COVARIANCE ESTIMATION OF SPATIO-TEMPORAL RANDOM VARIABLES WITH KRONECKER PRODUCT BASED MODELS

by

Can Hakan Dağıdır

B.S., Mathematics, Koç University, 2019

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Master of Science

Graduate Program in Mathematics

Boğaziçi University

2022

COVARIANCE ESTIMATION OF SPATIO-TEMPORAL RANDOM VARIABLES WITH KRONECKER PRODUCT BASED MODELS

APPROVED BY:

Assoc. Prof. Ümit Işlak (Thesis Supervisor)	.....
Assist. Prof. Mustafa Gökçe Baydoğan (Thesis Co-supervisor)	.....
Assoc. Prof. Ayhan Günaydın	.....
Assoc. Prof. Özlem Beyarslan	.....
Assist. Prof. İlker Arslan	.....

DATE OF APPROVAL: 29/07/2022

ABSTRACT

COVARIANCE ESTIMATION OF SPATIO-TEMPORAL RANDOM VARIABLES WITH KRONECKER PRODUCT BASED MODELS

Covariance estimation is a widely studied topic. Due to the nature of many problems, high-dimensional Spatio-temporal scenarios are considered frequently. In some cases, the true covariance matrices are also expected to have a Kronecker product-based representation. Especially in wind speed analysis, the true covariance matrix can be assumed to have spatial and temporal Kronecker factors. This thesis studies Kronecker product-based covariance estimation models. Also, a new Kronecker product-based method is proposed with experimentation results showing its performance.

ÖZET

KRONECKER ÇARPIMI TABANLI MODELLER İLE UZAY-ZAMANSAL RASSAL DEĞİŞKENLERİN KOVARYANS TAHMİNİ

Kovaryans tahmini birçok farklı alanda çalışılan bir konudur. Bu konudaki bazı problemlerde, doğaları gereği, çok boyutlu uzay-zamansal veriler incelenmektedir. Ayrıca, bazı senaryolarda gerçek kovaryans matrisinin başka kovaryans matrislerinin Kronecker çarpımları temelli gösterimler şeklinde yazılabileceği beklenmektedir. Özellikle rüzgar hızının incelendiği çalışmalarda gerçek kovaryans matrisinin uzaysal ve zamansal Kronecker çarpanları olduğu varsayılarak çalışmalar yürütülmüştür. Bu tezde kovaryans matrisi tahmini için geliştirilmiş Kronecker çarpımı temelli modeller incelenmektedir. Ayrıca Kronecker çarpımı bazlı yeni bir model tanımlanmış ve deney sonuçlarıyla kullanılabilirliği desteklenmiştir.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF SYMBOLS

$\mathbb{C}$	The Set of Complex Numbers
d	Number of features
d_s	Number of spatial features
d_t	Number of temporal features
D_n	Reshaped Risk Matrix of Sample Covariance Estimation
$\hat{K}_n^\lambda$	Reshaped PRLS Estimation Matrix
M	Reshaped True Covariance Matrix
$\hat{M}_n$	Reshaped Sample Covariance Matrix
n	Number of observations
$\mathcal{N}$	Normal Distribution
r	Seperation Rank
$\mathbb{R}$	The Set of Real Numbers
$\mathcal{S}_{d-1}$	Unit Euclidean Sphere in $\mathbb{R}^d$
σ_i	i-th Largest Singular Value of a Matrix
Σ	True Covariance Matrix
$\hat{\Sigma}_n$	Sample Covariance Matrix
$\hat{\Sigma}_n^\lambda$	PRLS Estimation Matrix
$\otimes$	Kronecker Product

LIST OF ACRONYMS/ABBREVIATIONS

FF	Flip-Flop
MLE	Maximum Likelihood Estimation
PCA	Principal Component Analysis
PRLS	Permuted Rank-Penalized Least Squares
SCM	Sample Covariance Matrix
SVD	Singular Value Decomposition
SVDT	Singular Value Threshold
TRKF	Temporally Reinforced Kronecker Factorization

1. INTRODUCTION

Covariance estimation is a widely studied topic in many areas such as geospatial studies [1], signal processing [2], and finance [3]. Also, there are many cases where using high-dimensional data is inevitable. Especially in geospatial studies [4, 5] and signal processing, [4,5] high dimension problem is addressed. Low-dimensional approximation of the covariance matrix is one way to handle this problem. Many studies propose different methods [6]. Sparse estimations [7], low-rank estimations [8] and Kronecker factorization [9–11] can be given as recent examples of these methods. Kronecker Product (KP) based models can also be categorized; many studies assume the existence of an underlying Kronecker product structure for the true covariance matrix, [9–11] while others try to find the nearest Kronecker product structured covariance matrix [12]. The matrices that can be expressed as a Kronecker product of other matrices with lesser dimensions are said to be separable. This thesis assumes the existence of the underlying KP sum structure and focuses on KP-based models, especially in wind speed analysis.

For multivariate wind speed analysis and wind-based electricity production, it is convenient to be interested in KP-based low-rank estimations of the covariance matrix. This is because of the high dimensional and Spatio-temporal nature of the wind analysis problem [11]. Wind speed analysis can naturally be a high-dimensional problem as the real-world data may come from many wind sensors spread across a region. Even when the number of spatial features is small, it is preferred to use interpolation to increase the dimension. Also, due to the nature of wind as a fluid, using a big window of temporal data is beneficial. When this problem is combined with wind-based electricity production, using polynomially derived wind speed features is likewise crucial. The benefit of using wind speed squared is straightforward, considering the classic kinetic energy formula. Additionally, it has been shown that third-degree polynomials of wind speed increase the performance of many production forecasting models [13]. All these results bring the use of very high-dimensional data.

In addition to the high dimensionality, again from the nature of the problem, decomposing the data to spatial and temporal variables is reasonable. With KP representation, we can separate the true covariance matrix to spatial and temporal covariance matrices. We also know that a temporal covariance matrix is a Toeplitz matrix and it has been shown that KP based models perform even better when one of the matrices are Toeplitz [9].

Throughout this thesis, we assume that the true covariance matrix Σ of a multivariate data with $d = d_s d_t$ dimensions has the form

$$\Sigma = \sum_{i=1}^{r} A_i \otimes B_i$$

where A_i and B_i are symmetric and positive semi-definite matrices with $\dim(A_i) = d_s \times d_s$ and $\dim(B_i) = d_t \times d_t$. We also inspect cases where one or both of these matrices are Toeplitz. For this type of separation, we inspect some estimation models in detail. It is easy to show that this representation exists for any matrix for a large enough r [10]. However, matrices that can be represented with small r can be estimated better with the discussed models. We analyse the Frobenius estimation error's norm and show bounds on that norm. Also we discuss the convergence rates of the estimations in different cases.

We then propose a new model, Temporally Reinforced Kronecker Factorization (TRKF), for reinforcing the covariance matrix estimations when a spatio-temporal or a similar decomposition for the data is available.

We run simulations for different cases and present their results for confirming the theoretical arguments. A wide range of Kronecker factor types, separation ranks, dimensions and number of observations are used to show the generality of the results. Also we tested the models on real world multivariate wind speed and wind-based electricity production data from different regions of Turkey. These experiments confirmed the credibility of the discussed models and show some valuable insights from the nature of wind.

The rest of this thesis is as follows. In Chapter 2 we provide some basics on preliminaries. These include results from linear algebra and probability theory. Chapter 3 discusses the theoretical background of our study, with the main emphasis being on the permuted rank penalized least squares. Further we introduce a new approach TRKF. In Chapter 4 we first verify the theoretical results with experiments, then show our proposed approach's performance. We conclude the thesis in Chapter 5 with a summary and possible directions.

2. PRELIMINARIES

2.1. Basic Definitions

The purpose of this section is to give some basic definitions that will be required in the rest of the thesis. Starting from certain definitions from elementary linear algebra, we then recall matrix norms and provide some further discussions.

Definition 2.1. *A matrix A is called **symmetric** if it equals its transpose, i.e. $A = A^T$, where A^T represents the transpose of A.*

Definition 2.2. *A matrix is called **Hermitian** if it is equal to the transpose of its complex conjugate, i.e. A is Hermitian if $A = \overline{A^T}$ where $\overline{A}$ represents the complex conjugate of A.*

Note that for real matrices, Hermitian and symmetric mean the same.

Definition 2.3. *Let A be a symmetric matrix of size $n \times n$. Then, A is said to be **positive-definite** if $\mathbf{x}^T A\mathbf{x} > 0$ for all $\mathbf{x} \in \mathbb{R}^n \setminus \{0\}$, and **positive semi-definite** if $\mathbf{x}^T A\mathbf{x} \geq 0$ for all $\mathbf{x} \in \mathbb{R}^n$.*

The definition for complex matrices is similar. If A is a Hermitian matrix of size $n \times n$, A is said to be positive-definite if $\overline{\mathbf{x}^T} A\mathbf{x} > 0$ for all $\mathbf{x} \in \mathbb{C}^n \setminus \{0\}$, and positive semi-definite if $\overline{\mathbf{x}^T} A\mathbf{x} \geq 0$ for all $\mathbf{x} \in \mathbb{C}^n$.

Definition 2.4. *A **covariance matrix** C is a symmetric and positive-definite matrix that contains the covariances between the elements of multivariate random variables. More precisely, if $X_1, \ldots, X_n$ are random variables whose second moments exist, $C_{ij} = Cov(X_i, X_j)$ for $i, j = 1, 2, \ldots, n$.*

Clearly, the main diagonal of a covariance matrix contains the variances of the variables.

Definition 2.5. *An $n \times n$ matrix A is said to be a* ***Toeplitz Matrix*** *if there are some constants $d_{1-n}, \ldots, d_{n-1}$ such that $A_{i,j} = d_{i-j}$ whenever $i, j \in \{1, 2, \ldots, n\}$.*

Toeplitz matrices are also known to be diagonal-constant matrices, because each descending diagonal entry of the matrix from left to right is constant. A visual example for a Toeplitz matrix is

$$\begin{bmatrix} a & b & c & d \\ e & a & b & c \\ f & e & a & b \\ g & f & e & a \end{bmatrix}.$$

Below we will be interested in these special matrices since some of the covariance matrices in our framework will be approximately Toeplitz (e.g. temporal covariance matrices).

Next we will briefly go over matrix norms which will be important in our study below. Let us begin with recalling vector norms for this purpose.

Definition 2.6. *Let V be a vector space over a field $\mathbb{F}$. A mapping*

$$\|\cdot\| : V \to \mathbb{R}$$

is a ***vector norm*** *if, for all $\mathbf{x}, \mathbf{y} \in V$ and all $c \in \mathbb{F}$, the followings hold:*

- $\|\mathbf{x}\| \geq 0$
- $\|\mathbf{x}\| = 0 \iff \mathbf{x} = 0$
- $\|c\mathbf{x}\| = |c|\|\mathbf{x}\|$
- $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|.\mathbf{y}\|$

Definition 2.7. *The* ***Euclidean norm*** *(ℓ_2-norm) of a vector $\mathbf{x} = [x_1 \ldots x_n]^T \in \mathbb{R}^n$ is*

$$\|\mathbf{x}\|_2 = \left(|x_1|^2 + |x_2|^2 + \cdots + |x_n|^2\right)^{\frac{1}{2}}.$$

Note that the Euclidean norm is a special case of the ℓ_p-norms.

Definition 2.8. *ℓ_p-**norm** of a vector $\mathbf{x} \in \mathbb{R}^n$ is $\|\mathbf{x}\|_p = (|x_1|^p + |x_2|^p + \cdots + |x_n|^p)^{\frac{1}{p}}$.*

A vector norm defined on a vector space of matrices is called a matrix norm.

Definition 2.9. ***The matrix-p norm** $\|A\|_p$ of a matrix A is defined to be $\|A\|_p = \sup_{x \neq 0} \frac{\|A\mathbf{x}\|_p}{\|\mathbf{x}\|_p}$.*

One nice and useful property enjoyed by matrix p-norms is the following.

Proposition 2.10. *Matrix p-norm is sub-multiplicative: If $A \in \mathbb{R}^{m \times n}$ and $\mathbf{x} \in \mathbb{R}^n$, then $\|A\mathbf{x}\|_p \leq \|A\|_p \|\mathbf{x}\|_p$.*

Definition 2.11. *The **spectral norm** of a matrix $A \in \mathbb{R}^{m \times n}$ is defined as*

$$\|A\|_2 = \sup_{\mathbf{x} \neq 0} \frac{\|A\mathbf{x}\|_2}{\|\mathbf{x}\|_2}.$$

Some further discussion on matrix norms will be included in the next section after reviewing singular values. We conclude this section with two more independent definitions that will be useful in the sequel.

Definition 2.12. *The operator **vec** on a matrix stacks the columns of the matrix on top of each other to create a vector.*

A visual example of the vec operation is shown below:

$$vec\left(\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix}\right) = \begin{bmatrix} a_{11} \\ a_{21} \\ a_{12} \\ a_{22} \\ a_{13} \\ a_{23} \end{bmatrix}.$$

The following probabilistic big oh notation will be used in our theoretical discussions.

Definition 2.13. ***Stochastic boundedness*** *is denoted with O_p. For a given sequence of random variables X_n, the equation $X_n = O_p(c_n)$ as $n \to \infty$ represents: $\forall \varepsilon > 0$, $\exists M, N > 0$ finite such that, $\mathbf{P}\left(\left|\frac{X_n}{c_n}\right| > M\right) < \varepsilon$ for all $n > N$.*

2.2. Singular Value Decomposition and Principal Component Analysis

In this section, we will define and discuss *Principal Component Analysis* (PCA), *Eigendecomposition*, and *Singular Value Decomposition* (SVD). PCA is a commonly used method for dimensionality reduction because it easily extracts the most "important" information from the data while reducing the dimension. First, with eigendecomposition, a new basis for the data is created. Then, the first *principal component* (PC) is created using the eigenvector corresponding to the largest eigenvalue. This method ensures that the first PC carries the maximum possible variance of the data.

Before going into further details, we first recall basics related to eigenvalues and explain eigendecomposition. Recall that for a square matrix A, the non-zero vectors $\mathbf{v}$ and corresponding scalars λ satisfying the equation $A\mathbf{v} = \lambda\mathbf{v}$ are called as the **eigenvectors** and **eigenvalues** of A, respectively.

It is well known that any symmetric matrix A can be decomposed as

$$A = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{-1}$$

where the columns of $\mathbf{V}$ are eigenvectors of A and where $\mathbf{\Lambda}$ is a diagonal matrix whose diagonal entries are the eigenvalues of A. This decomposition is called the **eigendecomposition** of A.

In this thesis, our main concern will be on positive semi-definite matrices. So let us mention a few properties of eigendecompositions of positive semi-definite matrices before going further:

- Eigendecomposition $\mathbf{V}\mathbf{\Lambda}\mathbf{V}^{-1}$ of any positive semi-definite matrix A always exists.
- Eigenvalues of such matrices are always non-zero, and eigenvectors are pairwise

orthogonal when their eigenvalues are different.

- Because eigenvectors corresponding to different eigenvalues are orthogonal, it is possible to store all the eigenvectors in an orthogonal matrix.
- $\mathbf{V}^T = \mathbf{V}^{-1}$

Now we will define a generalized eigendecomposition method, which loosens the square matrix condition.

Definition 2.14. ***Singular values*** *of a real matrix A are defined to be the square roots of the eigenvalues of the matrix A^TA.*

Definition 2.15. ***Singular Value Decomposition*** *of a (rectangular) matrix A is: $A = U\Delta V^T$, where the columns of U are the eigenvectors of AA^T, the columns of V are the eigenvectors of A^TA and Δ is a diagonal matrix whose diagonal entries are the singular values.*

The columns of U and V are called the left and right **singular vectors** of A, respectively. By convention, the singular values are ordered, i.e., $\sigma_1 \geq \sigma_2 \geq \cdots \geq 0$.

Let us also note that the **Thresholded Singular Value Decomposition** (SVT) is a simple modification to SVD. With SVT, small singular values are filtered out to obtain a low-rank approximation of a given matrix. It is also used for convex relaxation of rank minimization problems [14].

With the given definitions and remarks, we can talk more about PCA. A new basis for data can be created using the eigenvectors of the covariance matrix. As the eigenvectors are unit vectors, the corresponding eigenvalues are correlated with the explained variance of the data. Thus, choosing the eigenvectors corresponding to only the large enough eigenvalues creates a new representation of the data with fewer dimensions but almost the same information. Note that the old basis vectors had one element as 1 and all others as 0; in other words, each basis vector was only for one variable of the multivariate data. Also, the previous data had features that were correlated with each other. The new basis vectors let us contain information from multiple

variables in one dimension, and the new derived features (PCs) are not correlated. An additional note for the application is that instead of doing eigendecomposition to the full data, it is usually preferred to do partial SVD for computational reasons.

We now go back to matrix norms and conclude this section with some results and notes related to singular values and norms. The first enties the spectral norm to spectral values.

Proposition 2.16. *The spectral norm of a matrix A equals to the largest singular value σ_1 of A.*

In other words,

$$\|A\|_2 = \sqrt{\lambda_{\max}(A^T A)} = \max\{\sigma(A)\}$$

where $\sigma(A)$ is the set of singular values of A.

Definition 2.17. ***Frobenius norm*** *of an $m \times n$ matrix A is defined to be*

$$\|A\|_F = \sqrt{\sum_{i,j=1}^{n} {a_{ij}}^2} = \sqrt{\mathrm{tr}(A^T A)} = \sqrt{\sum_{i=1}^{min(m,n)} \sigma_i^2(A)}.$$

Note that we have

$$\|A\|_F = \sqrt{\mathrm{tr}(A^T A)} = \sqrt{\sum_{i=1}^{\min(m,n)} \sigma_i^2(A)}.$$

One nice property of the Frobenius norm is a certain permutation invariance. Namely, letting τ be a permutation in S_{mn}, if permute the elements of an $m \times n$ matrix A by using τ to obtain a new matrix A', we have

$$\|A\|_F = \sqrt{\sum_{i,j=1}^{n} {a_{ij}}^2} = \sqrt{\sum_{i,j=1}^{n} {a_{\tau(i)\tau(j)}}^2} = \|A'\|_F.$$

As we have defined SVD and the Frobenius norm, we give a fundamental result on low-rank matrix estimation.

Proposition 2.18 (Eckart-Young Theorem [15])**.** *Let A be a matrix with $rank(A) = k$. Non-increasingly ordered singular values of A are σ_i and corresponding left and*

right singular vectors are $\mathbf{u}_i$ *and* $\mathbf{v}_i$*, respectively. Then, any rank-r matrix* $A_r = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T$*, with* $r \leq k$*, satisfies*

$$A_r = \underset{A^\star : rank(A^\star)=r}{\arg\min} \|A - A^\star\|_F.$$

Note that Eckart-Young theorem states that rank-r approximation of any matrix can be found with a truncated SVD but in many scenarios this problem is not convex, that will be addressed in the following sections.

2.3. Kronecker Products

Let $A \in \mathbb{R}^{p_1 \times p_2}$ and $B \in \mathbb{R}^{q_1 \times q_2}$. Then we define the **Kronecker Product** as the operation $\otimes : (\mathbb{R}^{p_1 \times p_2}, \mathbb{R}^{q_1 \times q_2}) \to \mathbb{R}^{p_1 q_1 \times p_2 q_2}$, such that

$$A \otimes B = \begin{bmatrix} a_{1,1}B & \dots & a_{1,p_2}B \\ \vdots & \ddots & \vdots \\ a_{p_1,1}B & \dots & a_{p_1,p_2}B \end{bmatrix}$$

where $a_{i,j}$ is the element at i'th row and jth column of the matrix A.

An example visual that shows $A \otimes B$ for $A \in \mathbb{R}^{2\times 2}$ and$B \in \mathbb{R}^{3\times 3}$:

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \otimes \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{bmatrix} = \left[\begin{array}{ccc|ccc} a_{11}b_{11} & a_{11}b_{12} & a_{11}b_{13} & a_{12}b_{11} & a_{12}b_{12} & a_{13}b_{13} \\ a_{11}b_{21} & a_{11}b_{22} & a_{11}b_{23} & a_{12}b_{21} & a_{12}b_{22} & a_{13}b_{23} \\ a_{11}b_{31} & a_{11}b_{32} & a_{11}b_{33} & a_{12}b_{31} & a_{12}b_{32} & a_{13}b_{33} \\ \hline a_{21}b_{11} & a_{21}b_{12} & a_{21}b_{13} & a_{22}b_{11} & a_{22}b_{12} & a_{23}b_{13} \\ a_{21}b_{21} & a_{21}b_{22} & a_{21}b_{23} & a_{22}b_{21} & a_{22}b_{22} & a_{23}b_{23} \\ a_{21}b_{31} & a_{21}b_{32} & a_{21}b_{33} & a_{22}b_{31} & a_{22}b_{32} & a_{23}b_{33} \end{array}\right].$$

2.3.1. Kronecker Product Basic Properties

Proposition 2.19. *The following properties hold for Kronecker product (here* α, β *are scalars, and the matrices* A, B, C, D *are assumed to be compatible in terms of the given operations):*

(i) Distributive:

- $(A + B) \otimes C = (A \otimes C) + (B \otimes C)$
- $A \otimes (B + C) = (A \otimes C) + (A \otimes B)$

(ii) Scalars:

- $\alpha \otimes A = \alpha A$
- $A \otimes \alpha = \alpha A$
- $(\alpha A) \otimes (\beta B) = (\alpha\beta)(A \otimes B)$

(iii) Associative:

- $(A \otimes B) \otimes C = A \otimes (B \otimes C)$

(iv) Inverse:

- $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$

(v) Transpose:

- $(A \otimes B)^T = A^T \otimes B^T$

(vi) Multi Products: if AC and BD products exist (or they are well defined):

- $(A \otimes B)(C \otimes D) = (AC) \otimes (BD)$

(vii) Trace:

- $\operatorname{tr}(A \otimes B) = \operatorname{tr}(A) \operatorname{tr}(B)$

(viii) Rank:

- $rank(A \otimes B) = rank(A) rank(B)$.

Proofs of these results are elementary. We just prove the trace proposition as an exemplary work.

Proof. (of $\operatorname{tr}(A \otimes B) = \operatorname{tr}(A) \operatorname{tr}(B)$) Here we inherently assume that A, B are square. Let $\dim(A) = p \times p$ and $\dim(B) = q \times q$. Note that

$$\operatorname{tr}(A) = \sum_{i=1}^{p} a_{ii}.$$

Then, from directly from the definition of Kronecker product, we get

$$\operatorname{tr}(A \otimes B) = \sum_{i=1}^{p} \sum_{j=1}^{q} a_{ii} b_{jj} = \sum_{i=1}^{p} a_{ii} \sum_{j=1}^{q} b_{jj} = \operatorname{tr}(A) \operatorname{tr}(B).$$

□

2.3.2. Eigenvalues

In this subsection, we briefly go over the relation between eigenvalues and Kronecker products. The following result on deriving the eigenvalues and eigenvectors of a matrix from its Kronecker Product factors is classical.

Theorem 2.20. *Let A and B be square matrices with respective sizes $p \times p$ and $q \times q$. Assume that both A and B are of full rank. Let $\lambda_1^{(a)}, \lambda_2^{(a)}, \ldots, \lambda_p^{(a)}$ and $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_p$ be the eigenvalues and the corresponding eigenvectors of A. Let $\lambda_1^{(b)}, \lambda_2^{(b)}, \ldots, \lambda_q^{(b)}$ and $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_q$ be the eigenvalues and the corresponding eigenvectors of B. Then eigenvalues and the corresponding eigenvectors of the matrix $A \otimes B$ are: $\lambda_i^{(a)}\lambda_j^{(b)}$ and $\mathbf{u}_i \otimes \mathbf{v}_j$ for $i \in \{1, 2, \ldots, p\}$ and $j \in \{1, 2, \ldots, q\}$.*

Proof. From the definition of an eigenvalue, we get

$$A\mathbf{u}_i = \lambda_i^{(a)}\mathbf{u}_i \quad \text{and} \quad B\mathbf{v}_j = \lambda_j^{(b)}\mathbf{v}_j.$$

We also know that the following two equalities hold

$$\begin{aligned} (A\mathbf{u}_i) \otimes (B\mathbf{v}_j) &= (A \otimes B)(\mathbf{u}_i \otimes \mathbf{v}_j) \\ (\lambda_i^{(a)}\mathbf{u}_i) \otimes (\lambda_j^{(b)}\mathbf{v}_j) &= \lambda_i^{(a)}\lambda_j^{(b)}(\mathbf{u}_i \otimes \mathbf{v}_j). \end{aligned}$$

Combining these we obtain

$$(A \otimes B)(\mathbf{u}_i \otimes \mathbf{v}_j) = (A\mathbf{u}_i) \otimes (B\mathbf{v}_j) = (\lambda_i^{(a)}\mathbf{u}_i) \otimes (\lambda_j^{(b)}\mathbf{v}_j) = \lambda_i^{(a)}\lambda_j^{(b)}(\mathbf{u}_i \otimes \mathbf{v}_j).$$

□

2.3.3. Useful Results about Kronecker Product

Various properties of matrices are preserved under Kronecker products. We list three in the following proposition.

Proposition 2.21. *The following are true:*

$$\textit{If } A \textit{ and } B \textit{ are } \left\{ \begin{array}{c} \textit{Hermitian} \\ \textit{positive definite} \\ \textit{Toeplitz} \end{array} \right\}, \textit{ then } \; A \otimes B \;\; \textit{is} \left\{ \begin{array}{c} \textit{Hermitian} \\ \textit{positive definite} \\ \textit{Toeplitz} \end{array} \right\}.$$

Here we will discuss just the proof for the Hermitian case. For the proofs others see, for example, Van Loan [10]. The cited work contains further properties that are (and are not) preserved under Kronecker products.

For the proof of $(A \otimes B)$ is Hermitian when A, B are Hermitian, first we note without proof that the property $(A \otimes B)^* = A^* \otimes B^*$ holds. Next we recall that for $u, v \in \mathbb{C}^n$, inner product is defined as

$$\langle v \ , \ w \rangle := \sum_{i=1}^{n} \overline{u}_i v_i$$

where $\overline{u}$ is the complex conjugate of u. Now, assume A and B are Hermitian matrices. By definition of Hermitian adjoint, we have

$$\langle (A \otimes B)v \ , \ w \rangle = \langle v \ , \ (A \otimes B)^* w \rangle$$

and we want to show $\langle (A \otimes B)v \ , \ w \rangle = \langle v \ , \ (A \otimes B)w \rangle$ for any v, w. Thus, the following equations

$$\begin{aligned} \langle (A \otimes B)v \ , \ w \rangle &= \langle v \ , \ (A \otimes B)^* w \rangle \\ &= \langle v \ , \ (A^* \otimes B^*)w \rangle \\ &= \langle v \ , \ (A \otimes B)w \rangle \end{aligned}$$

finalize the proof.

2.4. Reshaping and Visualizing Matrices

As matrix reshaping has an important part in this thesis, we explain the procedure visually. Before defining the reshape function R of Van Loan [10], we start with visualizing the sample covariance matrix (SCM).

Let $\mathbf{v} \in \mathbb{R}^d$, be a multivariate normal random variable with $\mathbf{v} \sim N(0, \Sigma)$ and $d = d_s d_t$. Taking n i.i.d observations $\{\mathbf{v}_\alpha\}_{\alpha=1}^n$, the SCM can be calculated using

$$\hat{\Sigma}_n = \frac{1}{n} \sum_{\alpha=1}^{n} \mathbf{v}_\alpha \mathbf{v}_\alpha^T.$$

Letting $a_\alpha(i,j) = \mathbf{v}_\alpha(i)\mathbf{v}_\alpha(j)^T$, where $\mathbf{v}(i)$ represents the i-th element of the vector $\mathbf{v}$, we can visualize $\hat{\Sigma}_n$ in more detail

$$\hat{\Sigma}_n = \frac{1}{n} \sum_{\alpha=1}^{n} \begin{bmatrix} a_\alpha(1,1) & \dots & a_\alpha(1,d_t) & \dots & a_\alpha(1,2d_t) & \dots & a_\alpha(1,d_s d_t) \\ \vdots & & & & & & \\ a_\alpha(d_t,1) & \dots & a_\alpha(d_t,d_t) & \dots & \dots & \dots & \dots \\ \vdots & & & & & & \\ a_\alpha(2d_t,1) & \dots & a_\alpha(2d_t,d_t) & \dots & \dots & \dots & \dots \\ \vdots & \ddots & & & & & \\ \vdots & & & \ddots & & & \\ a_\alpha(d_s d_t,1) & \dots & \dots & \dots & \dots & \dots & a_\alpha(d_s d_t,d_s d_t) \end{bmatrix}.$$

We may consider this matrix as a $d_s \times d_s$ matrix of $d_t \times d_t$ blocks. This consideration will be useful for reshaping.

2.4.1. Blocking

Let $A \in \mathbb{R}^{d \times d}$ with $d = d_s d_t$. Visualization of A in more detail is as follows

$$A = \begin{bmatrix} a_{1,1} & \dots & a_{1,t} & \dots & a_{1,2d_t} & \dots & a_{1,d_s d_t} \\ \vdots & & & & & & \\ a_{d_t,1} & \dots & a_{d_t,d_t} & \dots & \dots & \dots & \dots \\ \vdots & & & & & & \\ a_{2d_t,1} & \dots & a_{2d_t,d_t} & \dots & \dots & \dots & \dots \\ \vdots & \ddots & & & & & \\ \vdots & & & \ddots & & & \\ a_{d_s d_t,1} & \dots & \dots & \dots & \dots & \dots & a_{d_s d_t,d_s d_t} \end{bmatrix}.$$

We consider this matrix as a $d_s \times d_s$ matrix of $d_t \times d_t$ blocks, and define $A_{[i,j]}$ as the (i,j)th element of the block matrix. Also for simplicity, we define another submatrix

$$A_{[j]} = \begin{bmatrix} \mathrm{vec}(A_{[1,j]})^T \\ \vdots \\ \mathrm{vec}(A_{[d_s,j]})^T \end{bmatrix}.$$

An example with $d_s = 3$ and $d_t = 2$, we can split A as

$$A = \left[\begin{array}{cc|cc|cc} a_{1,1} & a_{1,2} & a_{1,3} & a_{1,4} & a_{1,5} & a_{1,6} \\ a_{2,1} & a_{2,2} & a_{2,3} & a_{2,4} & a_{2,5} & a_{2,6} \\ \hline a_{3,1} & a_{3,2} & a_{3,3} & a_{3,4} & a_{3,5} & a_{3,6} \\ a_{4,1} & a_{4,2} & a_{4,3} & a_{4,4} & a_{4,5} & a_{4,6} \\ \hline a_{5,1} & a_{5,2} & a_{5,3} & a_{5,4} & a_{5,5} & a_{5,6} \\ a_{6,1} & a_{6,2} & a_{6,3} & a_{6,4} & a_{6,5} & a_{6,6} \end{array}\right].$$

Then we get

$$A_{[1,2]} = \begin{bmatrix} a_{1,3} & a_{1,4} \\ a_{2,3} & a_{2,4} \end{bmatrix} \text{ and } A_{[1]} = \begin{bmatrix} a_{1,1} & a_{2,1} & a_{1,2} & a_{2,2} \\ a_{3,1} & a_{4,1} & a_{3,2} & a_{4,2} \\ a_{5,1} & a_{6,1} & a_{5,2} & a_{6,2} \end{bmatrix}.$$

We also define a subvector representation. For a vector $\mathbf{v}$, define

$$\mathbf{v}(d_t, i) := [\mathbf{v}]_{(i-1)d_t+1:id_t}.$$

2.4.2. Reshaping

Van Loan [10] defines a reshaping operation for reformulating the matrix approximation with the Frobenius norm minimization problem to rank-one (or rank-r) approximation problem. With the notations from the previous section, we can define the reshaping operator R as

$$\mathrm{R}(A) := \begin{bmatrix} A_{[1]} \\ \vdots \\ A_{[d_s]} \end{bmatrix}.$$

To be more precise, $\mathrm{R} : \mathbb{R}^{d_s d_t \times d_s d_t} \to \mathbb{R}^{d_s^2 \times d_t^2}$ sets the $(i-1)d_s + j$ row of $\mathrm{R}(A)$ equal to $vec(A(i,j))^T$.

Reshaping has an important role in this thesis, so we show some examples of specific matrices in detail:

- Reshaping the sample covariance matrix $\hat{\Sigma}_n \in \mathbb{R}^{d_s d_t \times d_s d_t}$ gives us a matrix with dimensions $\dim(\mathrm{R}(\hat{\Sigma}_n)) = d_s^2 \times d_t^2$.
- The reshaped risk matrix of the SCM estimator $D_n = \mathrm{R}(\hat{\Sigma}_n) - \mathrm{R}(\Sigma)$ can be represented as:

$$\begin{aligned} D_n &= \begin{bmatrix} \hat{\Sigma}_{[1]} - \Sigma_{[1]} \\ \vdots \\ \hat{\Sigma}_{[d_s]} - \Sigma_{[d_s]} \end{bmatrix} \\ &= \frac{1}{n} \sum_{\alpha=1}^{n} \begin{bmatrix} \mathrm{vec}(\mathbf{v}_\alpha(d_t,1)\mathbf{v}_\alpha(d_t,1)^T)^T - \mathbb{E}[\mathrm{vec}(\mathbf{v}(d_t,1)\mathbf{v}(d_t,1)^T)^T] \\ \vdots \\ \mathrm{vec}(\mathbf{v}_\alpha(d_t,d_s)\mathbf{v}_\alpha(d_t,d_s)^T)^T - \mathbb{E}[\mathrm{vec}(\mathbf{v}(d_t,d_s)\mathbf{v}(d_t,d_s)^T)^T] \end{bmatrix}. \end{aligned}$$

3. THEORY

In this chapter, we first discuss some related work and show some results that are useful for the PRLS. Then we analyze the PRLS method in depth and prove some significant, relevant results. Finally, we will propose a new method TRKF, and then discuss discuss its methodology and improvements.

3.1. Historical Look

In this section, we first recall the canonical covariance estimation, SCM, and we discuss how that estimation can be improved in certain situations. Then we describe some of the related works that bring a better estimate. Also, we include some results which will be useful when discussing the PRLS. We begin with a simple penalized modification to SCM [16]. Secondly, we examine the first work that uses KP representation to estimate a matrix [10]. Then, we briefly describe Lu and Zimmerman's [17] "Flip-Flop" method for solving the MLE problem of Van Loan [10]. Before we start analyzing the PRLS method, we discuss the works of Werner et al. [9] in depth as their method outperforms all other methods in some situations.

3.1.1. Sample Covariance Matrix

Sample covariance matrix is the most basic approach for estimating the true covariance matrix. Assuming $\{\mathbf{v}_i\}_{\alpha=1}^{n}$ are n i.i.d. observations of a multivariate normal variable $\mathbf{v} \sim N(0, \Sigma)$. Then the sample covariance matrix $\hat{\Sigma}_n$ is:

$$\hat{\Sigma}_n = \frac{1}{n} \sum_{\alpha=1}^{n} \mathbf{v}_\alpha \mathbf{v}_\alpha^T.$$

When the observations are complete and represent the variables well enough, the SCM is a good estimator of the covariance matrix. However, this is not necessarily always the case. Especially when dealing with very high dimensions, not having a large enough sample to represent the true distribution is a common problem.

Many shortcomings of the SCM have lead to creation of many different estimations, especially for small sample size there are many low-rank estimation methods.

3.1.2. Penalized Sample Covariance Matrix

Lounici [16] studies the low rank estimation problem on high dimensional data with missing variables. They first reduce this problem to a convex minimization problem with adding a penalty term

$$\min_A \|\hat{\Sigma}_n - A\|_F^2 + \lambda\|A\|_1$$

where $\lambda > 0$ is the regularization parameter that penalizes the nuclear norm of the estimation. Note that with this penalty term, they get a convex problem and it is computationally efficient in high dimensions. They then modify this approach for handling missing variables, but that part is not covered for staying in the scope of this paper.

3.1.3. Not-Penalized Kronecker Estimator

Van Loan and Pitsianis [10] propose a reshaping (or as they call permutation) method for solving the norm minimization problems of the form

$$\min_{B,C} \|A - B \otimes C\|_F$$

where $A \in \mathbb{R}^{d\times d}$, $B \in \mathbb{R}^{d_s\times d_s}$ and $C \in \mathbb{R}^{d_t\times d_t}$ are matrices with $d = d_s d_t$.

They reformulate this approximation problem to a "rank-one" problem with reshaping the matrices. The reshaping is done with the mapping R as it was explained in the Section 2.4.2. They prove that

$$\|A - B \otimes C\|_F = \|\mathrm{R}(A) - \mathrm{vec}(B)\mathrm{vec}(C)^T\|_F.$$

Then, calculating the singular values of $\mathrm{R}(A)$ gives the best values of $\mathrm{vec}(B)$ and $\mathrm{vec}(C)$.

3.1.4. Flip-Flop Algorithm

Lu and Zimmerman [17] propose a more general algorithm for estimating any separable matrix. They solve a maximum likelihood estimation (MLE) problem with their proposed "Flip-Flop" (FF) method. That method recursively estimates each Kronecker factor while keeping the other one fixed. When one of the Kronecker factors are known, the rank-one minimization problem becomes convex. Also, it is worth noting that the FF method has many shortcomings in covariance estimation [9]. First, FF algorithm lacks asymptotic efficiency due to its recursiveness. It also prohibits the general linear structure that both Werner [9] and Hero [11] use to reduce the problem to a low-rank minimization problem. Also, the special properties of the covariance matrices, such as being positive definite, or sometimes Toeplitz, cannot be used for advantage in FF.

3.1.5. Werner's Kronecker Product Estimator

Werner et al. [9] propose multiple approaches for the covariance estimation problem. They start with assuming the separability of the covariance matrix to Kronecker factors. Their first model modifies the FF algorithm to a less recursive version. They note that their method gives remarkably close estimations with only three iterations compared to huge number of iterations. Also their method is invariant to the initial Kronecker factor values.

Their second method (R1LS) starts with criticizing the papers approximating via minimizing the Frobenius norm because of their lack of asymptotic efficiency. Thus, they are interested in minimizing a custom weighted norm that is dependent on the sample covariance matrix. They propose a method which reshapes the matrices just like Van Loan [10] to reformulate the problem to a rank-one approximation problem. After the reformulation the custom weighted norm minimization problem can be solved with SVD. Also, they note that when one of the Kronecker factors is Toeplitz, the estimator works even better. Note that, when separating wind speed to spatial and temporal covariance matrices, the temporal covariance matrix is of the form Toeplitz.

3.2. Permuted Rank-Penalized Least Squares

One of the main problems in MLE when approximating the covariance matrix is the non-convexness of the problem. As aforementioned, many different studies handle this problem in many different ways. By reshaping the covariance matrix as Van Loan [10] and adding a penalization like Lounici [16], one can reduce the approximation problem to a convex low-rank optimization problem. Also instead of assuming $\Sigma = A \otimes B$ and get a rank-one minimization problem, we assume $\Sigma = \sum_{i=1}^{r} A_i \otimes B_i$ to get a rank-r minimization problem.

Stating the problem once more, let $\mathbf{v} \in \mathbb{R}^d$, be a multivariate normal random variable with $\mathbf{v} \sim N(0, \Sigma)$ and $d = d_s d_t$. Taking n i.i.d observations $\{\mathbf{v}_\alpha\}_{\alpha=1}^n$, we can calculate the SCM $\hat{\Sigma}_n = \frac{1}{n}\sum_{\alpha=1}^{n} \mathbf{v}_\alpha \mathbf{v}_\alpha^T$. Then assume the true covariance matrix Σ has a KP expansion representation: $\Sigma = \sum_{i=1}^{r} A_i \otimes B_i$ for some matrices A_i and B_i with $\dim(A_i) = d_s \times d_s$ and $\dim(B_i) = d_t \times d_t$. Before starting to rank-r problem, we show some results for $\Sigma = A \otimes B$ then generalize it to the case where $\Sigma = \sum_{i=1}^{r} A_i \otimes B_i$.

Theorem 3.1. *[10, Theorem 2.1] Let $\Sigma \in \mathbb{R}^{d\times d}$ with $d = d_s d_t$. If $A \in \mathbb{R}^{d_t \times d_t}$ and $B \in \mathbb{R}^{d_s \times d_s}$, then*

$$\|\Sigma - A \otimes B\|_F = \|\mathrm{R}(\Sigma) - vec(A)vec(B)^T\|_F.$$

Proof. Using the notations and the reshaping function R from Section 2.4.2:

$$\begin{aligned}
\|\Sigma - A \otimes B\|_F^2 &= \sum_{j=1}^{d_s}\sum_{i=1}^{d_s} \|\mathrm{vec}(\Sigma_{[i,j]}) - a_{ij}\mathrm{vec}(B)\|_2^2 \\
&= \sum_{j=1}^{d_s}\sum_{i=1}^{d_s} \|\mathrm{vec}(\Sigma_{[i,j]})^T - a_{ij}\mathrm{vec}(B)^T\|_2^2 \\
&= \sum_{j=1}^{d_s} \|\Sigma_j - A_{[j]}\mathrm{vec}(B)^T\|_F^2 \\
&= \|\mathrm{R}(\Sigma) - \mathrm{vec}(A)\mathrm{vec}(B)^T\|_F^2.
\end{aligned}$$

□

The "rank-one" approximation problem can be solved with SVD.

Corollary 3.2. *[10, Corollary 2.2] If the matrix Σ with same properties has SVD as*

$$U^T \mathrm{R}(\Sigma) V = \Delta = diag(\sigma(\mathrm{R}(\Sigma)))$$

then the matrices A and B minimizing $\|\Sigma - A \otimes B\|_F$ are $vec(A) = \sigma_1 \mathbf{u}_1$ and $vec(B) = \sigma_1 \mathbf{v}_1$, where σ_1 is the largest singular value of $\mathrm{R}(\Sigma)$, $\mathbf{u}_1$ and $\mathbf{v}_1$ are the corresponding singular vectors.

With the above results, we use SVD to get a rank-one approximation of a given matrix. Generalizing the KP representation to KP series expansion:

Theorem 3.3. *[10] Let $\Sigma = \sum_{i=1}^{r} A_i \otimes B_i$, with the previous properties. Then the following holds*

$$\mathrm{R}(\Sigma) = \sum_{i=1}^{r} \mathrm{R}(A_i \otimes B_i) = \sum_{i=1}^{r} vec(A_i) vec(B_i)^T.$$

The matrix $\mathrm{R}(\Sigma)$ is a rank-r matrix, thus we get $A_i = \sigma_i \mathbf{u}_i$ and $B_i = \sigma_i \mathbf{v}_i$ for $i = 1, \ldots, r$. In this case, the minimizer is $K = A \otimes B$ where A and B are linear combinations of A_i and B_i, i.e.

$$A = k_1 A_1 + \cdots + k_r A_r$$

$$B = p_1 B_1 + \cdots + p_r B_r.$$

With this generalization, we reformulate the covariance matrix approximation problem to a rank-r approximation problem. Also, adding a penalty term gives us a convex relaxation. For simplicity in notation, let $\hat{M}_n := \mathrm{R}(\hat{\Sigma}_n)$ and $K := \sum_{i=1}^{r} \mathrm{R}(A_i \otimes B_i)$ with $\dim(A_i) = d_s \times d_s$ and $\dim(B_i) = d_t \times d_t$. Then the proposed reformation is

$$\hat{K}_n^{\lambda} = \underset{K \in \mathbb{R}^{d_s^2 \times d_t^2}}{\arg\min} \|\hat{M}_n - K\| + \lambda \|K\|_{\star}$$

where $\lambda > 0$ is the penalization parameter. This problem has a closed solution, and it can be obtained by solving a thresholded SVD

$$\hat{K}_n^{\lambda} = \sum_{i=1}^{\min(d_s^2, d_t^2)} \left(\sigma_i(\hat{M}_n) - \frac{\lambda}{2} \right)_+ \mathbf{u}_i \mathbf{v}_i^T.$$

Here we are making a singular value decomposition on $\hat{M}_n$ with $\hat{M}_n = U \Delta V$ and $\mathbf{u}_i$ and $\mathbf{v}_i$ are the i'th columns of the matrices U and V, respectively.

Before moving to properties of the estimator, below we summarize the procedure:

Algorithm 1 The PRLS Algorithm

Calculate the SCM: $\hat{\Sigma}_n = \frac{1}{n}\sum_{\alpha=1}^{n} \mathbf{v}_\alpha \mathbf{v}_\alpha^T$.

Reshape SCM: $\hat{M}_n := \mathrm{R}(\hat{\Sigma}_n)$

State the penalized minimization problem: $\hat{K}_n^\lambda = \min_{K \in \mathbb{R}^{d_s^2 \times d_t^2}} \|\hat{M}_n - K\| + \lambda \|K\|_\star$

Solve the problem with SVT: $\sum_{i=1}^{\min(d_s^2, d_t^2)} \left(\sigma(\hat{M}_n) - \frac{\lambda}{2}\right)_+ u_i v_i^T$

Inverse the reshaping: $\hat{\Sigma}_n^\lambda := \mathrm{R}^{-1}(\hat{K}_n^\lambda)$

$\hat{\Sigma}_n^\lambda$ is the estimation.

3.2.1. PRLS Preserving Necessary Conditions

The inverse reshaped matrix is the estimation of the covariance matrix. However, we need to show that it is a covariance matrix, i.e., it is symmetric and positive definite.

Theorem 3.4. *[11, Theorem 1] Let $\hat{\Sigma}_n^\lambda := \mathrm{R}^{-1}(\hat{K}_n^\lambda)$ denote the inverse reshaped matrix. Then,*

- *$\hat{\Sigma}_n^\lambda$ is symmetric with probability* 1.
- *When $n \geq d$, $\hat{\Sigma}_n^\lambda$ is a positive semi-definite matrix, with probability* 1.

3.2.2. Relationship between Risk Matrices

After showing that the PRLS procedure preserves the necessary conditions for the covariance matrix, it is time to show that the PRLS outperforms "the SCM estimator" in convergence rate. Note that the convergence rate of the SCM estimator is

$$\|\hat{\Sigma}_n - \Sigma\|_F^2 = O_P(\frac{d_s d_t}{n}).$$

Now we show a result that gives a relationship between the SCM risk norm and the PRLS risk norm.

Theorem 3.5. *[11, Theorem 2]*

Assuming $\lambda \geq \|\hat{M}_n - M\|_2$*, the below inequality holds*

$$\|\hat{K}_n^{\lambda} - M\|_F^2 \leq \inf_K \left\{ \|K - M\|_F^2 + \frac{(1+\sqrt{2})^2}{4}\lambda^2 rank(K) \right\}.$$

This result will be useful when setting a norm bound on the PRLS estimation error in Section 3.2.4.

3.2.3. Bound on the SCM Estimation Error

Let $D_n = \hat{M}_n - M = \mathrm{R}(\hat{\Sigma}_n - \Sigma)$. One can see that, $D_n \to 0$ a.s. as $n \to \infty$. This comes from the strong law of large numbers after fixing the spatial and temporal dimensions. In addition to that, a bound for $\|D_n\|_2$ can be set with the next theorem.

Theorem 3.6 (Operator Norm Bound on Reshaped SCM [11, Theorem 3])**.** *Assume* $\|\Sigma\|_2 < \infty$ *for all* d_s, d_t *and define* $N = \max(d_s, d_t, n)$*. Fix* $\varepsilon' < 1/2$*. Assume* $t \geq \max(\sqrt{4C_1 \ln(1 + \frac{2}{\varepsilon'})}, 4C_2 \ln(1 + \frac{2}{\varepsilon'}))$ *and* $C = \max(C_1, C_2) > 0$*. Then, with probability at least* $2N^{-\frac{t}{4C}}$*,*

$$\|D_n\|_2 \leq \frac{Ct}{1-2\varepsilon'} \max\{\frac{d_s^2 + d_t^2 + \log N}{n}, \sqrt{\frac{d_s^2 + d_t^2 + \log N}{n}}\}.$$

Before the proof, we need to prove a useful lemma.

Lemma 3.7 (Concentration of Measure for Coupled Gaussian Chaos [11, Lemma 2])**.** *Let* $\mathbf{x} = [x_1, \ldots, x_{d_s^2}]^T \in \mathcal{S}_{d_s^2-1}$ *and* $\mathbf{y} = [y_1, \ldots, y_{d_t^2}]^T \in \mathcal{S}_{d_t^2-1}$*. In the Sample Covariance Matrix notation, assume that* $\{\mathbf{v}_\alpha\}$ *are i.i.d. observations of multivariate normal random variable* $\mathbf{v} \sim \mathcal{N}(0, \Sigma)$*. Then for all* $\varepsilon \geq 0$

$$\mathbb{P}(\left|\mathbf{x}^T D_n \mathbf{y}\right|) \leq 2\exp\left(\frac{-n\varepsilon^2/2}{C_1\|\Sigma\|_2^2 + \varepsilon C_2 \|\Sigma\|_2}\right)$$

where $C_1 = \frac{4e}{\sqrt{6\pi}} \approx 2.5044$ *and* $C_2 = e\sqrt{2} \approx 3.8442$ *are absolute constants.*

Proof. Following the latest representation of D_n, we can write $\mathbf{x}^T D_n \mathbf{y}$ as

$$\mathbf{x}^T D_n \mathbf{y} = \frac{1}{n}\sum_{\alpha=1}^{n} \omega_\alpha,$$

where

$$\omega_k = \left(\sum_{i,j=1}^{s} \sum_{k,l=1}^{t} X_{i,j} Y_{k,l} \right) \left([\mathbf{v}_t]_{(i-1)t+k} [\mathbf{v}_t]_{(j-1)t+l} - \mathbb{E}\left[[\mathbf{v}_t]_{(i-1)t+k} [\mathbf{v}_t]_{(j-1)t+l} \right] \right)$$

with $X := \mathrm{R}(\mathbf{x})$ and $Y := \mathrm{R}(\mathbf{y})$.

After defining $K = X \otimes Y$, we get

$$\omega_\alpha = \mathbf{v}_\alpha^T K \mathbf{v}_\alpha - \mathbb{E}\left[\mathbf{v}_\alpha^T K \mathbf{v}_\alpha \right],$$

which is very difficult to analyze because of the correlated variables. Thus, we use "joint Gaussian property of the data" to simplify it.

As the stochastic equivalent of $\mathbf{v}_\alpha^T K \mathbf{v}_\alpha$ is $\mathbf{b}_\alpha^T \tilde{K} \mathbf{b}_\alpha - \mathbb{E}[\mathbf{b}_\alpha^T \tilde{K} \mathbf{b}_\alpha]$, where $\mathbf{b}_\alpha \sim \mathcal{N}(0, I_{d_s d_t})$ $\tilde{K} = \Sigma^{\frac{1}{2}} K \Sigma^{\frac{1}{2}}$. Note that this just changes v_αs with $N(0, I_{st})$ random variables and adds the mean to kronecker product part of the bilinear representation. Then these equalities and inequalities follow:

$$\begin{aligned}
\mathbb{E}|\omega_\alpha|^2 &= \mathbb{E}\left| b_\alpha^T \tilde{K} b_\alpha - \mathbb{E}[b_\alpha^T \tilde{K} b_\alpha] \right|^2 \\
&= \mathbb{E}\left| \sum_{i_1 \neq i_2} [b_\alpha]_{i_1} [b_\alpha]_{i_2} \tilde{K}_{i_1,i_2} + \sum_{i=1}^{d} ([b_\alpha]_{i_1}^2 - 1) \tilde{K}_{i_1,i_1} \right|^2 \\
&= \sum_{i_1 \neq i_2} \sum_{i_1' \neq i_2'} \mathbb{E}\left[[b_\alpha]_{i_1} [b_\alpha]_{i_2} [b_\alpha]_{i_1'} [b_\alpha]_{i_2'} \right] \tilde{K}_{i_1,i_2} \tilde{K}_{i_1',i_2'} \\
&+ \sum_{i_1} \sum_{i_1'} \mathbb{E}\left[([b_\alpha]_{i_1}^2 - 1)([b_\alpha]_{i_1'^2} - 1) \right] \tilde{K}_{i_1,i_1} \tilde{K}_{i_1',i_1'} \\
&= \sum_{i_1 \neq i_2} \tilde{K}_{i_1,i_2}^2 + 2 \sum_{i_1} \tilde{K}_{i_1,i_1}^2 \\
&= \|\tilde{K}\|_F^2 + \|diag(\tilde{K})\|_F^2 \\
&\leq 2\|\tilde{K}\|_F^2 \leq 2\|\Sigma\|_2^2 \|K\|_F^2 = 2\|\Sigma\|_2^2.
\end{aligned}$$

Note that for the last step we have used $\|K\|_F = \|X\|_F \|Y\|_F = 1$.

Now using a well known moment bound on Gaussian chaos [18, page 65] $\mathbb{E}|Z|^q \leq (q-1)^q (\mathbb{E}|Z|^2)^{q/2}$ and Stirling's formula [19] $q! = \sqrt{2\pi q} q^q e^{-q} e^{R_q}$ with $(12q+1)^{-1} \leq R_q \leq (12q)^{-1}$, we can show that, for all $k \geq 3$:

$$\mathbb{E}|\omega_\alpha|^k \leq \frac{k! W^{k-1} c_\alpha}{2},$$

where

$$W = e\sqrt{\mathbb{E}|\omega_\alpha|^2} \leq e\sqrt{2}\|\Sigma\|_2$$
$$c_\alpha = \frac{2e}{6\pi}\mathbb{E}|\omega_\alpha|^2 \leq \frac{4e}{\sqrt{6\pi}}\|\Sigma\|_2^2.$$

Before concluding our proof, recall one of the Bernstein's inequilites:

Proposition 3.8. *[20] For $\{X_i\}$ independent zero-mean random variables. Suppose $\|X_i\| \leq M$ almost surely, for all i. Then, for all $\varepsilon > 0$,*

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} X_i\right| \geq \varepsilon\right) \leq 2\exp\left(\frac{n\varepsilon^2}{2 + 2\varepsilon/3}\right). \tag{3.1}$$

Directly from the Bernstein's inequality, we can conclude the proof of the lemma:

$$\begin{aligned}\mathbb{P}\left(\left|\frac{1}{n}\sum_{\alpha=1}^{n}\omega_\alpha\right| \geq \varepsilon\right) &\leq 2\exp\left(\frac{-n^2\varepsilon^2/2}{nu_1 + Wn\varepsilon}\right) \\ &\leq 2\exp\left(\frac{-n\varepsilon^2/2}{C_1\|\Sigma\|_2^2 + C_2\|\Sigma\|_2\varepsilon}\right).\end{aligned}$$

□

Now we can start the proof of the Theorem 3.6.

Proof. Let $N_{d_s^2}$ and $N_{d_t^2}$ be ε'-nets on unit spheres $\mathbf{S}_{d_s^2-1}$ and $\mathbf{S}_{d_t^2-1}$, respectively. Let $\mathbf{u}_1$ and $\mathbf{v}_1$ be unit vectors from $\mathbf{S}_{d_s^2-1}$ and $\mathbf{S}_{d_t^2-1}$, respectively such that $|\mathbf{u}_1^T D_n \mathbf{v}_1| = \|D_n\|_2$. By definition of ε'-net, we can find unit vectors $\mathbf{u}_2$ and $\mathbf{v}_2$ from the corresponding spheres such that $\|\mathbf{u}_1 - \mathbf{u}_2\|_2 \leq \varepsilon'$ and $\|\mathbf{v}_1 - \mathbf{v}_2\|_2 \leq \varepsilon'$. Using that, we can write the below inequality:

$$\begin{aligned}|\mathbf{u}_1^T D_n \mathbf{v}_1| - |\mathbf{u}_2^T D_n \mathbf{v}_2| &\leq |\mathbf{u}_1^T D_n \mathbf{v}_1 - \mathbf{u}_2^T D_n \mathbf{v}_2| \\ &\leq |\mathbf{u}_1^T D_n(\mathbf{v}_1 - \mathbf{v}_2) + (\mathbf{u}_1 - \mathbf{u}_2)^T D_n \mathbf{v}_2| \\ &\leq |\mathbf{u}_1^T D_n(\mathbf{v}_1 - \mathbf{v}_2)| + |(\mathbf{u}_1 - \mathbf{u}_2)^T D_n \mathbf{u}_2| \\ &\leq |\mathbf{u}_1^T||\mathbf{v}_1 - \mathbf{v}_2|\|D_n\|_2 + |(\mathbf{u}_1 - \mathbf{u}_2)^T||\mathbf{v}_2|\|D_n\|_2 \qquad (3.2) \\ &\leq 2\varepsilon'\|D_n\|_2.\end{aligned}$$

The inequality (3.2) comes from the definition of spectral norm, and the fact that $\mathbf{u}_1$ and $\mathbf{v}_2$ being unit vectors.

Now, swapping $|\mathbf{u}_1^T D_n \mathbf{v}_1|$ with $\|D_n\|_2$, we get

$$\begin{aligned} \|D_n\|_2 - |\mathbf{u}_2^T D_n \mathbf{v}_2| &\leq 2\varepsilon' \|D_n\|_2 \\ (1-2\varepsilon')\|D_n\|_2 &\leq |\mathbf{u}_2^T D_n \mathbf{v}_2| \end{aligned}$$

for any $\mathbf{u}_2$ and $\mathbf{v}_2$ satisfying above conditions. Then we can generalize to

$$\begin{aligned} (1-2\varepsilon')\|D_n\|_2 &\leq \max\{|\mathbf{u}_2^T D_n \mathbf{v}_2| : \mathbf{u}_2 \in N_{d_s^2}, \mathbf{v}_2 \in N_{d_t^2}, \|\mathbf{u}_1 - \mathbf{u}_2\|_2 \leq \varepsilon', \|\mathbf{v}_1 - \mathbf{v}_2\|_2 \leq \varepsilon'\} \\ (1-2\varepsilon')\|D_n\|_2 &\leq \max\{|\mathbf{u}^T D_n \mathbf{v}| : \mathbf{u} \in N_{d_s^2}, \mathbf{v} \in N_{d_t^2}\} \\ \|D_n\|_2 &\leq (1-2\varepsilon')^{-1} \max_{\mathbf{u} \in N_{d_s^2}, \mathbf{v} \in N_{d_t^2}} |\mathbf{u}^T D_n \mathbf{v}|. \end{aligned}$$

Note that, above we just generalized to the whole set, it is an obvious inequality.

Also using a bound on cardinality [14]

$$\mathrm{card}(N_{d_s^2}) \leq \left(1 + \frac{2}{\varepsilon'}\right)^{d_s^2}$$

we get

$$\begin{aligned} \mathbb{P}(\|D_n\|_2 \leq \varepsilon) &\leq \mathbb{P}(\max_{\mathbf{u} \in N_{d_s^2}, \mathbf{v} \in N_{d_t^2}} |\mathbf{u}^T D_n \mathbf{v}| \geq \varepsilon'(1-2\varepsilon')) \\ &\leq \mathbb{P}(\bigcup_{\mathbf{u} \in N_{d_s^2}, \mathbf{v} \in N_{d_t^2}} |\mathbf{u}^T D_n \mathbf{v}| \geq \varepsilon'(1-2\varepsilon')) \\ &\leq \mathrm{card}(N_{d_s^2})\mathrm{card}(N_{d_t^2}) \times \max_{\mathbf{u} \in N_{d_s^2}, \mathbf{v} \in N_{d_t^2}} \mathbb{P}(|\mathbf{u}^T D_n \mathbf{v}| \geq \varepsilon'(1-2\varepsilon') \\ &\leq \left(1 + \frac{2}{\varepsilon'}\right)^{d_s^2 + d_t^2} \mathbb{P}(|\mathbf{u}^T D_n \mathbf{v}| \geq \varepsilon(1-2\varepsilon')). \end{aligned}$$

Then, using Lemma 3.7, we move further

$$\mathbb{P}(D_n \geq \varepsilon) \leq 2\left(1 + \frac{2}{\varepsilon'}\right)^{d_s^2 + d_t^2} \exp\left(\frac{-n\varepsilon^2(1-2\varepsilon')^2/2}{C_1\|\Sigma\|_2^2 + C_2\|\Sigma\|_2\varepsilon(1-2\varepsilon')}\right).$$

Now, we need to consider two different cases: Gaussian tail and exponential tail.

Assume Gaussian tail properties hold, *i.e.* $\varepsilon \leq \frac{c_1\|\Sigma\|_2}{C_2(1-2\varepsilon')}$. In this case, we can relax the bound to

$$\mathbb{P}(\|D_n\|_2 \geq \varepsilon) \leq 2\left(1 + \frac{2}{\varepsilon'}\right)^{d_s^2 + d_t^2} \times \exp\left(\frac{-n\varepsilon^2(1-2\varepsilon')/2}{2C_1\|\Sigma\|_2^2}\right).$$

Then, choosing $\varepsilon = \frac{t\|\Sigma\|_2}{1-2\varepsilon'}\sqrt{\frac{d_s^2+d_t^2+\log N}{n}}$, we get

$$\begin{aligned}\mathbb{P}\left(\|D_n\|_2 \geq \frac{t\|\Sigma\|_2)}{1-2\varepsilon'}\sqrt{\frac{d_s^2+d_t^2+\log N}{n}}\right) &\leq 2\left(1+\frac{2}{2\varepsilon'}\right)^{d_s^2+d_t^2}\exp\left(\frac{-t^2(d_s^2+d_t^2+\log N)}{4C_1}\right)\\ &\leq 2\left(\left(1+\frac{2}{\varepsilon'}\right)e^{-t^2/(4C_1)}\right)^{d_s^2+d_t^2}N^{-t^2/(4C_1)}\\ &\leq 2N^{-t^2/(4C_1)}.\end{aligned}$$

Gaussian tail case is proven. Now let's assume that the tail is exponential, *i.e.* $\varepsilon \geq \frac{C_1\|\Sigma\|_2}{C_2(1-2\varepsilon')}$, and then set

$$\varepsilon = \frac{t\|\Sigma\|_2}{1-2\varepsilon'}\frac{d_s^2+d_t^2+\log N}{n}.$$

With that ε, we finalize our proof:

$$\begin{aligned}\mathbb{P} &\left(\|D_n\|_2 \geq \frac{t\|\Sigma\|_2}{1-2\varepsilon'}\frac{d_s^2+d_t^2+\log N}{n}\right)\\ &\leq 2\left(1+\frac{2}{\varepsilon'}\right)^{d_s^2+d_t^2}\exp\left(\frac{-t(d_s^2+d_t^2+\log N)}{4C_2}\right)\\ &\leq 2\left(\left(1+\frac{2}{\varepsilon'}\right)e^{\frac{-t}{4C_2}}\right)N^{-\frac{t}{4C_2}}\\ &\leq 2N^{-\frac{t}{4C_2}}.\end{aligned}$$

Note that, we have used the assumption $t \geq 4C_2\ln(1+\frac{2}{\varepsilon'})$. After combining both cases and letting $C = \|\Sigma\|_2$, we complete the proof. Also, note that $t > 1$ and $\frac{tC_2}{C_1} > 1$. □

3.2.4. Bound on the PRLS Estimation Error

Now we have bound on the reshaped PRSL estimation error from Theorem 3.5 and another bound on the reshaped SCM estimation error from Theorem 3.6. Using these two results we can set a bound on the risk of PRLS estimator.

Theorem 3.9 (Frobenius Norm Bound on Estimation Error [11, Theorem 4])**.** *Define* $N = \max(d_s, d_t, n)$. *Set* $\lambda = \lambda_n = \frac{2Ct}{1-2\varepsilon'}\max\{\frac{d_s^2+d_t^2+\log N}{n}, \sqrt{\frac{d_s^2+d_t^2+\log N}{n}}\}$ *with* t *satisfying the conditions of Theorem 3.6. Then, with probability at least* $1-2N^{-\frac{t}{4C}}$

$$\|\hat{\Sigma}_n^{\lambda} - \Sigma\|_F^2 \leq \inf_{K:rank(K)\leq r}\|K-M\|_F^2$$

$$+C'r\ max\left\{\left(\frac{d_s^2+d_t^2+\log N}{n}\right)^2, \frac{d_s^2+d_t^2+\log N}{n}\right\},$$

where

$$C' = \left(Ct\frac{1+\sqrt{2}}{1-2\epsilon'}\right)^2 = \left(3(1+\sqrt{2})Ct\right)^2 > 0.$$

Proof. Define the event:

$$E_r = \left\{\|\hat{K}_n^\lambda - M\|_F^2 > \inf_{K:\text{rank}(K\leq r)} \|K-M\|_F^2 + \frac{(1+\sqrt{2})^2}{4}\lambda_n^2 r\right\}.$$

Theorem 3.5 implies that on the event $\lambda_n \geq 2\|D_n\|_2$, with probability 1, for any $1 \leq r \leq r_0$ we have

$$\|\hat{K}_n^\lambda - M\|_F^2 \leq \inf_{K:\text{rank}(K)\leq r} \|K-M\|_F^2 + \frac{(1+\sqrt{2})^2}{4}\lambda_n^2 r.$$

Using the above inequality and Theorem 3.6, we get

$$\begin{aligned}
\mathbb{P}(E_r) &= \mathbb{P}(E_r \cap \{\lambda_n \geq 2\|D_n\|_2\}) + \mathbb{P}(E_r \cap \{\lambda_n < 2\|D_n\|_2\}) \\
&\leq \mathbb{P}(E_r|\lambda \geq 2\|D_n\|_2)\mathbb{P}(\lambda_n \geq 2\|D_n\|_2) + \mathbb{P}(\lambda_n < 2\|D_n\|_2) \\
&= \mathbb{P}(\|D_n\|_2 > \frac{Ct}{1-2\varepsilon'} \times \max\left\{\frac{d_s^2+d_t^2+\log N}{n}, \sqrt{\frac{d_s^2+d_t^2+\log N}{n}}\right\} \\
&\leq 2N^{-t/4C}.
\end{aligned}$$

□

To summarize, we showed some models motivated by the well-known Eckart-Young theorem [15]. Then presented Van Loan's [10] method for reducing the Eckart-Young problem to a rank-one minimization problem. After that, we described Hero's [11] convex-relaxed improving to Van Loan's results, in detail. We have presented the necessary propositions for the penalized low-rank estimation to give a mathematically valid covariance estimation. Following that, we have analyzed the relationship between risk matrices of argued methods. Then, we gave proof for an operator norm bound on the reshaped SCM estimation. Finally, we have shown a Frobenius norm for the PRLS estimation error and proved that the norm converges with a known high probability.

3.3. Temporally Reinforced Kronecker Factorization

In wind analysis and many other Spatio-temporal topics, it is possible to find different settings with similar temporal characteristics. Also, even though there is a huge amount of data being observed for a long time, many new sensors are still being placed every day. In addition, there sometimes may be problems with the continuity of the data, which leaves us with only a small number of observations. However, we may have prior knowledge about the data, and we want to improve our model with this knowledge. We consider using previously analyzed temporal properties of an existing data to reinforce another data with fewer observations to make better covariance estimations. We named our method Temporally Reinforced Kronecker Factorization (TRKF).

Before explaining TRKF, we first assume the existence of the separable true covariance matrix. Then, we present a new covariance estimate which we call Spatio-temporally Decomposed Kronecker Product (SDKP). For SDKP, we first decompose the data sample to spatial and temporal data. The spatial data contains only the spatial features (e.g., locations) as variables and the temporal data contains features as time indices. The spatial data is straightforward, but the temporal data is actually an estimate representing the temporal relationship of the data. As an example let the data X that we are interested in contains the features of the form $L_i^{(t_j)}$ for $i = 1, \ldots, p$ and $j = 1, \ldots, q$. Then the spatial data will contain the features L_i for $i = 1, \ldots, p$ and the temporal data will contain $t_0, \ldots, t_q$ as features. After this decomposition, we calculate the SCM of both parts and take their Kronecker product as our SDKP covariance estimate. We argue that this estimation is a little smoother version of the SCM and more robust to outliers. However, unlike previously discussed methods, we do not expect this method to perform well under a small sample size.

Now we can define TRKF, assume that the true covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$ is separable, i.e. $\Sigma = A \otimes B$, for $A \in \mathbb{R}^{d_s \times d_s}$ and $B \in \mathbb{R}^{d_t \times d_t}$, with $d = d_s d_t$. Then we argue that, when the number of observations is low, instead of using the SCM, we may reinforce the SDKP model with an externally obtained temporal covariance

matrix. In SDKP, we calculate the temporal and spatial covariance matrices, then we switch or modify the temporal covariance matrix with another temporal covariance matrix obtained by the same method from another sample that carries similar temporal characteristics.

4. EXPERIMENTATION

We compare the SCM and Kronecker-based methods on synthetic simulations and real-world applications. We calculate the error matrices for the true covariance matrix for comparison. Also, we observe the explained variance of the components for each approximation in detail.

4.1. Simulation

4.1.1. Data Generation

For simulation, we generate two types of true covariance matrices of the form $\Sigma = \sum_{i=1}^{3} A_i \otimes B_i$ with dimension $d = 600$ where $d_s = 30$ and $d_t = 20$. The first matrices have their Kronecker factors with no appointed structure other than being symmetric. We randomly generate matrices X with i.i.d entries from $\mathcal{N}(0, 1)$ distribution then create the positive semi-definite Kronecker factors by setting $A_i = XX^T$. We generate the second type of matrices similarly; the only difference is one of the Kronecker factors (B_i) being Toeplitz. We call these matrices of type 1 and type 2 throughout this section.

We generate 10 different matrices of each type. Then for each matrix, we generate 5 different random data of size 500. For each $n \in \{5, 10, 20, 40, 60, 80, 100\}$, we pick 3 random subsamples from each data.

4.1.2. Covariance Estimations

We first calculate the SCM for each subsample. Then, we reshape the SCM and do partial SVD and SVT to calculate the R1LS and the PRLS, respectively. We measure the performance of each estimation with the Frobenius norm and matrix-1 norm. Also, we observe the explained variance of the PRLS estimation and compare it with the principal components obtained from the SCM.

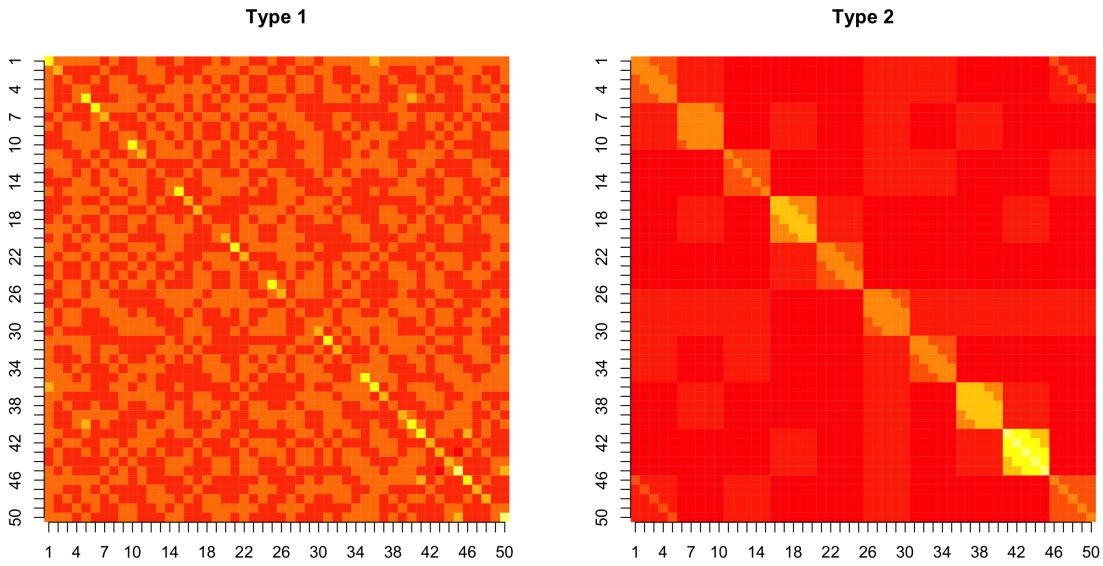

Figure 4.1. A visual representation of the true covariance matrices. The first one is a type-1 matrix that has symmetric Kronecker factors. The second one has a Toeplitz Kronceker factor. The yellow color indicates a higher value.

Comparing the true covariance matrices in Figure 4.1, we can see that having a Toeplitz factor makes the matrix much more structured. With this structure, it is possible to explain most of the variance with only one component.

In Figure 4.2, we can observe that the approximations made with Kronecker factors are more accurate than the classical approach. In general, the PRLS and the R1LS do not dominate each other's performance for $n < 40$. However, for $n > 40$, the PRLS almost consistently outperforms the R1LS. Both of these methods always give better estimates than the SCM. Also, we have tried different methods for generating random matrices, and they gave similar results, with some having the PRLS dominance for all n, but we did not include them here.

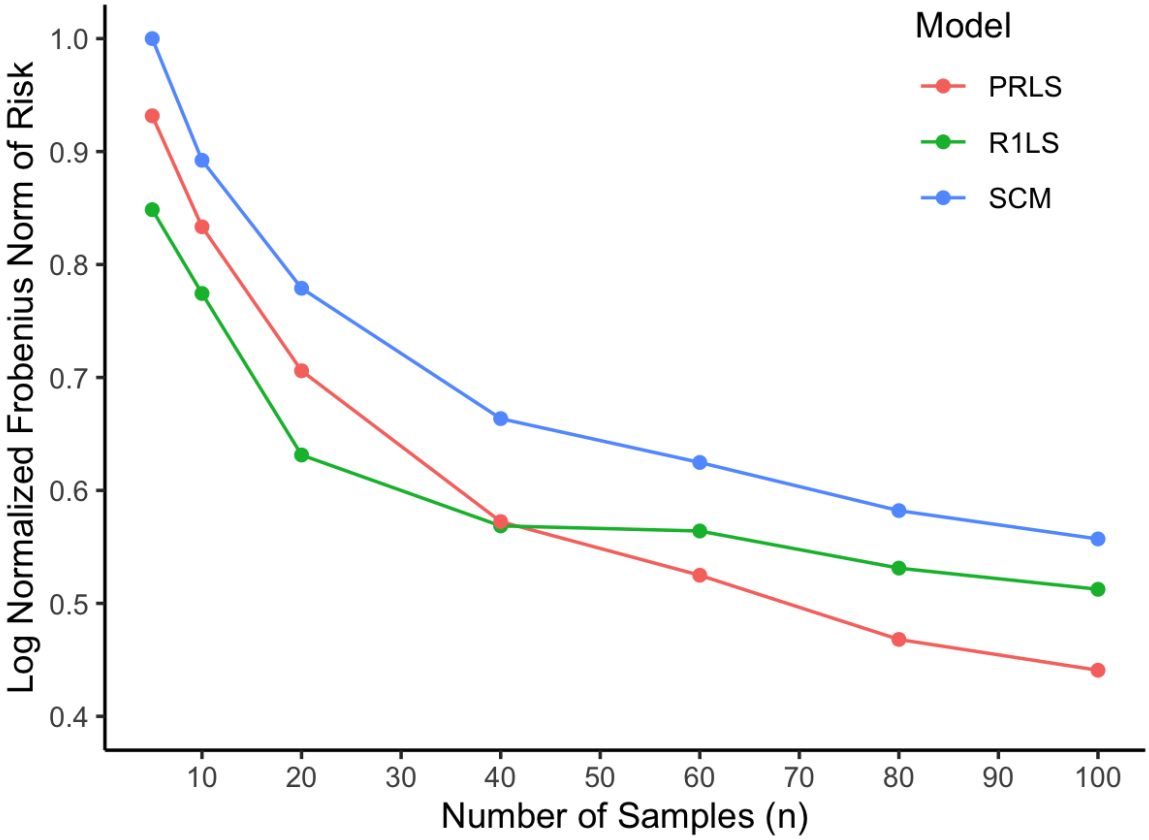

Figure 4.2. A simulation result example. The Frobenius norm comparison of estimations for a type-1 matrix with $d_s = 30, d_t = 20$. The Frobenius norms of the risks are scaled after log-normalizing for visual easiness.

In Figure 4.3, we see that almost all of Kronecker's spectrum energy is contained in the first three components, with the first component carrying around 67% of the variance. The energy of the eigenspectrum is much more spread, with the first PC carrying only 1% of the energy. With this picture in mind, we again underline Kronecker factorization's power. Kronecker-based models not only make a better estimate, but they also have a low rank compared to the classical approach. This advantage grows even more in the type 2 matrices. To sum up, with reshaping, we were able to extract most of the variance in the least amount of PCs.

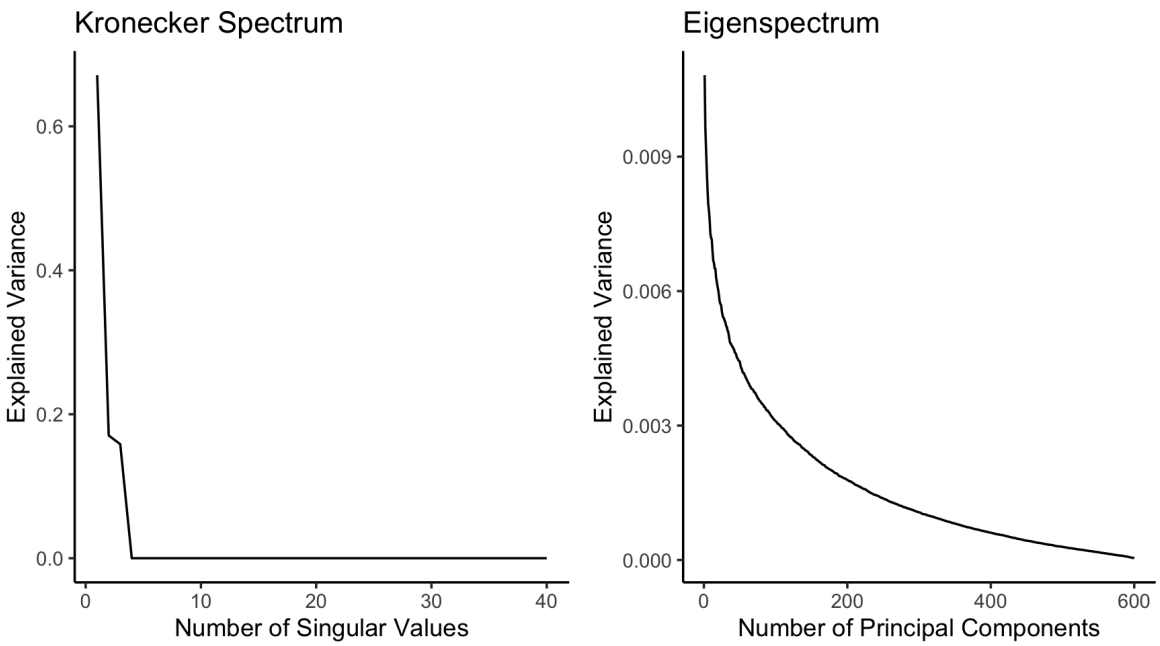

Figure 4.3. Explained variance of the PRLS and the eigenspectrum of the true covariance matrix.

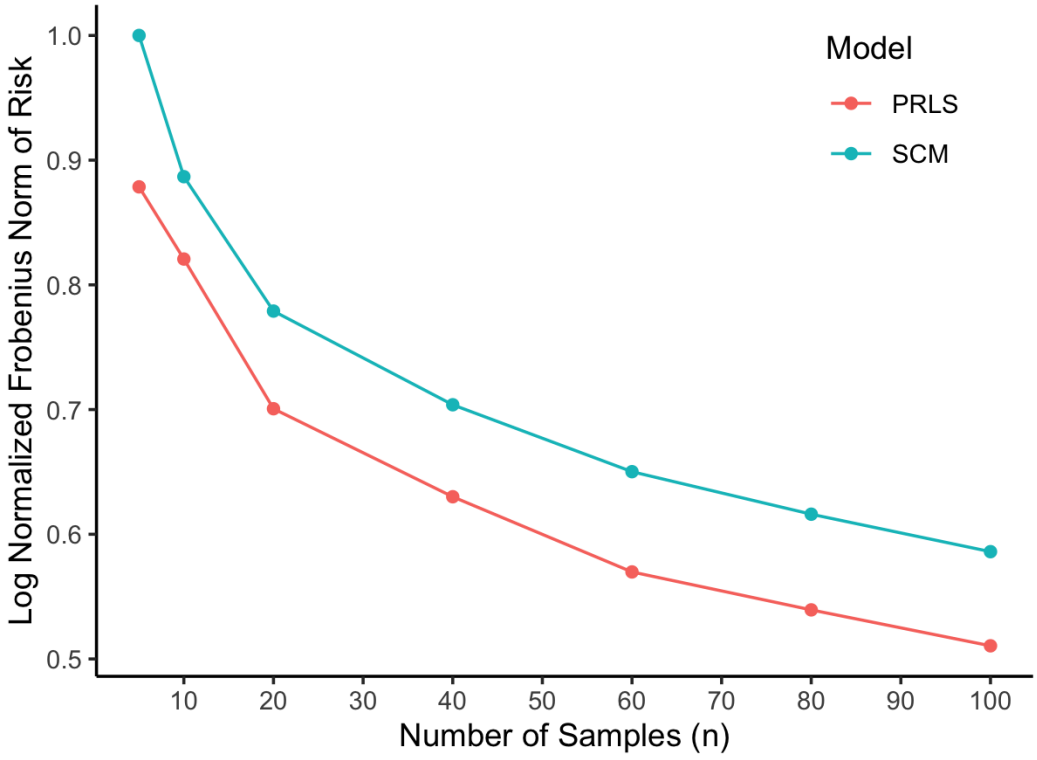

Figure 4.4. A simulation result example for a type-2 matrix. The Frobenius norm of the risks are log-normalized and scaled for visual easiness.

In Figure 4.4, For matrices of type 2, we again observe better approximations for all n. Observing Figure 4.5, we see the Kronecker spectrum becoming even denser with a Toeplitz factor. This was expected as we got a repetitive structure in the covariance matrix with a Toeplitz factor. As the first component contains almost the entire spectrum energy, approximations made with the PRLS and the R1LS do not differ significantly.

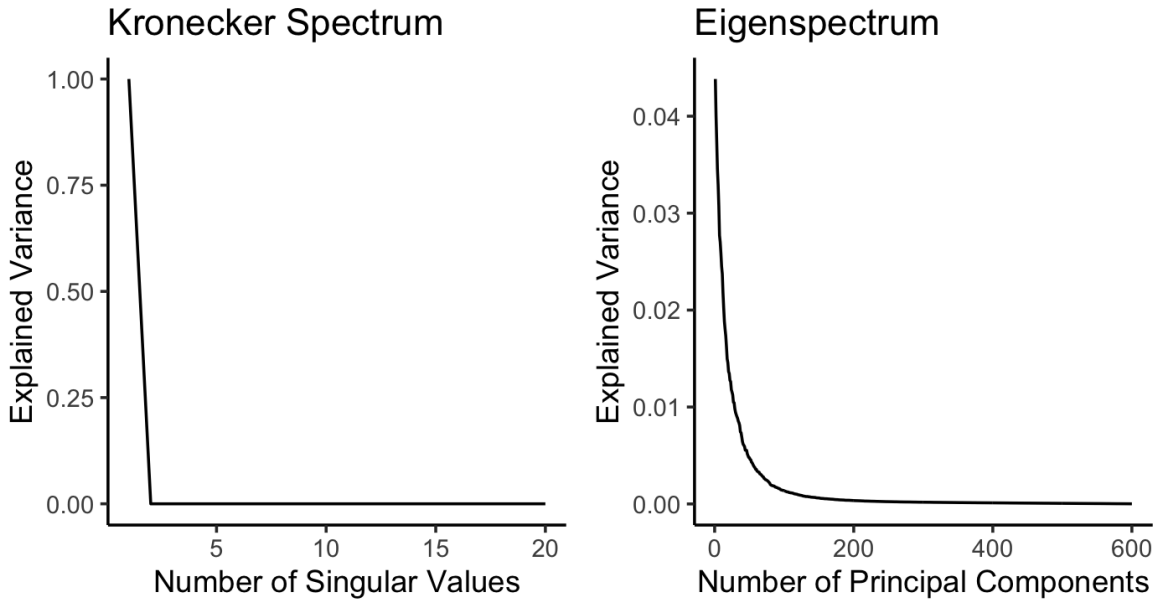

Figure 4.5. Explained variance comparison for a type-2 matrix.

We ran 150 simulations for each matrix type and every n. Almost all of them had similar results that verified the figures. We confirmed that better approximations could be made with the Kronecker structure even with a tiny number of samples. Also, we observed that Kronecker-based models could explain the data with significantly fewer components. Having a Toeplitz matrix as a Kronecker factor creates a more structured matrix. We regard this situation when dealing with real-world applications, too. Thus this improvement for the Toeplitz matrices is valuable for us.

Also, as a small note, in almost all of our simulations, we observe that instead of doing SVT with tuning λ, doing SVD with tuning expected r performs slightly better. Also, it will be more practical than tuning λ if the data is not well-prepared.

4.2. Wind Speed

4.2.1. Data

We use wind speed data from 6 different regions on the western side of Turkey to forecast wind-based electricity production from wind turbines in these regions [21]. The regions of the sensors and the turbines are: Aliağa, Bares, Dinar, Geycek, Söke and Soma. We will number them from station 1 to 6, alphabetically. Hourly observations are obtained from each station and turbine. The total number of sensors of the stations vary; Station-1 has nine sensor on a three-by-three grid, the stations numbered from 2 to 5 have 16 sensors on four-by-four grids, and the last station has 12 sensors on a three-by-four grid. Also, we used a time window of 7 hours long.

We process the raw data with the following steps: First, we combine u and v speed vectors to make a velocity scalar. Then, we scale the production values to fit between 0 and 1; we use logistic regression because wind turbines have a maximum production capacity. Finally, we transform the data into a Spatio-temporal format, also thinking A as the spatial and B as the temporal factor in $A \otimes B$ format.

4.2.2. Covariance Estimations

Here we analyze the covariance approximations. From the law of large numbers, we expect the true covariance matrix to be closely approximated by the SCM for $n = 36000$. We assume this matrix represents the true covariance matrix. Then for different values of n between 10 and 200, we compare the approximations: SCM and the PRLS with calculating the Frobenius norm of the risk matrix. Also, we observe the explained variance of the the PRLS to verify our simulated results once more.

In Figure 4.6, we observe that the PRLS gives a better estimate in all stations. We tried different values for n between 10 and 200 and randomly picked 10 samples for each n. The figure shows the case when $n = 100$ but this results hold for every n and every random sample. Observing the explained variance in Figure 4.7, we see that the

PRLS approximations explained almost the full variance with one component. This also creates a great advantage for dimension reduction studies.

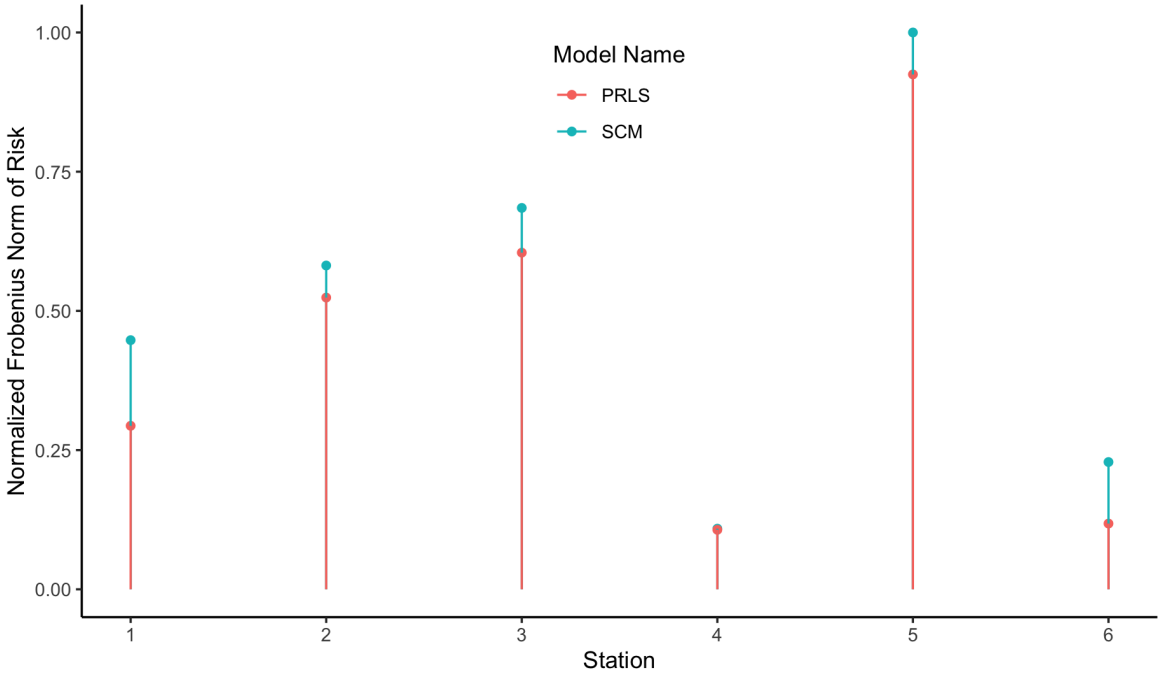

Figure 4.6. Comparison of the PRLS and the SCM models. We can see that the PRLS significantly outperforms the SCM in every station.

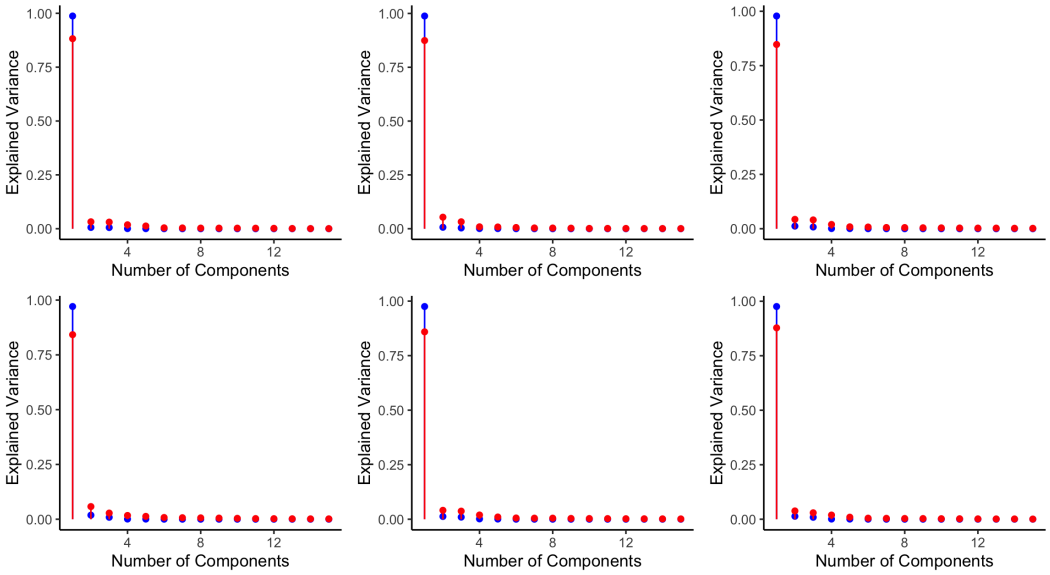

Figure 4.7. Explained variance comparison between the true covariance matrix and the PRLS. Blue color indicates the PRLS.

4.3. Temporally Reinforced Kronecker Factorization

In this section, we show that TRKF outperforms previously discussed models significantly. For experimentation, we have used the same wind data as the previous section, 6 different stations from Turkey. For the first experiment, we removed one location from each station to use it as the temporal covariance matrix generator with a big sample. We then applied TRKF model to the other locations with the removed location as a reinforcer. For the second experiment, we used one station's temporal covariance to reinforce another. For the final experiment, we used an old but large sample from the station to reinforce the same station.

We use n observations ranging from 10 to 200. For each n, we take 5 different random samples from the data. We tried different values of λ for the PRLS and included $r = 2$ in the final comparisons.

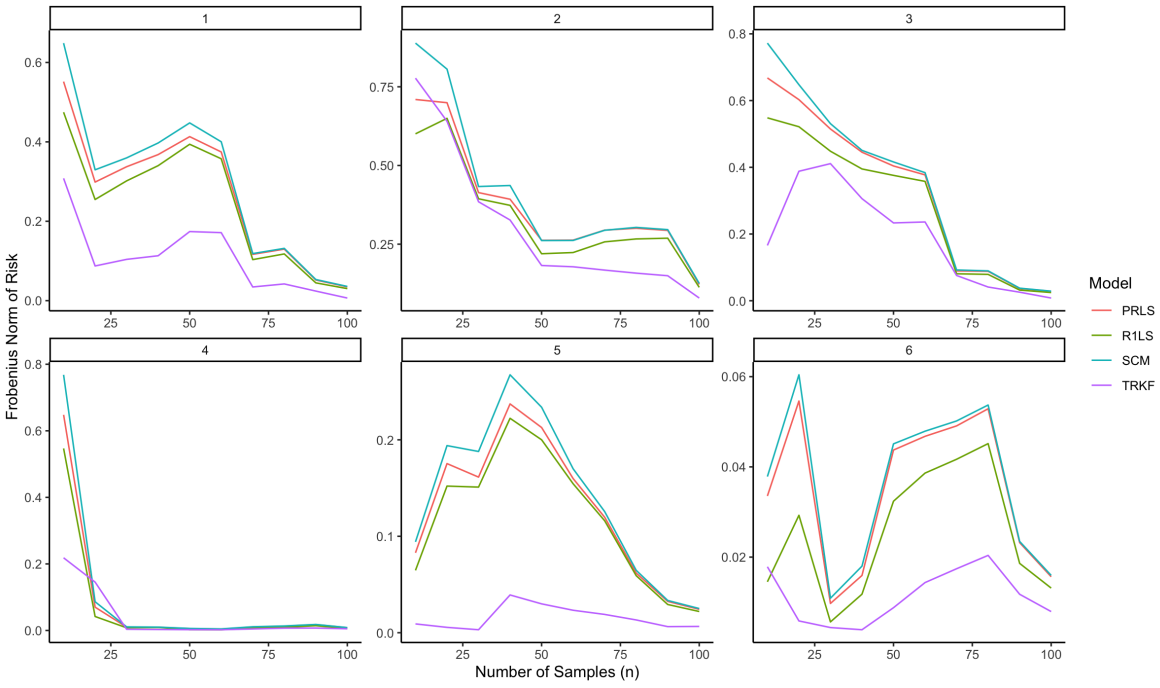

Figure 4.8. An example result from the first experiment. The numbers above the plots represent the corresponding stations.

Observing Figure 4.8, we can see that TRKF outperforms all other models, especially for a sample size case. We have observed similar results for almost every case. We were expecting these results as we could use a reasonable estimate of the temporal knowledge with a nearby station. We obtained even better results for the third experiment but did not include it because it was even more in favor of TRKF. However, for the second experiment, the results were mixed; some stations were helping each other well, and others were not.

To conclude, we have shown that we could create temporal and covariance matrices by decomposing the data. This decomposition helped us improve the performance as there are many settings with similar temporal matrices.

5. CONCLUSION

We have analyzed different covariance estimation methods using KP or KP expansion representations of the true covariance matrix. We used previously proposed results to show that the Frobenius norm minimization problem, which is not convex, can be reduced to a convex low-rank minimization problem with a KP expansion representation assumption for the true covariance matrix. We have inspected the PRLS method in detail and gave results about its convergence to the true covariance. Then we represented some simulation results that verified the theoretical findings. Also, we have used real-world wind speed data to show that these results are applicable in real life. We also want to note that having a dense Kronecker spectrum may also benefit feature selection studies. For future work, we may change the Gaussian random data assumption. This change is especially desirable because we have observed that wind speed data does not usually have a Gaussian distribution. In our observations, we have seen a distribution similar to the Weibull distribution. Also, Covariance matrices with more than two Kronecker factors may be studied. Using different types of matrices as Kronecker factors is also worth noting for the future. Some methods for modifying non-positive definite matrices to become p.s.d. may also be studied. Finally, combining temporal knowledge from different fields may be possible with our proposed TRKF method.

REFERENCES

1. Ossiander, M., M. Peszynska, L. Madsen, A. Mur and W. Harbert, "Estimation and Simulation for Geospatial Porosity and Permeability Data", *Environmental and Ecological Statistics*, Vol. 24, pp. 109–130, 2017.

2. Park, S. and R. W. Heath, "Spatial Channel Covariance Estimation for mmWave Hybrid MIMO Architecture", *50th Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, CA, United States of America, 2016.

3. Steland, A., "Shrinkage for Covariance Estimation: Asymptotics, Confidence Intervals, Bounds and Applications in Sensor Monitoring and Finance", *Statistical Papers*, Vol. 59, pp. 1441–1462, 2018.

4. Kyriakidis, P. C. and A. G. Journel, "Geostatistical Space–time Models: a Review", *Mathematical geology*, Vol. 31, No. 6, pp. 651–684, 1999.

5. Li, B., M. G. Genton and M. Sherman, "Testing the Covariance Structure of Multivariate Random Fields", *Biometrika*, Vol. 95, No. 4, pp. 813–829, 2008.

6. Janková, J. and S. van de Geer, "Confidence intervals for high-dimensional inverse covariance estimation", *Electronic Journal of Statistics*, Vol. 9, No. 1, pp. 1205 – 1229, 2015.

7. Friedman, J., T. Hastie and R. Tibshirani, "Sparse Inverse Covariance Estimation with the Graphical Lasso", *Biostatistics*, Vol. 9, pp. 432–441, 2008.

8. Stein, M. L., "Limitations on Low Rank Approximations for Covariance Matrices of Spatial Data", *Spatial Statistics*, Vol. 8, pp. 1–19, 2014.

9. Werner, K., M. Jansson and P. Stoica, "On Estimation of Covariance Matrices With Kronecker Product Structure", *IEEE Transactions on Signal Processing*,

Vol. 56, pp. 478–491, 2008.

10. Van Loan, C. F. and N. Pitsianis, "Approximation with Kronecker Products", *Linear Algebra for Large Scale and Real-time Applications*, pp. 293–314, Springer, 1993.

11. Tsiligkaridis, T. and A. O. Hero, "Covariance Estimation in High Dimensions via Kronecker Product Expansions", *IEEE Transactions on Signal Processing*, Vol. 61, pp. 5347–5360, 2013.

12. Genton, M. G., "Separable Approximations of Space-time Covariance Matrices", *Environmetrics*, Vol. 18, pp. 681–695, 2007.

13. Wang, L., J. Liu and F. Qian, "Wind Speed Frequency Distribution Modeling and Wind Energy Resource Assessment Based on Polynomial Regression Model", *International Journal of Electrical Power and Energy Systems*, Vol. 130, p. 106964, 2021.

14. Cai, T. T., C.-H. Zhang and H. H. Zhou, "Optimal Rates of Convergence for Covariance Matrix Estimation", *The Annals of Statistics*, Vol. 38, 2010.

15. Eckart, C. and G. Young, "The Approximation of One Matrix by Another of Lower Rank", *Psychometrika*, Vol. 1, No. 3, pp. 211–218, 1936.

16. Lounici, K., "High-dimensional Covariance Matrix Estimation with Missing Observations", *Bernoulli*, Vol. 20, No. 3, pp. 1029–1058, 2014.

17. Lu, N. and D. L. Zimmerman, "The Likelihood Ratio Test for a Separable Covariance Matrix", *Statistics and Probability Letters*, Vol. 73, pp. 449–457, 2005.

18. Ledoux, M. and M. Talagrand, *Probability in Banach Spaces: Isoperimetry and Processes*, Vol. 23, Springer Science & Business Media, 1991.

19. Rauhut, H., K. Schnass and P. Vandergheynst, "Compressed Sensing and Redun-

dant Dictionaries", *IEEE Transactions on Information Theory*, Vol. 54, No. 5, pp. 2210–2219, 2008.

20. Boucheron, S., G. Lugosi and P. Massart, *Concentration Inequalities*, Oxford University Press, 2013.

21. National Centers for Environmental Prediction, National Weather Service, NOAA, U.S. Department of Commerce, "NCEP GFS 0.25 Degree Global Forecast Grids Historical Archive", Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory, 2015, `https://rda.ucar.edu/datasets/ds084.1/`, accessed on February 7, 2022.