0, 0]$ and covariance $\Sigma = \mathbf{V}^T \mathbf{D} \mathbf{V}$, where $\mathbf{V} \in \mathbb{R}^{2 \times 2} = QR(\text{randn}(2, 2))$ is a randomly generated orthonormal matrix and diagonal matrix $\mathbf{D} \in \mathbb{R}^{2 \times 2}$ has two positive diagonal entries which were also randomly generated. QR represents QR factorization, and $\text{randn}(2, 2)$ is a 2×2 matrix whose entries are i.i.d. from a normal distribution. However, instead of using a fixed \mathbf{U} over time, it was mildly perturbed as follows:

$$\mathbf{U}_{t+1} \leftarrow QR(\mathbf{U}_t + \gamma \times \text{randn}(n, r)) \quad (25)$$

where γ is a small constant. In (25), the initial \mathbf{U} is randomly generated.

Methodology: The experiment was done as follows: With $T = 20$, $n = 20$, $r = 2$, and $\tau = 5$, we generated $(T + \tau)$ sets of matrices \mathbf{U}_t , \mathbf{Y}_t and \mathbf{X}_t . The first T observation covariances \mathbf{X}_t , $1 \leq t \leq T$ were used as training data and the last τ steps were used as the test set. TVCA2 and PCA were applied on the training set to estimate \mathbf{U} . The estimated \mathbf{U} from TVCA2 and PCA were evaluated on the testing set in terms of ARE. The experiment was repeated 50 times, and the final results reported are the average over the 50 runs.

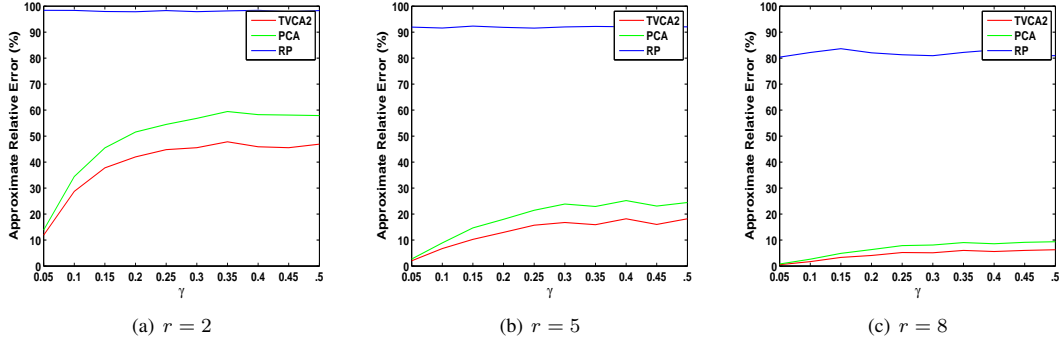


Figure 1: Approximation Relative Error (ARE) on artificial data in different dimensions r and increasing noise level γ . TVCA2 outperforms PCA and RP, especially in low dimensions and high noise levels.

Table 2: ARE(%)(std) on test set of artificial data when $r = 2$

Step	$\gamma = 0.05$		$\gamma = 0.1$	
	TVCA2	PCA	TVCA2	PCA
1	15.01(6.92)	17.19(6.00)	37.85(10.89)	43.53(10.5)
2	16.85(6.22)	18.83(6.34)	46.54(12.45)	51.6(11.33)
3	20.54(6.32)	22.77(6.57)	51(11.96)	54.95(11.9)
4	22.82(7.35)	24.59(7.10)	56.5(11.87)	59.64(12.05)
5	25.00(6.10)	27.14(6.15)	58.92(13.5)	61.88(13.32)

Results: Figure 1 shows the performance, in terms of ARE (lower is better), across different noise levels γ for fixed dimensionality r . As seen in the figure, TVCA2 outperforms PCA and significantly outperforms RP. The improvements are most significant for lower dimensions, e.g., $r = 2$. All three methods improve with increase in dimensionality of the latent space ($r = 5, 8$), and PCA almost catches up with TVCA2 when $r = 2$. Further, the improvements of TVCA2 over PCA is more pronounced when the noise level γ is high.

Table 2 showed the ARE results for the five steps in the test set along with standard deviation (std) when $r = 2$. In Table 2, the numbers in the leftmost column represents the time step. The performance of both TVCA2 and PCA deteriorates further into the future, but TVCA2 outperforms PCA for all 5 steps for both noise levels considered ($\gamma = 0.05, 0.1$).

Figure 6.1 shows the shape of all the latent covariances for all 25 time steps, including 20 in the training set and 5 in the test set. The ground truth are the generated \mathbf{Y}_t plotted as blue ellipses. For TVCA2 and PCA, their latent covariances were calculated based on the leading 2 components. TVCA2 is plotted as red ellipses and PCA is plotted as green ellipses. A visual comparison shows that TVCA2 is able to track most of the ground truth on training set as well as testing set much better than PCA, which also explains the quantitative results in Figure 6.1 and Table 2.

6.2 Stocks Data

We considered two real world stock market datasets, each spanning 14 years at a daily resolution. The first dataset, S&P500, contains all 381 stocks in the current S&P index which has been in the S&P index from 1995 to 2008. The second dataset, NYSE, is a widely used dataset [9, 2, 4] consisting of 36 stocks at daily resolution spanning from 1971 to 1984. For both datasets, the first 10 years' data was used as the training set, and the last 4 years' data used for testing.

Methodology: For the experiments, the covariance of the daily return, given by $\text{Return} = \frac{\text{closing price} - \text{opening price}}{\text{opening price}} \times 100\%$ (see details in [3]) was considered for both datasets. For each dataset, we constructed monthly average of the daily covariances, and each average monthly covariance was considered as an observed covariance matrix \mathbf{X}_t . For both datasets, the training set had 10 years, leading to 120 observed covariances, i.e., $T = 120$. Further, the test set spanned

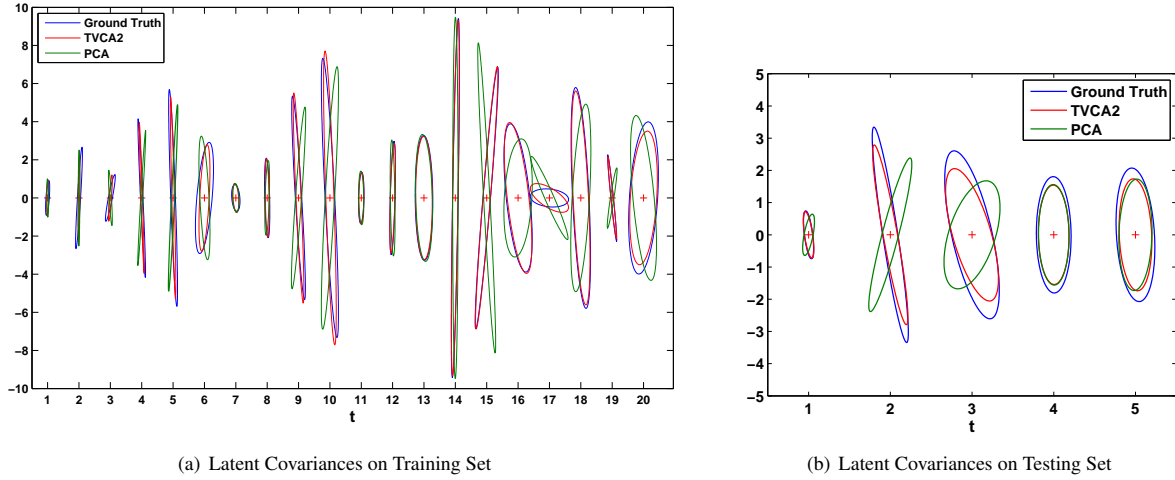


Figure 2: Latent Covariances of Artificial Data. TVCA2 is seen to track the true latent covariance better than PCA both in the training and test set. (Best viewed in color)

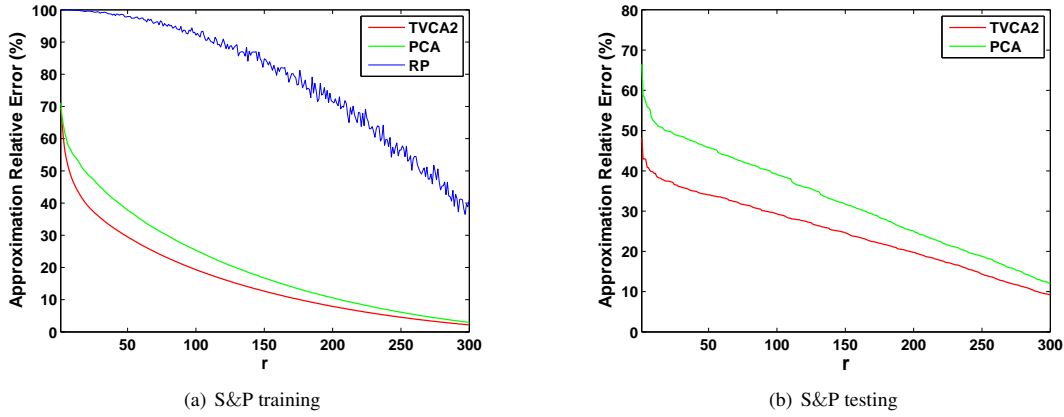


Figure 3: The change of ARE versus R on the testing set of S&P500.

4 years, thus having 48 observed covariances.

Results: The training and test set performance on S&P500 and NYSE are shown respectively in Figures 3 and 4, where the x -axis is the dimensionality r and the y -axis is the ARE (lower is better). On both datasets, TCVA2 outperforms PCA and significantly outperforms RP. The performance improvements of TCVA2 over PCA is more pronounced in S&P500 as compared to NYSE possibly since S&P500 is a higher dimensional dataset with $n = 381$ as opposed to NYSE for which $n = 36$. Another reason is that the stock market has been volatile from 1995 to 2008, while from 1971 to 1985 the stock market was relatively less volatile. Further, the improvements are more prominent in the test set. For the S&P500 dataset, the improvements are more significant in the lower dimensions and PCA catches up as the dimensionality is increased. All these observations demonstrate the potential advantages TVCA2 can provide over PCA. In Figures 5 and 6, we plot the latent covariance matrices (level sets) obtained from TVCA2 in dimensions $r = 2$ and $r = 3$ for S&P500. We plot one covariance matrix \mathbf{Y}_t for each year corresponding to the first month of the year. The variations in the latent covariance matrix \mathbf{Y}_t over time are clear from the figure. More interestingly, the latent covariances for S&P 500 even in such a low dimensions seem to capture the two major financial bubbles and market meltdowns as seen around 2001 (dot-com bubble) and 2008 (housing bubble). Similarly, the latent covariance for NYSE captures the stock market crash around 1974 resulting from the collapse of the Bretton Woods system along

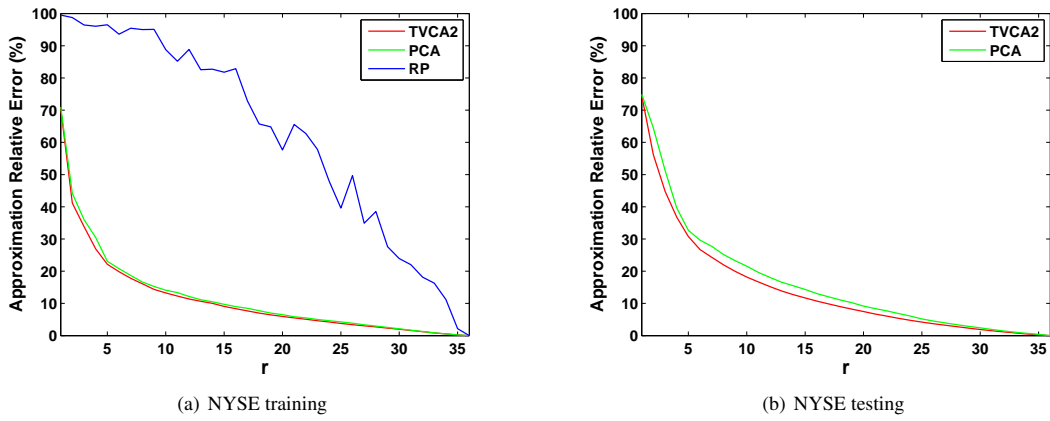


Figure 4: The change of ARE versus R on the testing set of NYSE.

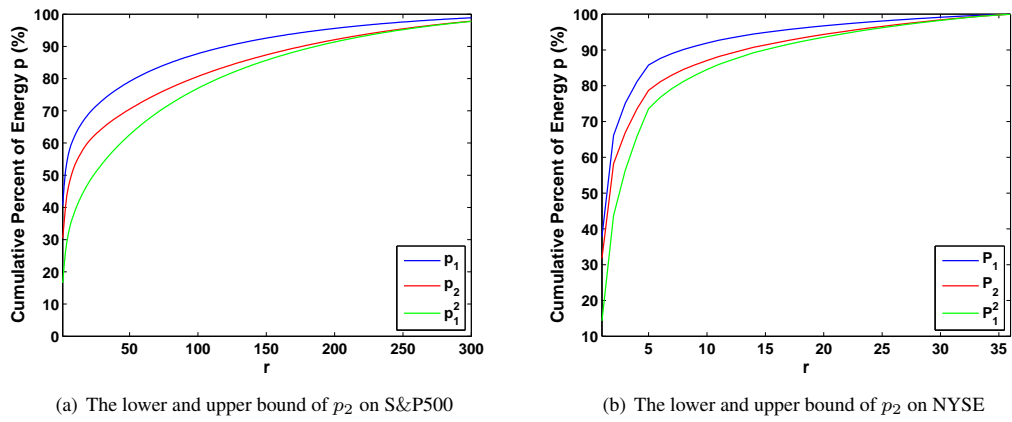
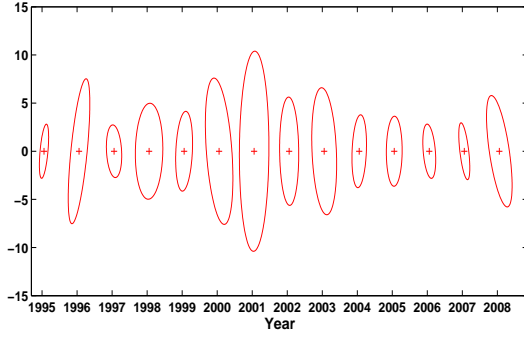
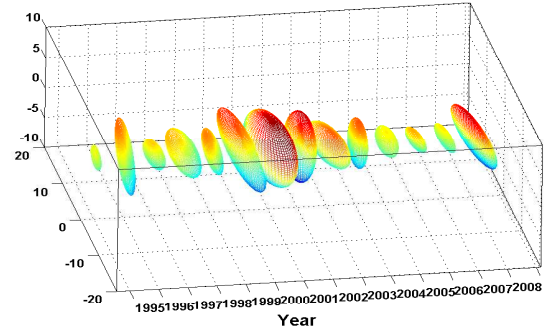


Figure 5: The lower and upper bound of p_2

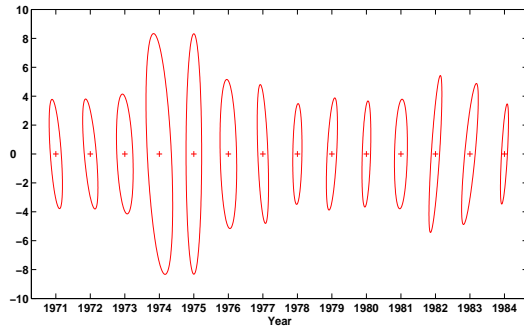


(a) 2D Latent Covariances S&P 500

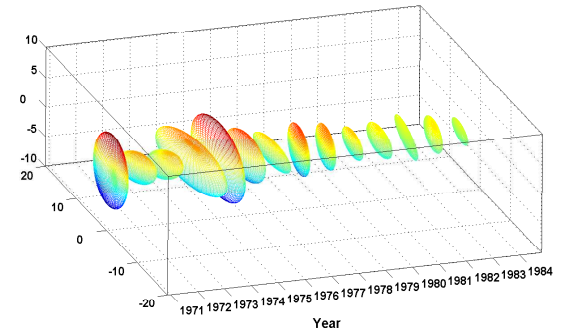


(b) 3D Latent Covariances S&P 500

Figure 6: Latent Covariances over time for S&P500 from 1995 to 2008. The two financial meltdowns in 2001 and 2008 are prominently captured in the latent low dimensional space. (Best viewed in color)



(a) 2D Latent Covariances NYSE



(b) 3D Latent Covariances NYSE

Figure 7: 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)

with the ‘Nixon Shock’ and the devaluation of the US dollar. In addition, Fig. 5 plots the upper and lower bounds of p_2 , which experimentally corroborates the result in *Proposition 2* (b).

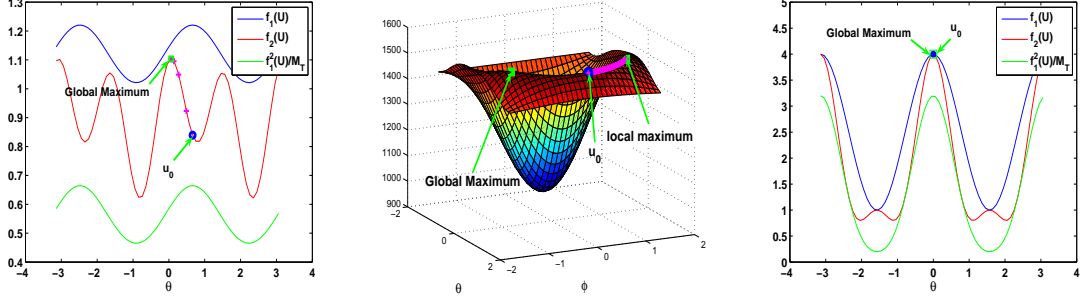
We also evaluated the efficacy of Algorithm 2 for a given ARE upper bound. The results on two stock market datasets are shown in Table 6.2 and 6.2 respectively. Given an ARE upper bound δ (first row), the corresponding dimensionality r (second row) is computed in the initialization step in Algorithm 2. The true ARE on the training and testing sets are shown in the subsequent rows. The fact that the training set ARE is guaranteed to satisfy the given upper bound is ensued by *Proposition 3*(c). As seen in the table, the bound is indeed satisfied in all cases.

$\delta(\%)$	30	20	10	5
r	73	115	187	287
training ARE(%)	24.20	17.05	9.01	4.72
testing ARE(%)	31.30	28.48	20.59	11.08

Table 3: Results of Algorithm 2 on S&P 500.

$\delta(\%)$	30	20	10	5
r	5	8	16	23
training ARE(%)	22.52	15.59	8.01	4.31
testing ARE(%)	36.53	29.40	18.44	10.01

Table 4: Results of Algorithm 2 on NYSE.



(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 8: Optimizing $f_2(\mathbf{U})$ in TVCA2 based on TVCA1 initialization and iterative updates. Objective $f_2(\mathbf{U})$ for TVCA2 is shown in red; the lower and upper bounds based on $f_1(\mathbf{U})$ for TVCA1 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.

6.3 Numerical simulation of lower and upper bound

We study TVCA2, TVCA1, and Algorithm 1 on small low dimensional problems to get additional insights into workings of the idea, including cases where the approach can and cannot find the global maxima of $f_2(\mathbf{U})$. It is important to recall that while $f_2(\mathbf{U})$ is a convex function for unconstrained \mathbf{U} , the model requires maximizing $f_2(\mathbf{U})$ 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 TVCA2 in Figure 8. In Figure 8(a), we consider 3 time steps for a 2-dimensional covariance matrix, with

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

The vector \mathbf{u} is parameterized as $\mathbf{u} = [\sin(\theta), \cos(\theta)]$, and the x -axis denotes θ . Note that $f_2(\mathbf{u})$ 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 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 $f_2(\mathbf{U})$ 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 8(a). However, the initialization does not always lead to the global maximum as shown in Figure 8(b). In Figure 8(b), we consider

$$\mathbf{X} = [\mathbf{X}_1 | \mathbf{X}_2 | \mathbf{X}_3] = \begin{bmatrix} 29.7995 & 2.5707 & 1.7377 & 21.8515 & -2.2068 & 2.0377 & 8.5273 & -2.5322 & 1.1011 \\ 2.5707 & 30.1445 & -0.0292 & -2.2068 & 22.8371 & 0.0490 & -2.5322 & 9.6724 & -0.9796 \\ 1.7377 & -0.0292 & 24.1799 & 2.0377 & 0.0490 & 21.1336 & 1.1011 & -0.9796 & 6.4754 \end{bmatrix},$$

and the vector \mathbf{u} is parameterized as $\mathbf{u} = [\sin(\theta), \cos(\theta), \sin(\phi), \cos(\theta) \cos(\phi)]$. In Figure 8(b), θ and ϕ are the x -axis and y -axis respectively, and $f_2(\mathbf{u})$ 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 8(c) shows a case where the initialization itself achieves the global maximum of TVCA2. In Figure 8(c), we consider

$$\mathbf{X} = [\mathbf{X}_1 | \mathbf{X}_2] = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

and \mathbf{u} is parameterized as in Figure 8(a). For this scenario, if \mathbf{u}_0 denotes the initialization obtained from TVCA1, we see that $f_1^{\max} = f_1(\mathbf{u}_0) = f_2(\mathbf{u}_0)$, implying $f_2(\mathbf{u}_0) = f_2^{\max}$.

7 Conclusions

The ability to study and analyze time varying covariance matrices is becoming important in several domains. Since data lives in a high-dimensional space, directly working with high dimensional covariance matrices may be problematic. Further, in several domains, the latent dynamics is expected to be in a lower dimensional space. In this paper, we have introduced a framework called TVCA2 for modeling time varying covariance matrices in low dimensions. While the framework has similarities with existing approaches to tensor decomposition, we present a novel and unique analysis of TVCA2 in terms of a more tractable framework called TVCA1. The key optimization problem in the context of TVCA1 is an EVD problem, and can be readily obtained. Interestingly, the solution to TVCA1 can be used to construct lower and upper bounds for the global maximum for TVCA2. The analysis leads to an effective initialization scheme for TVCA2. We also present an algorithm which iteratively improves the objective till convergence. We consider the algorithm in two different settings, given a fixed dimensionality and given an upper bound on the relative error w.r.t. the global maximum. The corresponding algorithms converge to local maxima of the objective with clear approximation guarantees w.r.t. the global maximum. We also discuss non-trivial conditions under which a global maximum will be achieved. We illustrate the effectiveness of the approach on synthetic data as well as two real world stock market datasets, each spanning 14 years.

While the algorithms presented in the paper follow a somewhat standard approach of starting with a reasonable initialization followed by iterative updates, the analysis presented in the paper relates the result obtained by the algorithm to the global maximum of the problem. 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 over time 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 TVCA2 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] Z. Bodie, A. Kane, and A. J. Marcus. *Investments, 6 eds.* McGraw-Hill/Irwin, 2004.
- [4] T. M. Cover. Universal portfolios. *Mathematical Finance*, 1(1):1–29, 1991.
- [5] S. Dasgupta. Experiments with random projection. In *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*, pages 143–151, 2000.
- [6] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.
- [7] K. Fukunaga. *Introduction to Statistical Pattern Recognition, 2nd edition.* Academic Press, 1990.
- [8] G. H. Golub and C. V. Loan. *Matrix Computations, 3rd ed.* Johns Hopkins University Press, 1996.
- [9] D. Helmbold, R. Schapire, Y. Singer, and M. Warmuth. Online portfolio selection using multiplicative weights. *Mathematical Finance*, 8(4):325–347, 1998.

- [10] K. Inoue and K. Urahama. Equivalence of non-iterative algorithms for simultaneous low rank approximations of matrices. In *CVPR*, pages 154–159, 2006.
- [11] 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.
- [12] T. G. Kolda and B. W. Bader. Tensor decompositions and applications. *SIAM Review*, 51(3):455–500, 2009.
- [13] T. G. Kolda, B. W. Bader, and J. P. Kenny. Higher-order web link analysis using multilinear algebra. In *ICDM*, pages 242–249, 2005.
- [14] P. M. Kroonenberg. *Applied Multiway Data Analysis*. Wiley, 2008.
- [15] 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.
- [16] 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.
- [17] L. D. Lathauwer, B. D. Moor, J. Vandewalle, and J. V. On the best rank-1 and rank-(r_1, r_2, \dots, r_n) approximation of higher-order tensors. *SIAM Journal on Matrix Analysis and Applications*, 21(4):1324 – 1342, 2000.
- [18] C. V. Loan and etc. Workshop report 1 future directions in tensor-based computation and modeling. 2009.
- [19] 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.
- [20] J. T. Scraggs 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.
- [21] S. Tadjudin and D. A. Landgrebe. Covariance estimation with limited training samples. *IEEE Transactions on Geoscience and Remote Sensing*, 37(4), July 1999.