## Covariance Estimation for Matrix-valued Data

Yiichi Zhang\(^a\), Weining Shen\(^b\), and Dehan Kong\(^c\)

\(^a\)Department of Statistics, North Carolina State University, Raleigh, NC; \(^b\)Department of Statistics, University of California, Irvine, Irvine, CA; \(^c\)Department of Statistical Sciences, University of Toronto, Toronto, Canada

### ABSTRACT
Covariance estimation for matrix-valued data has received an increasing interest in applications. Unlike previous works that rely heavily on matrix normal distribution assumption and the requirement of fixed matrix size, we propose a class of distribution-free regularized covariance estimation methods for high-dimensional matrix data under a separability condition and a bandable covariance structure. Under these conditions, the original covariance matrix is decomposed into a Kronecker product of two bandable small covariance matrices representing the variability over row and column directions. We formulate a unified framework for estimating bandable covariance, and introduce an efficient algorithm based on rank one unconstrained Kronecker product approximation. The convergence rates of the proposed estimators are established, and the derived minimax lower bound shows our proposed estimator is rate-optimal under certain divergence regimes of matrix size. We further introduce a class of robust covariance estimators and provide theoretical guarantees to deal with heavy-tailed data. We demonstrate the superior finite-sample performance of our methods using simulations and real applications from a gridded temperature anomalies dataset and an S&P 500 stock data analysis. Supplementary materials for this article are available online.

### 1. Introduction
Matrix-valued data have received considerable interest in various applications. In environmental studies, the outcome of interest (e.g., temperature, humidity, air quality) is measured over a range of geographical regions. It is hence natural to represent the resulting data in a matrix form with two dimensions corresponding to latitude and longitude. Examples of matrix-valued data also include two-dimensional digital imaging data, brain surface data, and colorimetric sensor array data. There have been a few recent studies on regression analysis for matrix-valued data (Zhou and Li 2014; Wang and Zhu 2017; Kong et al. 2020; Hu et al. 2020, 2021).

In this article, we are interested in estimating the covariance of the matrix-valued data. Covariance estimation is a fundamental problem in multivariate data analysis. A large collection of statistical and machine learning methodologies including the principal component analysis, linear discriminant analysis, regression analysis and clustering analysis, require the knowledge of the covariance matrices. Denote \( \mathbf{X} \in \mathbb{R}^{p \times q} \) two dimensional matrix-valued data, and \( \text{vec}(\cdot) \) the vectorization operator that stacks the columns of a matrix into a column vector. The covariance of \( \mathbf{X} \) is defined as \( \text{cov}(\text{vec}(\mathbf{X})) = \mathbf{\Sigma}^* \in \mathbb{R}^{pq \times pq} \). A naive estimate of \( \mathbf{\Sigma}^* \) is the sample covariance. However, when \( pq > n \), it performs poorly. Wachter (1978), Johnstone (2001), and Johnstone and Lu (2009) showed that when \( pq/n \rightarrow c \in (0, \infty) \), the largest eigenvalue of the sample covariance matrix is an inconsistent estimator of the largest eigenvalue for the population covariance matrix, and the eigenvectors of the sample covariance matrix can be nearly orthogonal to the truth.

To overcome the ultra-high-dimensionality, structural assumptions are needed to estimate the covariance consistently. Various types of structured covariance matrices have been introduced such as bandable covariance matrices, sparse covariance matrices, and spiked covariance matrices. Several regularization methods have been developed accordingly to estimate these matrices, including banded methods (Bickel and Levina 2008a; Wu and Pourahmadi 2009), tapering methods (Furrer and Bengtsson 2007; Cai, Zhang, and Zhou 2010), and thresholding methods (Bickel and Levina 2008b; El Karoui 2008; Cai and Liu 2011). Another issue with the sample covariance is that it does not use the knowledge that the data actually lie in a two-dimensional matrix space. To address this issue, it is common to impose a separability assumption on the covariance of \( \text{vec}(\mathbf{X}) \), that is, \( \text{cov}(\text{vec}(\mathbf{X})) = \mathbf{\Sigma}_1^* \otimes \mathbf{\Sigma}_2^* \), where \( \mathbf{\Sigma}_1^* \in \mathbb{R}^{p \times p} \) and \( \mathbf{\Sigma}_2^* \in \mathbb{R}^{q \times q} \) represent covariances among the rows and columns of the matrices, respectively. The separability assumption helps provide a stable and parsimonious alternative to an unrestricted version of \( \text{cov}(\text{vec}(\mathbf{X})) \), and equally importantly, renders for a simple-yet-meaningful scientific interpretation. For example, when analyzing temperature measurements over a geographical region, this assumption helps decompose the variability in the measurements onto spatial directions (e.g., longitude and latitude).
To account for the separability assumption when estimating the covariance of matrix-valued data, a class of methods were proposed in the literature, all based on assuming a matrix normal distribution for the data. This idea was first proposed by Dawid (1981), and then explored by Dutilleul (1999) as they introduced an iterative algorithm for maximum likelihood estimation. Werner, Jansson, and Stoica (2008) developed two alternative estimation methods and derived the Cramér-Lower bound for the problem in a compact form. Beyond the matrix case, Galecki (1994) and Hoff (2011) considered separable covariance matrices estimation for tensor data under the tensor normal model. To summarize, all the aforementioned methods rely heavily on the matrix normal distribution assumption since their estimation procedures are obtained using maximum likelihood estimation (MLE). Moreover, these methods can only handle the matrices with fixed dimensions, especially for the development of asymptotic theory. It remains unclear how those methodologies can be generalized under realistic situations where the data do not satisfy a matrix normal distribution (or any presumed distribution) and how the asymptotic theory works for matrices with high dimensions.

In this article, we consider the covariance estimation problem for matrix-valued data under a much more challenging but realistic scenario. First, our method is distribution-free, which significantly differs from all the previous likelihood approaches. Second, we allow the dimensions of matrix-valued data to be much larger than the sample size, for example, they can diverge at the exponential rate of the sample size. Under this scenario, even if the matrix normal assumption is true, the MLE still does not exist due to overfitting. Our solution is to impose a bandable assumption on $\Sigma^*_1$ and $\Sigma^*_2$. This assumption has been widely adopted for time series with scientific applications (Visser and Molenaar 1995). The resulting bandable covariance structure exhibits a natural order among variables, thus, can naturally depict the spatial and/or temporal correlation of the matrix-valued data. We then incorporate the separable and bandable properties into one unified estimation framework, and propose an efficient computational algorithm to obtain banded and tapering covariance estimates. The convergence rates of the proposed estimators are derived and shown to be minimax optimal under the high-dimensional setting and appropriate tail conditions. A phase transition phenomenon of optimal bandwidth selection is revealed by analyzing the impact of parameter complexity on the minimax optimality regime. Our proof makes use of some new matrix analysis techniques including an $\epsilon$-net argument that assesses the impact of doubly bandable covariance structure, the newly-derived random matrix inequality (Zajkowski 2020) and the unilateral singular space perturbation bound (Cai and Zhang 2018), which shed new insights on high-dimensional regularized covariance estimation while accounting for matrix structure in the data. To deal with potentially heavy-tailed data, we further propose truncation-based robust banded and tapering covariance estimators. The truncation level is subtly analyzed to achieve an appropriate balance in bias-variance tradeoff, and the corresponding convergence rate is derived.

The rest of the article is organized as follows. We introduce our banded and tapering covariance estimates of matrix-valued data in Section 2. Section 3 provides theoretical support of our method. In Section 4, we further propose a robust banded and tapering covariance estimation procedure to deal with heavy-tailed data and provide theoretical guarantees. Simulations are conducted in Section 5 to evaluate the finite-sample performance of the proposed methods. In Section 6, we apply our method to a gridded temperature anomalies dataset. We end with some discussions in Section 7. Technical proofs, additional theoretical and numerical results, and an additional S&P 500 stock data analysis are presented in the supplementary materials.

Notation: We summarize the notation used throughout the article here. For a vector $v \in \mathbb{R}^d$, we denote its Euclidean norm by $\|v\|$. For a matrix $A = [A_{ij}] \in \mathbb{R}^{d_1 \times d_2}$, we denote $\text{tr}(A)$ its trace and $\|A\|_F$ its Frobenius norm. We also define the following matrix norms,

$$
\|A\|_2 \equiv \sup\{\|Ax\|_2, \|x\|_2 = 1\},
$$

$$
\|A\|_1 \equiv \sup\{\|Ax\|_1, \|x\|_1 = 1\} = \max_i \sum_j |A_{ij}|,
$$

$$
\|A\|_\infty \equiv \sup\{\|Ax\|_{\max}, \|x\|_{\max} = 1\} = \max_j \sum_i |A_{ij}|, \quad (1.1)
$$

$$
\|A\|_\max \equiv \max_{ij} |A_{ij}|, \quad \|A\|_{1,1} = \sum_{ij} |A_{ij}|.
$$

For two matrices $A \in \mathbb{R}^{d \times q}$ and $B \in \mathbb{R}^{p \times d}$, their Kronecker product $A \otimes B$ is a $pq \times pq$ matrix. Denote $I_d$ a $d \times d$ matrix with all elements equal to 1. Let $\circ$ be the Hadamard product of two matrices, that is, element-wise product. We use $a \wedge b$ and $a \vee b$ as shorthand notation of $\min(a, b)$ and $\max(a, b)$, respectively. For an arbitrary set $S$, we use $|S|$ to denote the cardinality of $S$. We let $\text{sgn}(\cdot)$ be the sign function and $\lfloor \cdot \rfloor$ be the floor function. We write $a \preceq b$ if there exists a universal constant $C > 0$ such that $a \leq Cb$.

2. Methodology

Denote $X \in \mathbb{R}^{p \times q}$ a two-dimensional random matrix, and $\text{vec}(\cdot)$ a vectorization operator that stacks the columns of a matrix into a vector. Let $\Sigma^* \in S^+_{pq \times pq}$ be the true covariance matrix of $\text{vec}(X_i)$, where $S^+_{d \times d}$ denotes the space of $d \times d$ positive definite matrices. Assume that $[X_i : 1 \leq i \leq n]$ are iid matrix-valued samples generated from $X$. The main interest of the article is to estimate $\Sigma^*$ from the sampled data, where we allow $p > n$ and $q > n$.

A naive estimator of $\Sigma^*$ is the sample covariance $\hat{\Sigma} = n^{-1} \sum_{i=1}^{n} \{\text{vec}(X_i) - \text{vec}(\tilde{\mu})\}\{\text{vec}(X_i) - \text{vec}(\tilde{\mu})\}^\top$, where $\tilde{\mu} = n^{-1} \sum_{i=1}^{n} X_i$. Although the sample covariance $\hat{\Sigma}$ is well-behaved for fixed $p$ and $q$, it has undesired properties when $pq > n$. In particular, the sample covariance matrix is singular, and it may not be a consistent estimator of $\Sigma^*$. In addition, the eigenvalues are often overdispersed and may be inconsistent (Bickel and Levina 2008a, 2008b).

Therefore, to estimate $\Sigma^*$ in high-dimensional settings, we impose an additional assumption that the covariance of $\text{vec}(X)$ is separable, that is,

$$
\Sigma^* = \Sigma^*_{2} \otimes \Sigma^*_{1} \in S^+_{pq \times pq},
$$

where $\Sigma^*_1$ is a positive-definite matrix and $\Sigma^*_2$ is a $q \times q$ matrix with all entries equal to 1. This assumption is motivated by the idea that the spatial and/or temporal correlation of the matrix-valued data can be naturally captured by $\Sigma^*_1$, and the remaining spatial correlation is modeled by $\Sigma^*_2$.
where $\Sigma^*_1 \in S^{p \times p}_+$ and $\Sigma^*_2 \in S^{q \times q}_+$ represent covariances among the rows and columns of the matrices, respectively. The separability assumption provides a stable and parsimonious alternative to an unrestricted version of $\Sigma^*$, and reduces the number of parameters from $pq(pq + 1)/2 \to \{p(p + 1)/2 + q(q + 1)/2\}$. This assumption is commonly used in modeling matrix-valued data (Dawid 1981; Hoff 2011) and is justified for multiple matrix-variate distributions. For example, matrix normal distributions. For example, matrix normal distribution

Another example is the matrix variate distributions. For example, matrix normal distribution

Remark 2.1. The separability of true covariance is a key assumption in our framework, and is recommended to be tested in the data preprocessing stage. As a common assumption in spatial statistics, neuroimaging and functional data analysis, many methods have been proposed to test this assumption. For example, Lu and Zimmerman (2005) developed likelihood ratio tests and Filipiak, Klein, and Roy (2016) considered score test for separability of the covariance matrices under the matrix Gaussian assumption; Aston, Pigoli, and Tavakoli (2017) provided projected-based bootstrap tests under both (parametric) matrix Gaussian and nonparametric conditions. In our real data applications, we implement Aston, Pigoli, and Tavakoli (2017)’s projected-based bootstrap test because it is both theoretically guaranteed and computationally fast under high-dimensional scenario. In addition, the test is distribution-free, which is suitable for our nonparametric framework.

2.1. Banded Covariance

To estimate the covariance matrix when $p > n$ or $q > n$, regularizing large empirical covariance matrices has been widely used in literature (Bickel and Levina 2008). One popular way is to band the sample covariance matrix. For any $A = [A_{l,m}]_{d \times d}$ and $k > 0$, define

$$B_d(k) = \left\{ A \in \mathbb{R}^{d \times d} : A_{l,m} = 0 \text{ for any } |l - m| > k, 1 \leq l, m \leq d \right\},$$

and

$$B_k(A) = \left\{ A_{l,m} : I(|l - m| \leq k) \right\}_{d \times d},$$

where $I(\cdot)$ is an indicator function. We propose to solve the following optimization problem for a given pair of tuning parameters $(k_1, k_2)$:

$$\left(\hat{\Sigma}^B_1(k_1), \hat{\Sigma}^B_2(k_2)\right) = \underset{\Sigma_1 \in B_p(k_1), \Sigma_2 \in B_q(k_2)}{\text{argmin}} \| \hat{\Sigma} - \Sigma_2 \otimes \Sigma_1 \|_{F}^2.$$  

(2.2)

And the banded covariance estimate corresponding to $(k_1, k_2)$ is $\hat{\Sigma}^B_1(k_1, k_2) = \hat{\Sigma}^B_2(k_2) \otimes \hat{\Sigma}^B_1(k_1)$. Here $k_1$ and $k_2$ control the regularization level of banding.

Surprisingly, the above problem has a closed form solution. Define $\hat{\Sigma}^B_1(k_1, k_2)$ to be a $pq \times pq$ matrix satisfying $\hat{\Sigma}^B_1(k_1, k_2) = \hat{\Sigma} \circ \{B_k(1_2) \otimes B_k(1_2)\}$, where $1_p$ and $1_q$ are matrices of all 1’s with dimensions $p \times p$ and $q \times q$, respectively. We call $\hat{\Sigma}^B_1(k_1, k_2)$ a doubly banded matrix of $\hat{\Sigma}$ with bandwidths $k_1$ and $k_2$. We have the following proposition whose proof is deferred to the supplementary materials.

**Proposition 2.2.** Solving (2.2) is equivalent to solving the following optimization problem:

$$\left(\hat{\Sigma}^B_1(k_1), \hat{\Sigma}^B_2(k_2)\right) = \underset{\Sigma_1 \in B_p(k_1), \Sigma_2 \in B_q(k_2)}{\text{argmin}} \| \hat{\Sigma} - \Sigma_2 \otimes \Sigma_1 \|_{F}^2.$$  

(2.3)

This proposition provides an efficient way to solve the optimization problem (2.2). In particular, one can first obtain $\hat{\Sigma}^B(k_1, k_2)$ by doubly banding $\hat{\Sigma}$, and then solve the rank one unconstrained Kronecker product approximation (2.3) based on the method in Van Loan and Pitsianis (1993) and Pitsianis (1997). Solutions $\hat{\Sigma}^B_1(k_1), \hat{\Sigma}^B_2(k_2)$ are identified up to scale, while $\hat{\Sigma} = \hat{\Sigma}^B_2(k_2) \otimes \hat{\Sigma}^B_1(k_1)$ is unique. The procedure of solving (2.3) implies that our proposed method is a *spectral method*, which has been applied to a wide class of statistical problems. We refer interested readers to Chen et al. (2021) for a recent survey therein. More discussions on solving (2.3) are deferred to Section 5.4 of the supplementary materials.

To implement the rank one unconstrained Kronecker product approximation, we adopt the algorithm proposed in Batselier and Wong (2017). This algorithm is implemented using the Matlab package TKPSVD, which can be downloaded at https://github.com/kbatseli/TKPSVD.

There are two tuning parameters $k_1$ and $k_2$ involved in our estimation procedure. Theoretically, we will show in Section 3 that when $k_1$ and $k_2$ are chosen appropriately, our estimator will be consistent even when both $p$ and $q$ diverge at the exponential order of the sample size $n$. In practice, we apply the resampling procedure proposed in Bickel and Levina (2008) to select the optimal bandwidths $k_1$ and $k_2$. In particular, we randomly split the original data into a training set and a test set, with sample sizes $n_1$ and $n_2 = n - n_1$, respectively. We use the training set to estimate the covariance matrix $\hat{\Sigma}^B_1(k_1, k_2)$ using our procedure, and compare with the sample covariance matrix of the test sample $\hat{\Sigma}^t$. We repeat the random split procedure for $N$ times, and let $\hat{\Sigma}^t_1(k_1, k_2)$ and $\hat{\Sigma}^t_2$ denote the estimates from the $v$th split for $v = 1, \ldots, N$. We select the $(k_1, k_2)$ that minimizes

$$R(k_1, k_2) = N^{-1} \sum_{v=1}^{N} \| \hat{\Sigma}^t_1(k_1, k_2) - \hat{\Sigma}^t_1 \|_{F}.$$  

Similar to Bickel and Levina (2008), we choose $n_1 = \lfloor \frac{n}{4} \rfloor$ and $n_2 = n - n_1$. We use $N = 10$ random splits throughout the article. Denote $\hat{k}_1$ and $\hat{k}_2$ the selected optimal bandwidths. Our final banded covariance estimate is $\hat{\Sigma}^B_1(\hat{k}_1, \hat{k}_2) = \hat{\Sigma}^B_2(\hat{k}_2) \otimes \hat{\Sigma}^B_1(\hat{k}_1)$, where

$$\left(\hat{\Sigma}^B_1(\hat{k}_1), \hat{\Sigma}^B_2(\hat{k}_2)\right) = \underset{\Sigma_1 \in B_p(\hat{k}_1), \Sigma_2 \in B_q(\hat{k}_2)}{\text{argmin}} \| \hat{\Sigma} - \Sigma_2 \otimes \Sigma_1 \|_{F}^2.$$  

(2.4)
2.2. Tapering Covariance

Another popular technique for covariance matrix regularization is tapering (Bickel and Levina 2008b; Cai, Zhang, and Zhou 2010). For any matrix \( A = [A_{lm}]_{d \times d} \) and any \( k \geq 0 \), we define \( T_k(A) = [T_k(A)_{lm}]_{d \times d} \), where

\[
T_k(A)_{lm} = \begin{cases} A_{lm} & \text{when } |l - m| \leq |k/2|, \\ 0 & \text{otherwise}. \\end{cases}
\]

Consider

\[
\widehat{\Sigma}_T(k_1, k_2) = \widehat{\Sigma} \circ \{T_{k_2}(1_q) \otimes T_{k_1}(1_p)\}.
\]

Analogous to (2.3), we propose to solve

\[
(\widehat{\Sigma}^T_1(k_1), \widehat{\Sigma}^T_2(k_2)) = \text{argmin} \left\| \widehat{\Sigma}_T(k_1, k_2) - \Sigma_2 \otimes \Sigma_1 \right\|_F^2.
\]

(2.6)

Then we obtain the tapering covariance estimate as \( \widehat{\Sigma}_T(k_1, k_2) = \widehat{\Sigma}_2 \otimes \widehat{\Sigma}_1(k_1) \).

Remark 2.3. For the doubly banded covariance, we have \( \widehat{\Sigma}_T(k_1, k_2) = \widehat{\Sigma} \) when \( k_1 \geq p \) and \( k_2 \geq q \). However, for the doubly tapering covariance, we need \( k_1 \geq 2p, k_2 \geq 2q \) to make \( \widehat{\Sigma}_T(k_1, k_2) = \widehat{\Sigma} \). Therefore, in the theoretical analysis, we always assume \( k_1 \) and \( k_2 \) are at most the same order as \( p \) and \( q \), respectively.

Remark 2.4. When \( p \) or \( q \) equals 1, the matrix-valued data degenerate to the vector-valued data, which we refer to as degenerate regime. By the forms of our proposed optimization problems in (2.2) and (2.6), it is easy to see that our proposed estimators also simplify to Bickel and Levina (2008a)'s banded estimator. Similarly, with those estimators under the degenerate regime, our proposed estimators are not guaranteed to be positive semidefinite or positive-definite. It is possible to apply the eigenvalue truncation technique (see, e.g., Remark 3 in Cai, Zhang, and Zhou 2010) to resolve this issue. In particular, we can replace the negative eigenvalues of our proposed estimator with 0 or a small positive constant.

3. Theoretical Results

3.1. Notation

Following Bickel and Levina (2008b), we define the following uniformity class of approximately banded covariance matrices

\[
\mathcal{F}(\varepsilon_0, \alpha) = \left\{ \Sigma : \max_m \sum_l |\sigma_{lm}| : |l - m| > k \leq C_0 k^{-\alpha} \text{ for all } k \geq 1, \right. \\
\left. \text{and } 0 < \varepsilon_0 \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq 1/\varepsilon_0 \right\},
\]

with fixed constants \( C_0, \alpha \geq 0 \) and some \( \varepsilon_0 < 1 \), and \( \lambda_{\min}(\Sigma), \lambda_{\max}(\Sigma) \) are the smallest and largest eigenvalues of \( \Sigma \), respectively. We further define another important class of covariance matrices (Cai, Zhang, and Zhou 2010),

\[
M(\varepsilon_0, \alpha) = \left\{ \Sigma : |\sigma_{lm}| \leq C_1 |l - m|^{-\alpha - 1} \text{ for } l \neq m, \right. \\
\left. \text{and } 0 < \varepsilon_0 \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq 1/\varepsilon_0 \right\},
\]

with fixed constants \( C_1, \alpha > 0 \) and some \( \varepsilon_0 < 1 \). One can see that \( M(\varepsilon_0, \alpha) \) is a subset of \( \mathcal{F}(\varepsilon_0, \alpha) \) when \( C_1 \leq \alpha C_0 \). In other words, \( M(\varepsilon_0, \alpha) \) is a more restrictive class. Since we focus on the bandable covariance matrix classes \( \mathcal{F}(\varepsilon_0, \alpha) \) and \( M(\varepsilon_0, \alpha) \), the true covariance matrix in the subsequent theoretical analysis \( \mathbf{\Sigma}_* \) will always have eigenvalues bounded away from \( +\infty \), which is consistent with the previous works dealing with bandable covariance estimation for vector-valued data (Bickel and Levina 2008a; Cai, Zhang, and Zhou 2010; Cai and Yuan 2012; Cai, Ren, and Zhou 2016). A more detailed discussion on bounded eigenvalues of \( \mathbf{\Sigma}_* \) is given in Section S.2.1 of the supplementary materials.

3.2. Main Results

3.2.1. Convergence Rate of Proposed Estimators

In this section, we derive the upper bound for the convergence rates of our banded and tapering covariance matrix estimates. We derive the \( \ell^2 \) convergence rates of our proposed estimators under Frobenius norm. By Markov’s inequality, the convergence rates of our proposed estimators also simplify to Bickel and Levina (2008a)’s banded estimator and Cai, Zhang, and Zhou (2010)’s tapering estimator under this degenerate regime. Similarly, with those estimators under the degenerate regime, our proposed estimators are not guaranteed to be positive semidefinite or positive-definite. It is possible to apply the eigenvalue truncation technique (see, e.g., Remark 3 in Cai, Zhang, and Zhou 2010) to resolve this issue. In particular, we can replace the negative eigenvalues of our proposed estimator with 0 or a small positive constant.

\[
\text{Pr} \left[ \left| \mathbf{v}^\top \left( \text{vec}(\mathbf{X}) - \mathbb{E} \{ \text{vec}(\mathbf{X}) \} \right) \right| > t \right] \leq e^{-x^2}. \tag{3.3}
\]

Recall \( \mathbf{X}_1, \ldots, \mathbf{X}_n \) are iid \( p \times q \) random matrix samples and vec(\( \mathbf{X}_i \)) \( \in \mathbb{R}^{pq} \) their vectorizations. Denote \( x_{l_1 l_2}^{(i)} \) the \( l_1 l_2 \)th entry of \( \mathbf{X}_i \) for \( 1 \leq l_1 \leq p \) and \( 1 \leq l_2 \leq q \). Denote \( \preceq \) and \( \succeq \) inequalities up to multiplicative universal constants and \( \approx \) an equality up to a multiplicative universal constant. For sequences \( a_n, b_n \), we also write \( a_n = \Theta(b_n) \) if \( a_n \asymp b_n \).
derive the convergence rates of the covariance estimates under these two cases when \( X_i \) is a random vector, respectively.

Denote \( \xi_2^B(k_2) \otimes \xi_1^B(k_1) \) the proposed banded estimator and \( \tilde{\xi}_2^T(k_2) \otimes \tilde{\xi}_1^T(k_1) \) the proposed tapering estimator. To unify the notation, we denote \( \eta \in \{B, T\} \). The following theorem presents an upper bound on the convergence rate for the proposed banded and tapering covariance estimators obtained from (2.2) and (2.6) under Scenario (i).

**Theorem 3.1.** Let \( \text{vec}(X_1), \text{vec}(X_2), \ldots, \text{vec}(X_n) \) be iid sub-Gaussian random vectors in \( \mathbb{R}^{p^2} \) with true covariance \( \Sigma^* = \Sigma_2^* \otimes \Sigma_1^* \). Let \( I_{\eta,q}(k) = I(\eta = B, k < d - 1) + I(\eta = T, k < 2d - 2) \) for \( \eta \in \{B, T\}, d \geq 1 \), and let \( \bar{a}_a = \{2a_d \) when \( \Sigma^*_a \in \mathcal{F}(\varepsilon_0, a_1), \Sigma^*_a \in \mathcal{F}(\varepsilon_0, a_2) \) for \( a \in [1, 2] \).

Then we have,

\[
\mathbb{E} \left( \frac{1}{pq} \left( \frac{2k_1 + k_2}{pm} + I_{\eta,q}(k_1) \cdot k_1^{-2a} \right) \right) \leq \left\{ \begin{array}{ll}
pk_1 + qk_2 \leq n; \\
\frac{1}{pq} \left( \frac{k_1^2}{n} + \frac{k_2^2}{pm} + I_{\eta,q}(k_1) \cdot k_1^{-2a} \right) + I_{\eta,q}(k_2) \cdot k_2^{-2a}; & pk_1 + qk_2 > n.
\end{array} \right.
\]

(3.5)

To better understand the error rate in (3.5), consider a simple example where \( q \asymp 1 \) and \( \Sigma^*_1, \Sigma^*_2 \in \mathcal{M}(\varepsilon_0, \alpha) \). Then the rate becomes \( \min \{ n^{-\alpha/2}, p/n \} \) with \( k_1 = \min \{ n^{-\alpha/2}, 2p \} \) and \( k_2 = 2q \), which matches exactly with the minimax optimal rate in Cai, Zhang, and Zhou (2010). Another example is when \( p = q \asymp \sqrt{n} \) and \( \alpha_1 = \alpha_2 \). Then (3.5) becomes \( (k_1/pn) + I_{\eta,q}(k_1) \cdot k_1^{-2a} \asymp \min \{ (pn)^{-\alpha_1}, \max(1/n) \} \) under the optimal choice for \( k_1 = k_2 \), which is \( \min \{ (pn)^{1/2}, \max(1/n) \} \). In general, the selection of \( k_1, k_2 \) to attain the optimal convergence rate can be quite complicated depending on the divergence regimes of \( p, q \); and we provide a detailed discussion in Section S.2.3.

**Scenario (ii):** The sub-Gaussian assumption (i) can be relaxed to the finite fourth moment condition (ii), with a sacrifice of the convergence rate. The following theorem presents an upper bound on the convergence rate for the proposed banded and tapering covariance estimators obtained from (2.2) and (2.6) under Scenario (ii).

**Theorem 3.2.** Let \( \text{vec}(X_1), \text{vec}(X_2), \ldots, \text{vec}(X_n) \) be iid random vectors in \( \mathbb{R}^{p^2} \) with true covariance \( \Sigma^* = \Sigma_2^* \otimes \Sigma_1^* \). Assume \( \mathbb{E}(\sum_{i,j=1}^n \sum_{m=1}^d x_{i,j}^m x_{i,m}^j)^2 \leq M < +\infty \), where \( M \) is a constant that does not depend on \( n, l_1, l_2, m_1, m_2 \). Let \( I_{\eta,q}(k) \) and \( \bar{a}_a \) be the same as in Theorem 3.1. Then we have,

\[
\mathbb{E} \left( \frac{1}{pq} \left( \frac{2k_1 + k_2}{pm} + I_{\eta,q}(k_1) \cdot k_1^{-2a} \right) \right) \leq \left\{ \begin{array}{ll}
\frac{k_1k_2}{n} + I_{\eta,q}(k_1) \cdot k_1^{-2a} \\
+ I_{\eta,q}(k_2) \cdot k_2^{-2a}.
\end{array} \right.
\]

(3.6)

The selection of \( k_1, k_2 \) to attain the optimal convergence rate of (3.6), under different divergence regimes of \( p, q \), is discussed in Section S.2.5, 9 of the supplementary materials.

### 3.2.2. Overall Lower Bound

In this section, we give an overall lower bound for the convergence rates of covariance matrix estimates for a special scenario, where \( \Sigma^*_1 \) and \( \Sigma^*_2 \) are in \( \mathcal{M}(\varepsilon_0, \alpha) \) and \( \text{vec}(X) \) follows iid sub-Gaussian distribution. This scenario is a special case of Scenarios (i) and (ii) with either \( \Sigma^*_1 \in \mathcal{F}(\varepsilon_0, a_1), \Sigma^*_2 \in \mathcal{F}(\varepsilon_0, a_2) \) or \( \Sigma^*_1 \in \mathcal{M}(\varepsilon_0, a_1), \Sigma^*_2 \in \mathcal{M}(\varepsilon_0, a_2) \), as considered in Section 3.2.1. So the lower bound we present here can be compared with the upper bounds presented in both Theorems 3.1 and 3.2.

**Theorem 3.3.** Let \( \mathcal{P}^n_{a_1, a_2} \) denote the class of distributions of \( \{\text{vec}(X_i)\}_{i=1}^n \), such that \( \text{vec}(X_1), \text{vec}(X_2), \ldots, \text{vec}(X_n) \) are iid sub-Gaussian random vectors in \( \mathbb{R}^{p^2} \) with any true covariance \( \Sigma^* = \Sigma_2^* \otimes \Sigma_1^* \), where \( \Sigma^*_1 \in \mathcal{M}(\varepsilon_0, a_1), \Sigma^*_2 \in \mathcal{M}(\varepsilon_0, a_2) \). Let \( \hat{\Sigma}_n \) be any possible covariance estimator based on \( \{\text{vec}(X_i)\}_{i=1}^n \), we have

\[
\inf_{\Sigma^*_a \in \mathcal{P}^n_{a_1, a_2}} \sup_{p \in \mathcal{P}^n_{a_1, a_2}} \mathbb{E} \left( \frac{1}{pq} \left( \frac{2k_1 + k_2}{pm} + I_{\eta,q}(k_1) \cdot k_1^{-2a} \right) \right) \geq \max \left\{ \min \left\{ \frac{p}{mq}, (nq)^{-\frac{1}{2m}} \right\}, \min \left\{ \frac{q}{np}, (np)^{-\frac{1}{2m}} \right\} \right\}.
\]

(3.7)

### 3.3. Additional Theoretical Results

We summarize other major theoretical findings as nine take-away messages below.

1. By matching the derived lower and upper bounds, we are able to obtain sufficient conditions for which our obtained convergence rate is minimax optimal (similarly to Cai, Zhang, and Zhou (2010), we focus on the case that \( \Sigma_1, \Sigma_2 \) are in the \( \mathcal{M}(\varepsilon_0, \alpha) \) class). In particular, consider two regimes of \( p, q; (1) \) degenerate regime, where \( p \wedge q = O(1) \); (2) moderate high-dimensional regime, where \( 1 \leq p \wedge q \leq \max \left\{ \frac{n}{q}, \frac{q}{n} \right\} \). We can show that under Scenario (i), our proposed estimator is rate-optimal under both regimes; and under Scenario (ii), our estimator is rate-optimal under the degenerate regime. These rate optimality findings are further verified by a simulation study in Section S.2.9, supplementary materials.

2. Under Scenario (i), the proposed estimator is always rate-optimal when \( p, q \asymp \sqrt{n} \). In addition, when the bandable levels \( \alpha_1, \alpha_2 \) are large enough, the proposed estimator is also rate-optimal under the more challenging situation when both \( p, q \) diverge at an asymptotic order close to \( n \). For example, when \( \alpha_1 = \alpha_2 = 2 \), the proposed estimator is rate-optimal when \( p = q = n^{0.7} \). See Section S.2.7.3 of the supplementary materials for more details.

3. Our results reveal an interesting phase transition phenomenon in the sense that the optimal rate can be achieved without banding when \( p \) and \( q \) do not diverge fast enough in \( n \). Take optimal \( k_1 \) as an example. If \( p \) diverges not sufficiently fast compared to \( q \), then no banding is needed for \( k_1 \). For example, consider \( \alpha_1 = \alpha_2 = 2 \). Under Scenario (i) with the
proposed banded estimator, we let \( p, q \) satisfy the moderate high-dimensional regime. If \( p < (nq)^{1/6} \), then the optimal \( k_1 = p-1 \), which means no banding on \( p \) direction is needed. If \( p > (nq)^{1/6} \), then the optimal \( k_1 \approx (nq)^{1/6} \ll p \). Similar symmetric results can be obtained for the optimal \( k_2 \).

To provide more insights, we now focus on the example under Scenario (i) with \( \alpha_1 = \alpha_2 = 2 \), \( M \)-class \( \Sigma_1, \Sigma_2 \), and give some graph illustration. We let \( p = n^{\beta_1} \) and \( q = n^{\beta_2} \), where different \( \beta_1, \beta_2 > 0 \) represent different divergence regimes of \( p, q \). Results in Section S.2.3 show that the upper bounds of Theorems 3.1 after selecting the optimal \( k_1, k_2 \), as well as the lower bound of Theorem 3.3, always have the polynomial form: \( n^{-r} \). As \( r = -\log_n (n^{-r}) \), we call \( r \) the negative log convergence rate (NLCR) and use it to measure the convergence rate of both upper and lower bounds. Similarly, the optimal divergence regimes of \( k_1, k_2 \) always have the form: \( n^r \). We call \( r' \) the log divergence rate (LDR) and use it to measure the divergence rate of optimal \( k_1 \).

In Figure 1(a) and (b), we illustrate the NLCRs of the lower and upper bounds of the proposed estimator with different divergence regimes of \( p, q \) under Scenario (i). The lower and upper bounds are matched in the rate-optimal region in Figure 1(b), in which the regime \( p = q = n^{0.7} \) is included. When \( p, q \) become larger, for example, \( p = q = n^{0.8} \), the proposed estimator is no longer rate-optimal by the current theoretical results.

In Figure 1(c), the LDR of optimal \( k_1 \) is in the no-banding region when \( p \) is slowly divergent. When \( p \) diverges faster than the phase-transition rate \( (nq)^{1/6} \), the LDR of optimal \( k_1 \) is out of the no-banding region, and thus the banding on the \( p \) direction is needed. The no-banding region in Figure 1(c) becomes wider when \( \beta_2 \) becomes larger. This is because the phase-transition rate \( (nq)^{1/6} \) diverges faster when \( q \) diverges faster.

We leave more graphical illustrations in Section S.2.8 of the supplementary materials.

5. In general, parameter complexity (such as matrix size and bandable level) has a complex effect on the error rate, the optimal bandwidth selection, and the rate-optimal regime. For example, under Scenario (i), with larger bandable levels \( \alpha_1, \alpha_2 \), a higher dimensional regime of \( p, q \) can be shown to be rate-optimal by Theorem 3.1. A detailed discussion is given in Section S.2.7, supplementary materials.

6. There are two technical reasons that make our estimator sub-optimal when \( p \) and \( q \) are both very large (e.g., \( \gg \sqrt{n} \)). One is that the spectral-norm bound in Lemma S.8.8 is loosened from \( \frac{k_1}{q} + \frac{k_2}{p} \) to \( \frac{pk_1^2}{q^2} + \frac{qk_2^2}{p^2} \), when \( pk_1 + qk_2 > n \), that is, if the spectral-norm bound in Lemma S.8.8 can be kept as the tight bound \( \frac{k_1}{q} + \frac{k_2}{p} \), then the resulting optimal upper bounds by Theorem 3.2 will always match the lower bound in Theorem 3.3, regardless of \( p, q \)’s divergent regime. Another reason is that the lower bound in Theorem 3.3 may not be tight. Improving these bounds is an interesting future work direction.

7. For the degenerate regime (one of \( p, q \) is \( O(1) \)), the rate optimality of both upper and lower bounds, essentially agrees with those of Bickel’s linear banded estimator (Bickel and Levina 2008b; Xiao and Bunea 2014) and Cai’s linear tapering estimator in Cai, Zhang, and Zhou (2010). See Section S.2.6.1 of the supplementary materials for more details.

8. We have derived the error bounds for individual matrix component estimation in Section S.2.10 of the supplementary materials. The results show that our individual matrix estimator, multiplied by a proper constant, which is needed for the identifiability purpose, converges to the truth at the obtained rates under the Frobenius norm.

9. We extend the rate results under the Frobenius norm to the spectral norm in Section S.2.11 of the supplementary materials. It turns out to be very challenging to obtain spectral norm convergence results for our current estimator. Therefore, we consider a rank one Kronecker product approximation based on minimizing the spectral norm rather than the Frobenius norm in our original estimation procedure. We then derive an upper bound for the convergence rate of the new estimator under the spectral norm.
4. Robust Covariance Estimation

Heavy-tailed data are commonly encountered in many applications. When modeling heavy-tailed data, Theorems 3.1 and 3.2 may not be suitable since they require sub-Gaussian tail and finite fourth moment conditions. In this section, we propose robust banded/tapering covariance estimators that improve the proposed nonrobust estimators and enjoy desired theoretical properties for heavy-tailed data.

We first give a precise definition of heavy tail condition for the sample matrix $X_i, 1 \leq i \leq n$. Following state-of-the-art robust covariance estimation literature (Avella-Medina et al. 2018; Lu, Han, and Liu 2021), we quantify heavy-tailedness of $X_i$ by elementwise $2\gamma$-finite moment order, such that for any $1 \leq l_1, m_1 \leq p$ and $1 \leq l_2, m_2 \leq q$,

$$
E\left[\left|x_{l_1m_1}^{(i)} \cdot x_{l_2m_2}^{(i)}\right|^\gamma\right] \leq M < +\infty.
$$

We will show that the robust banded/tapering covariance estimators enjoy the following two nice properties. First, our proposed robust estimators converge to the truth under the Frobenius norm, even when data only have a finite moment order of $\gamma > 1$. Compared to the proposed nonrobust estimators, Theorem 3.2 needs at least $\gamma \geq 2$, that is, a finite fourth or higher moment. Additionally, even when $\gamma \geq 2$, the convergence rate in Theorem 3.2 is always slower or equal to the convergence rate in Theorem 4.1. Second, we show our proposed robust estimators have an adaptive convergence rate with different levels of heavy-tailedness. In particular, when the data $X_i$ has a higher moment condition $\gamma$, the convergence rate becomes faster, and closer to the rate in Theorem 3.1 under the sub-Gaussian condition, while the proposed nonrobust estimators do not have such properties.

We adopt the idea in Fan, Wang, and Zhu (2021) and truncate the sample covariance as a preliminary step. When $\text{vec}(X_i) \in \mathbb{R}^{p^2}$, the modified estimator is defined as

$$
\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} \left[\text{vec}(\tilde{X}_i) - \text{vec}(\hat{\mu})\right] \cdot \left[\text{vec}(\tilde{X}_i) - \text{vec}(\hat{\mu})\right]^T,
$$

where $\tilde{X}_i$ satisfies

$$
\tilde{x}_{h_{l_2}}^{(i)} = \text{sgn}(x_{h_{l_2}}^{(i)})(x_{h_{l_2}}^{(i)} | \wedge \tau) \quad (4.1)
$$

with some selected $\tau > 0$ for all $1 \leq l_1 \leq p, 1 \leq l_2 \leq q$, and $\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} \tilde{X}_i$. Same as Fan, Wang, and Zhu (2021), we assume $\mathbb{E}[X_i] = 0$ for the ease of exposition. For general covariance estimation problem, the results in Fan, Wang, and Zhu (2021) suggest that if elementwisely $\mathbb{E}\left[\left|x_{h_{l_2}}^{(i)} \cdot x_{h_{l_2}}^{(i)}\right|^\gamma\right] \leq M < +\infty$, where $M$ is a constant that does not depend on $n, l_1, m_1, l_2, m_2$, then $|\hat{\Sigma} - \Sigma^*|_{\text{max}} = O_P\left(\sqrt{\log \max(p, q)}/n\right)$ by taking $\tau \asymp \left[\log\max(p, q)\right]/n$, while sample covariance estimator needs the sub-Gaussian assumption to achieve max-norm convergence.

Similarly, instead of using the sample covariance $\tilde{\Sigma}$, we replace $\tilde{\Sigma}$ with $\hat{\Sigma}$ when implementing our banded and/or tapering estimation procedure. In particular, define

$$
\hat{\Sigma}_B(k_1, k_2) = \hat{\Sigma} \circ \left( B_{k_1}(1_p) \otimes B_{k_2}(1_p) \right);
$$

$$
\hat{\Sigma}_T(k_1, k_2) = \hat{\Sigma} \circ \left( T_{k_2}(1_q) \otimes T_{k_1}(1_p) \right),
$$

as the banded/tapering matrix of $\hat{\Sigma}$ with parameters $k_1$ and $k_2$. We propose the robust banded/tapering estimators as $\hat{\Sigma}_B(k_1, k_2)$ and $\hat{\Sigma}_T(k_1, k_2)$. Theorem 4.1 and 4.2 are solutions from

$$
\left(\hat{\Sigma}_1 \otimes \hat{\Sigma}_2\right)(k_1, k_2) = \arg\min \left\|\hat{\Sigma}_1(k_1, k_2) - \Sigma_1 \otimes \Sigma_2\right\|^2_{\text{F}}.
$$

In practice, we adopt the resampling scheme in Section 2.1 to select $\tau$ from a candidate pool $\mathcal{T} = \{|x_i|; i \in \mathcal{P}\}$, where $|x_i|$ is the $i$ percentile value among the set of all possible absolute values of coordinates in any $X_i$, that is, $\{|x_i|; 1 \leq i \leq n, 1 \leq l_1 \leq p, 1 \leq l_2 \leq q\}$ and $\mathcal{P}$ is a candidate pool of percentiles under consideration. To solve (4.2) and select the tuning parameters $(k_1, k_2)$, we adopt the same procedures as the ones used in nonrobust banded/tapering covariance estimation in Section 2.1.

Before presenting our asymptotic result, we list two regularity assumptions for the truncated data $\tilde{X}_1, \ldots, \tilde{X}_n$. Denote $\hat{\Sigma}_n^* \equiv \text{cov}(\tilde{X}_i)$ and its doubly banded or tapering version $\Sigma_n^* = (k_1, k_2)$ with $n \in \{B, T\}$.

Assumption 4.1. We assume eigenvalues of $\Sigma_n^*$ satisfying $\epsilon^* < \lambda_{\min}(\Sigma_n^*) \leq \lambda_{\max}(\Sigma^*_n) < 1/\epsilon^*$ for some constant $0 < \epsilon^* < 1$.

Assumption 4.2. There exists $J_n$ such that $\|\text{vec}(\tilde{X}_i)\|_{\chi^2_{p}} = \sup_{|\|v\|\leq 1} \|v^\top \text{vec}(\tilde{X}_i)\|_{\chi^2_{p}} \leq J_n\tau$, where $1 \gtrsim J_n \gtrsim \sqrt{pq}$ and $J_n$ only depends on $n$.

Here $\|\cdot\|_{\chi^2_{p}}$ is the sub-Gaussian norm defined in Section S.1.3 of the supplementary materials. Equation (4.1) implies that each element of $\text{vec}(\tilde{X}_i)$ can be upper bounded by $C_{\gamma}\tau$ under sub-Gaussian norm, with some constant $C_{\gamma} > 0$. Assumption 4.2 further introduces a general thresholding factor $J_n$ such that upper bound $J_n\tau$ holds uniformly for any $\|v^\top \text{vec}(\tilde{X}_i)\|_{\chi^2_{p}}$ when $\|v\| = 1$, where $J_n \gtrsim 1$. A trivial bound of $J_n$ is $J_n \gtrsim \sqrt{pq}$ by triangle inequality. In many conditions, based on the probabilistic structure of $X_i$, $J_n$ can be further reduced to a constant. We defer more discussions to Section S.3.1 of the supplementary materials.

Our main result is given in the following theorem, and the proof is included in the supplementary materials.

**Theorem 4.1.** Let $\text{vec}(X_1), \text{vec}(X_2), \ldots, \text{vec}(X_n)$ be iid random vectors in $\mathbb{R}^{p^2}$ with true covariance $\Sigma^* = \Sigma_2 \otimes \Sigma_1^*$, Assume $\mathbb{E}(x_{h_{l_2}}^{(i)} \cdot x_{h_{l_2}}^{(i)}) \leq M < +\infty$ where $\gamma > 1$ is the order of heavy-tailedness, and $M$ is a constant that does not depend on $n, l_1, m_1, l_2, m_2$. Define $I_{\eta, \alpha}(k)$ and $\alpha_{\alpha}(k)$ the same as in Theorem 3.1, and define the error terms, $r_1^{R_{\gamma}}(k_1, k_2, p, q, n) = (k_1k_2)^{1/\gamma} \left( k_1^{p/\gamma} + k_2^{q/\gamma}\right)^{1-1/\gamma}$, and $r_2^{R_{\gamma}}(k_1, k_2, p, q, n) = (k_1k_2)^{1/\gamma} \left( k_1^{p/\gamma} + k_2^{q/\gamma}\right)^{1-1/\gamma}$.

Consider the following four cases: C1. $pk_1 + qk_2 \lesssim n$ and $1 < \xi < 2$; C2. $pk_1 + qk_2 \lesssim n$ and $\xi \geq 2$; C3. $pk_1 + qk_2 > n$ and $1 < \xi < 2$; C4. $pk_1 + qk_2 > n$ and $\xi \geq 2$. Under Assumptions 4.1–4.2, by choosing the optimal $\tau$ given in Section S.3.3 in the supplementary materials, we have
We report several quantities, \( \| \hat{\Sigma} - \Sigma \|_1, \| \hat{\Sigma} - \Sigma \|_2 \), where \( \hat{\Sigma} \) can be our proposed estimators and the other five comparison estimators. These quantities characterize the estimation errors for the covariance matrices. We also report the \( \hat{k}_1 \) and \( \hat{k}_2 \) for our proposed methods and the doubly banded/tapering estimators; and \( \hat{k} \) for Bickel and Levina (2008b)'s banded estimator and Cai, Zhang, and Zhou (2010)'s tapering estimator. We summarize the averages of these quantities over 100 Monte Carlo repetitions in Table 1 and Tables S1–S7 in Section S.6.1 of the supplementary materials. Their associated standard errors are summarized in Section S.6.2 of the supplementary materials.

From these tables, we can see that our proposed methods always perform better than all comparison methods in terms of estimation errors. In addition, the doubly banded and tapering estimators perform generally better than the other three comparison estimators (sample/banded/tapering). For Case 1, the oracle \( k_1 \) and \( k_2 \) for the banded estimator are both 1. Noticing that \( B_1(1_p) = T_2(1_p) \), the oracle \( k_1 \) and \( k_2 \) for the tapering estimator are both 2, so our method can select \( \hat{k}_1 \) and \( \hat{k}_2 \) accurately. For Case 2, when \( \rho_1 \) and \( \rho_2 \) increase, the selected bandwidths \( \hat{k}_1 \) and \( \hat{k}_2 \) for the proposed method also increase.

### 6. Gridded Temperature Anomaly Data Analysis

We analyze a gridded temperature anomalies dataset collected by the U.S. National Oceanic and Atmospheric Administration (NOAA) (Shen 2017; Gu and Shen 2020) in this section. Another case study of a stock price dataset is presented in Section S.5 of the supplementary materials. The temperature anomalies dataset contains the monthly air and marine temperature measurements from January 1880 to 2017 with a \( 5^\circ \times 5^\circ \) resolution.
latitude-longitude resolution. It can be downloaded at ftp://ftp.ncdc.noaa.gov/pub/data/noaaglobaltemp/operational.

In our study, we focus on the temperature anomalies (the difference between an observed temperature and the baseline/normal value) in the past 20 years over the region marked in deep blue as shown in Figure S14 (supplementary materials) to avoid the missing values and to make sure the resulting data are in a matrix form (with two dimensions representing longitude and latitude). We have preprocessed the data to remove the mean trend and the dependence over the time. This is implemented by (i) first fitting a separate linear model for each spatial coordinate over the time and then removing the estimated time trend; and (ii) “thinning” the sequence of monthly measurements by taking a monthly record from a window of every 5 months. In Figure S15 on the supplementary materials, we use $5^\circ \times 5^\circ$ box centered at $57.5^\circ W$ longitude and $7.5^\circ S$ latitude as an example to show the effect of preprocessing. In (a) and (b), we show the data before and after the detrending; and in (c), we plot the estimated auto-correlation function for the thinned sequence. It can be seen that both detrending and thinning work quite well for that region. Similar results were also obtained for other spatial regions in our dataset.

After preprocessing, we obtain a dataset of 15 (latitude) $\times$ 68 (longitude) matrix with a sample size of $n = 48$. We check the separability of our dataset’s covariance structure via two procedures. First, we use the projection-based empirical bootstrap test (Aston, Pigoli, and Tavakoli 2017), implemented by function empirical_bootstrap_test in R package covsep. The $p$-value of the test is 0.133, which confirms the validity of the separability assumption. Second, we compare the prediction error of covariance estimators with and without the separability. In particular, we randomly split the dataset into a training set with $n_1 = 48/2 = 24$ samples and a test set with the remaining

### Table 1. Simulation results for $(p, q, \rho_1, \rho_2) = (100, 100, 0.5, 0.5)$ with the MA(1) covariance structure over 100 replications.

| $(n, p, q, \rho_1, \rho_2)$ | Method | $|| \mathbf{\Sigma} - \mathbf{\Sigma}_1 ||_F$ | $|| \mathbf{\Sigma} - \mathbf{\Sigma}_1 ||_1$ | $|| \mathbf{\Sigma} - \mathbf{\Sigma}_1 ||_2$ | $\hat{k}$ | $\hat{k}_1$ | $\hat{k}_2$ |
|--------------------------|--------|---------------------------------|---------------------------------|---------------------------------|--------|--------|--------|
| $(50, 100, 100, 0.5, 0.5)$ | Sample | 1428.40 | 1605.38 | 244.51 | 57.5 | 50, 20, 30, 0.8, 0.8 | 200, 100, 100, 0.5, 0.5 |
|                           | Banded | 91.39  | 4.08   | 2.59  | 1.09 | 1.01 | 1.01 |
|                           | Tapering | 93.60 | 4.49   | 2.80  | 2.00 | 1.01 | 1.01 |
|                           | Double B | 47.82 | 4.32   | 2.58  | 1.05 | 1.05 | 1.05 |
|                           | Double T | 77.87 | 3.96   | 3.46  | 1.78 | 1.82 | 1.82 |
|                           | Proposed B | 8.41 | 0.79   | 0.49  | 1.73 | 1.87 | 1.87 |
|                           | Proposed T | 8.71 | 0.81   | 0.51  | 2.00 | 2.00 | 2.00 |
| $(200, 100, 100, 0.5, 0.5)$ | Sample | 1004.87 | 1031.87 | 130.38 | 20.0 | 20.0 | 20.0 |
|                           | Banded | 88.68 | 3.37 | 2.29 | 1.05 | 1.05 | 1.05 |
|                           | Tapering | 89.76 | 3.63 | 2.42 | 2.00 | 2.00 | 2.00 |
|                           | Double B | 33.58 | 2.89 | 1.70 | 1.01 | 1.01 | 1.01 |
|                           | Double T | 51.88 | 4.85 | 2.37 | 2.00 | 2.00 | 2.00 |
|                           | Proposed B | 5.85 | 0.53 | 0.33 | 1.65 | 1.84 | 1.84 |
|                           | Proposed T | 6.13 | 0.56 | 0.35 | 2.00 | 2.00 | 2.00 |

NOTE: The averages of $|| \mathbf{\Sigma} - \mathbf{\Sigma}_1 ||_F$, $|| \mathbf{\Sigma} - \mathbf{\Sigma}_1 ||_1$ and $|| \mathbf{\Sigma} - \mathbf{\Sigma}_1 ||_2$ for the proposed estimators (Proposed B and Proposed T), doubly banded and tapering estimators (Double B and Double T), Bickel’s banded estimator (Banded), Cai’s tapering estimator (Tapering) and the sample covariance estimator (Sample) are reported. The averages of $\hat{k}_1$ and $\hat{k}_2$ for the proposed and doubly banded/tapering estimators, the averages of $\hat{k}$ for Bickel’s banded estimator and Cai’s tapering estimator are also reported.

### Table 2. Simulation results for heavy-tailed data with $(\rho_1, \rho_2) = (0.1, 0.1), (0.5, 0.5), (0.8, 0.8), (n, p, q) = (50, 20, 30)$, and AR(1) covariance structure over 100 replications.

| $(n, p, q, \rho_1, \rho_2)$ | Method | $|| \mathbf{\Sigma} - \mathbf{\Sigma}_1 ||_F$ | $|| \mathbf{\Sigma} - \mathbf{\Sigma}_1 ||_1$ | $|| \mathbf{\Sigma} - \mathbf{\Sigma}_1 ||_2$ | $\hat{k}$ | $\hat{k}_1$ | $\hat{k}_2$ |
|--------------------------|--------|---------------------------------|---------------------------------|---------------------------------|--------|--------|--------|
| $(50, 20, 30, 0.1, 0.1)$ | Sample | 245.38 | 590.03 | 209.47 | 1.13 | 1.13 | 1.13 |
|                           | Proposed B | 12.64 | 3.02 | 1.97 | 1.24 | 1.24 | 1.24 |
|                           | Proposed T | 11.79 | 1.66 | 1.32 | 1.10 | 1.10 | 1.10 |
|                           | Robust B | 9.05 | 1.75 | 1.05 | 1.45 | 1.45 | 1.45 |
|                           | Robust T | 8.63 | 1.17 | 0.85 | 1.20 | 1.20 | 1.20 |
| $(50, 20, 30, 0.5, 0.5)$ | Sample | 245.70 | 574.99 | 209.81 | 1.86 | 1.86 | 1.86 |
|                           | Proposed B | 24.15 | 12.72 | 7.10 | 2.05 | 2.05 | 2.05 |
|                           | Proposed T | 22.08 | 6.67 | 5.26 | 1.80 | 1.80 | 1.80 |
|                           | Robust B | 16.05 | 7.71 | 4.38 | 2.37 | 2.37 | 2.37 |
|                           | Robust T | 16.06 | 6.91 | 4.33 | 2.00 | 2.00 | 2.00 |
| $(50, 20, 30, 0.8, 0.8)$ | Sample | 249.51 | 511.77 | 212.37 | 5.65 | 5.65 | 5.65 |
|                           | Proposed B | 92.58 | 141.98 | 63.37 | 7.63 | 7.63 | 7.63 |
|                           | Proposed T | 83.38 | 117.68 | 56.57 | 7.81 | 7.81 | 7.81 |
|                           | Robust B | 48.68 | 49.61 | 28.75 | 8.60 | 8.60 | 8.60 |
|                           | Robust T | 48.56 | 49.26 | 30.45 | 9.55 | 9.55 | 9.55 |

NOTE: The averages of $|| \mathbf{\Sigma} - \mathbf{\Sigma}_1 ||_F$, $|| \mathbf{\Sigma} - \mathbf{\Sigma}_1 ||_1$ and $|| \mathbf{\Sigma} - \mathbf{\Sigma}_1 ||_2$ for robust estimators (Robust B and Robust T), our proposed estimators (Proposed B and Proposed T) and the naive sample covariance estimator (Sample) are summarized in this table. The averages of $\hat{k}_1$ and $\hat{k}_2$ for the proposed robust/nonrobust methods, the averages of $\hat{k}$ for the proposed robust methods are also reported.
The threshold parameter is chosen to be 99.99% and 99.995% for proposed banded and tapering estimators here. The large quantiles in the selected suggest the existence of outliers, which is a different data abnormality structure from the moment-based structure. This finding is also supported by the Q-Q plot in Figure S15(d) of the supplementary materials. In this case, our robust estimation procedure still works well because the initial truncation step can help remove these outliers. Meanwhile, our theoretical results mainly focus on the moment-based condition and it is still unclear how to extend these results when there exist outliers and the moment-based condition is violated. We leave it for future investigation.

We plot the estimated covariance matrices in Figure 2, including those over the latitude direction obtained by banding and tapering in (a) and (d), over the longitude direction in (b) and (e), and the overall covariance matrices in (c) and (f). All these matrices are scaled such that the maximum entry is 1. The results clearly suggest that the bandable assumption fits the data well, which is expected since the association between temperatures at two distant geographic areas is very weak. To further evaluate this assumption, we compare the magnitudes of entries removed from regularization (regularized entries) and unregularized entries, in latitude banded covariance \( \hat{\Sigma}^{R, B} \left( k_1^T \right) \) and longitude banded covariance \( \hat{\Sigma}^{R, B} \left( k_2^L \right) \). Over the latitude direction, the \( l \)th entry in \( \hat{\Sigma}^{R, B} \left( k_1^T \right) \) is a regularized entry

**Figure 2.** Temperature data analysis: Plots for the estimated covariance matrices obtained by proposed robust banded/tapering estimators: Panel (a) is the scaled robust banded latitude direction covariance \( \hat{\Sigma}^{R, B} \left( k_1^L \right) \), Panel (b) is the scaled robust banded longitude direction covariance \( \hat{\Sigma}^{R, B} \left( k_2^L \right) \) and Panel (c) is overall banded covariance \( \hat{\Sigma} \left( \hat{\Sigma} \left( k_1^L, k_2^L \right) \right) \). Panel (d) is the scaled robust tapering latitude direction covariance \( \hat{\Sigma}^{R, T} \left( k_1^L \right) \), Panel (e) is the scaled robust tapering longitude direction covariance \( \hat{\Sigma}^{R, T} \left( k_2^L \right) \) and Panel (f) is the overall tapering covariance \( \hat{\Sigma}^{R, T} \left( k_1^L, k_2^L \right) \).

### Acknowledgments

The authors wish to thank the reviewers for their constructive comments and suggestions, which greatly improved the quality of the paper. The research was supported by a grant from the National Science Foundation (NSF) under awards DMS-1712842, DMS-1712843, and DMS-1712844. The computations were performed on the Agave High Throughput Computing (HTC) system at the University of Colorado Boulder. The authors also thank the editor and the associate editor for their careful review and constructive comments. The authors would like to express their gratitude to the editor and the associate editor for their careful review and constructive comments.
if \(|l_1 - m_1| \leq \hat{k}^B_1 = 3\), and is an unregularized entry otherwise. 
Similarly, we can define the regularized and unregularized entries in \( \hat{\Sigma}_2^{R,B}(k_i^B) \). Due to (2.2), all regularized entries in \( \hat{\Sigma}_1^{R,B}(k_i^B) \) and \( \hat{\Sigma}_2^{R,B}(k_i^B) \) are zero. For comparison, we also implement our proposed robust estimation procedure without banding (using same threshold \( \tau^{R,B} \)), and obtain \( \hat{\Sigma}_1^{R,B}(15) \) and \( \hat{\Sigma}_2^{R,B}(68) \). We then plot the histograms of regularized entries in \( \hat{\Sigma}_1^{R,B}(15) \), unregularized entries in \( \hat{\Sigma}_1^{R,B}(15) \), and unregularized entries in \( \hat{\Sigma}_1^{R,B}(k_i^B) \) in Figure S16(a) of the supplementary materials. Similarly to Figure 2, maximum magnitudes of \( \hat{\Sigma}_1^{R,B}(15) \) and \( \hat{\Sigma}_1^{R,B}(k_i^B) \) are scaled to 1. The histogram clearly shows that the magnitudes of regularized entries in \( \hat{\Sigma}_1^{R,B}(15) \) are generally much smaller than those of unregularized entries in \( \hat{\Sigma}_1^{R,B}(15) \) and \( \hat{\Sigma}_1^{R,B}(k_i^B) \). This finding confirms the bandable structure over the latitude direction. Similar findings are observed in Figure S16(b) of the supplementary materials for the bandable structure over the longitude direction.

In Figure 2, for the latitude direction, there are two clusters of areas with strong correlation around 50°E and 83°W. For the longitude direction, there is a cluster of areas around 40°N that has a strong correlation (lower-right corner in (a) and (d)). Those coordinates correspond to Great Lakes (U.S.) and the Caspian Sea. To further illustrate the use of our estimated covariance for the matrix-valued data, we focus on the 5° × 5° box centered at 107.5°W longitude and 22.5°N latitude (target region: west coast side of Mexico), and estimate its covariance with the temperature at other regions. The results obtained by banding and tapering are summarized in Figure S17 of the supplementary materials. From the plot, there are two regions that have a strong correlation with the target region. The first one marked in red is the target region itself, which suggests a strong self-correlation in the neighborhood areas around the target region. The second marked in blue corresponds to ocean area (northeastern direction) near Hawaiian islands and they have a strong negative correlation, which may be related to the recent studies on Land–Ocean Surface Temperature ratio (Lambert, Webb, and Joshi 2011).

7. Discussion

In this article, we propose banded and tapering covariance estimators for matrix-valued data under a separability condition. We adopt an efficient computational algorithm and derive the convergence rates of our covariance estimates under various scenarios. To deal with heavy-tailed data, we further propose robust banded and tapering covariance estimators, and show their theoretical advantages.

Bandable covariance structure plays an important role in our methodology. In practice, it is possible that the true covariance matrix is only bandable over one direction. We note this scenario can be naturally handled by our method. For instance, suppose \( \Sigma^o_1 \in \Sigma^o \subset \Sigma^o_1 \) is not bandable, then we can choose the bandwidth as \( k_1 = p \) for the proposed banded estimator, and \( k_1 = 2p \) for the proposed tapering estimator (see Remark 2.3), to eliminate banding or tapering over the row direction.

In terms of theory, although Theorems 3.1 and 3.2 are established under which both \( \Sigma^*_1 \) and \( \Sigma^*_2 \) are assumed belonging to regularity class \( \mathcal{F}(\varepsilon_0, \alpha) \) or \( \mathcal{M}(\varepsilon_0, \alpha, \alpha) \), we find that the proof of Theorems 3.1 and 3.2 can adapt to this new case when \( \Sigma^*_2 \in \mathcal{F}(\varepsilon_0, \alpha_2) \) or \( \mathcal{M}(\varepsilon_0, \alpha_2, \alpha_2) \) with \( \alpha_2 > 0 \), and \( \Sigma^*_1 \) only needs to satisfy \( ||\Sigma^*_1||_2 \leq C_{max} \) with some constant \( C_{max} > 0 \). Keeping other conditions unchanged and selecting \( k_1 = p \) for the proposed banded estimator and \( k_1 = 2p \) for the proposed tapering estimator, the error bound in Theorem 3.1 for sub-Gaussian scenario, becomes

\[
E \left( \left| \hat{\Sigma}_2^{o}(k_2) \hat{\Sigma}_1^{o}(k_1) - \Sigma^o \right|^2 \right) \approx \frac{p^2}{\eta} + \left( \frac{\eta}{\eta + q} + \frac{qk_2}{p} \right) \left\{ \left( \frac{pk_2}{\eta} \right)^2 + \left( \frac{pk_2}{\eta + q} \right)^2 \right\} + \left( \frac{qk_2}{p} \right) \left| \Sigma^o \right|^2 \approx \frac{pk_2}{\eta} + \frac{qk_2}{p} \left| \Sigma^o \right|^2.
\]

Additionally, the error bound in Theorem 3.2 for finite fourth moment scenario becomes

\[
E \left( \left\| \hat{\Sigma}_2^{o}(k_2) \hat{\Sigma}_1^{o}(k_1) - \Sigma^o \right\|_F^2 \right) \approx \frac{p^2}{\eta} + \left( \frac{\eta}{\eta + q} + \frac{qk_2}{p} \right) \left\{ \left( \frac{pk_2}{\eta} \right)^2 + \left( \frac{pk_2}{\eta + q} \right)^2 \right\} + \left( \frac{qk_2}{p} \right) \left\| \Sigma^o \right\|_F^2 \approx \frac{pk_2}{\eta} + \frac{qk_2}{p} \left\| \Sigma^o \right\|_F^2.
\]

Besides the spectral norm consistency problem discussed in Section S.2.11 of the supplementary materials, there are a number of important directions for further investigation. First, it is still unclear whether the convergence of our proposed estimators achieves the lower bound when \( p, q \) are beyond the rate-optimal regimes that we have discussed in Section 3.3. Second, as discussed in Section S.2.11 of the supplementary materials, an efficient algorithm to solve the spectral-norm Kronecker product approximation is of interest. Third, extension of the current approach to tensor-valued data is highly nontrivial both in both computation and theory. Therefore, we leave it for future research.

Supplementary Materials

The supplementary materials include code examples for the simulations in the main paper, an additional S&P 500 stock data example, the discussions of the main theorems, additional theoretical results, and all technical proofs.

Acknowledgments

The authors would like to thank the editor, associate editor, and three reviewers for their constructive comments, which have substantially improved the article.

Funding

Shen's research was partially supported by Simons Foundation Award 51262. Kong's research was partially supported by the Natural Science and Engineering Research Council of Canada.

ORCID

Weining Shen http://orcid.org/0000-0003-3137-1085

References

Aston, J. A., Pigoli, D., and Tavakoli, S. (2017), “Tests for Separability in Nonparametric Covariance Operators of Random Surfaces,” The Annals of Statistics, 45, 1431–1461. [2622,2628]

Avella-Medina, M., Battey, H. S., Fan, J., and Li, Q. (2018), “Robust Estimation of High-Dimensional Covariance and Precision Matrices,” Biometrika, 105, 271–284. [2626]
