Proceedings of Machine Learning Research 77:208–223, 2017 ACML 2017

Rate Optimal Estimation for High Dimensional Spatial Covariance Matrices

Yi Li LI.YI3@HUSKY.NEU.EDU
Northeastern University

Aidong Adam Ding A.DING@NEU.EDU
Northeastern University

Jennifer Dy JDY@ECE.NEU.EDU
Northeastern University

Editors: Yung-Kyun Noh and Min-Ling Zhang

Abstract

Spatial covariance matrix estimation is of great significance in many applications in climatology, econometrics and many other fields with complex data structures involving spatial dependencies. High dimensionality brings new challenges to this problem, and no theoretical optimal estimator has been proved for the spatial high-dimensional covariance matrix. Over the past decade, the method of regularization has been introduced to high-dimensional covariance estimation for various structured matrices, to achieve rate optimal estimators. In this paper, we aim to bridge the gap in these two research areas. We use a structure of block bandable covariance matrices to incorporate spatial dependence information, and study rate optimal estimation of this type of structured high dimensional covariance matrices. A double tapering estimator is proposed, and is shown to achieve the asymptotic minimax error bound. Numerical studies on both synthetic and real data are conducted showing the improvement of the double tapering estimator over the sample covariance matrix estimator.

Keywords: high-dimensional statistics, covariance estimation

1. Introduction

Investigation of the covariance structure is one of the central topics in high dimensional statistics Pourahmadi (2014); Cai et al. (2016); Fan et al. (2016). The covariance matrix, as well as its inverse, plays a significant role in many applications, including classification problem in discriminant analysis Anderson (2003), portfolio selection Markowitz (1952), Gaussian graphical models Wang et al. (2016) and dimensionality reduction (e.g., principal component analysis). However, it has been shown that the sample covariance matrix has inferior performance in the high dimensional setting Johnstone (2001). On the other hand, regularization methods, including banding, tapering, and thresholding the sample covariance matrix, have been successfully applied in this area as the improvement against the inferior performance of the sample covariance matrix estimator in estimating covariance in high dimensions Bickel and Levina (2008a,b). For the high dimensional covariance estimation problem Cai et al. (2016); Fan et al. (2016), theoretical asymptotic rate optimality has been proven for various classes of structured matrices, such as bandable matrices Cai et al. (2010); Cai and Yuan (2012) and Toeplitz matrices Xiao and Wu (2012); Cai et al. (2013). Those structures

however do not describe spatial dependence and the theoretical results of rate optimal estimation for high-dimension spatial covariance matrix is yet to be proven.

Spatial data are encountered in a wide range of disciplines. With the increasing complex data model being investigated, for example in climate science Benestad et al. (2008), high dimensional spatial covariance matrix estimation is a crucial element of spatial data analysis. The spatial covariance matrix is useful in both supervised regression problem Cressie (1993); Christensen (2002); Montero et al. (2015) and unsupervised problem Benestad et al. (2015), as well as other areas like spatial econometrics LeSage and Pace (2009).

For high dimensional covariance matrix estimation, there are two types of regularized estimators based on different matrix specifications. One specification does not depend on the index structure, for which the thresholding Bickel and Levina (2008b) estimator is used. The other specification is based on the index structure (i.e., some feature dependency structure is assumed), which corresponds to banding and tapering Bickel and Levina (2008a) estimators. As is noted in Bickel and Levina (2008b), in the application of spatial data, banding or tapering the covariance matrix can be applied as long as there is an appropriate metric (for example, some underlying distance) on variable indexes. The tapering estimator has been studied in spatial statistics; while most literature focus on computational issues, see Kaufman (2008); Shaby and Ruppert (2012) and the references therein, the theoretical rate optimality is not established, as there was not a good class of structured matrices describing the spatial covariance matrix. Hence the theoretical optimality results on bandable matrices and Toeplitz matrices (which were developed for time series data) do not apply.

As noted in Zhu and Liu (2009), the theoretical treatment of spatial matrix estimation is more difficult than that of time series data (partially) because there is no natural ordering (or indexing) of the observation. In this paper, we show that spatial covariance matrices can exhibit a *block bandable* structure under the default ordering method in Zhu and Liu (2009). Bandable covariance matrix is a class of matrices that has large values when the elements are close to its diagonal and decays gradually to its upper-right and lower-left corners, which is applicable to covariance matrices of some time series. A block bandable matrix has many blocks of bandable submatrices, while the blocks also decay further from the diagonal in the pattern of a bandable matrix. We establish the minimax risk bound for block bandable matrix estimation under the Frobenius norm, which is a commonly used norm in covariance estimation problems Lam and Fan (2009); Ravikumar et al. (2011). We show that a double-tapering estimator achieves the minimax rate, which indicates that it preforms best in the worst possible case allowed in the problem, thus is indeed theoretically rate optimal. Numerical performance of the double-tapering estimator is shown through simulation studies and two applications to image compression and climate downscaling.

2. Block Bandable Spatial Covariance Estimation

In this section, we first define the class of block bandable matrices, and show its connection with the covariance matrices of two-dimensional spatial data. Then we describe the regularization estimator based on this structure of covariance matrices.

2.1. Block Bandable Spatial Covariance

Spatial data is recorded based on its geographical location, e.g., latitude and longitude. One of the main challenge for this kind of data in $\mathbb{R}^d$ with $d \geq 2$ is that, they are not ordered as the case in $\mathbb{R}^1$ such as in time series data. As a consequence, the simple structures like Toeplitz matrices Xiao and

Wu (2012); Cai et al. (2013) are not applicable for spatial data whose dimension is higher than one Zhu and Liu (2009).

We consider the case that the spatial dependence decays with respect to their spatial distances, which can be viewed as a type of generalization of the parametric isotropic covariance models Montero et al. (2015); Sherman (2011). In this case, when we index the spatial observations in a regular scheme, the resulting covariance matrices exhibits a certain trackable structure.

For example, we consider a grid of 10×10 spatial stations as shown in the right panel of Figure 1 where we index the spatial stations by raster scan. If the dependence decays exponentially with the underlying spatial distance, the covariance structure is shown as the left panel in Figure 1 (whiter pixels correspond to higher values, and redder corresponds to lower values).

Generally speaking, if we have $p = p_1 \times p_2$ spatial locations with p_1 rows and p_2 columns, and index them by raster scan order from 1 to p [1], the resulting $p \times p$ dimensional covariance matrix has $p_1 \times p_1$ blocks of sub-matrices. Each sub-matrix is of $p_2 \times p_2$ dimensions and is a bandable matrix Bickel and Levina (2008a); Cai et al. (2010). In this way, the spatial observatory station in the i^{th} row and j^{th} column corresponds to the $[p_2(i-1)+j]^{th}$ indexed element used for constructing the covariance matrix. Moreover, the values on the sub-diagonals of the blocks closer to the main diagonal tend to be greater than the values on the sub-diagonals of those blocks more apart from the main diagonal. We term this kind of matrices the *block bandable matrices*, of which the class is formally defined next. Thus, the block bandable matrices can be used to model the dependency structure of two-dimensional data that decays according to its spatial distance. It is worth noting that matrices that exhibit block structure plays a significant role in spatial modeling, such as the circulate embedding of random fields Lord et al. (2014).

Let $\Sigma_{p\times p}$ be the covariance matrix of dimensions $p \times p$ with $p = p_1 p_2$. It consists of block sub-matrices Σ^{st} with $1 \le s \le p_1$, and $1 \le t \le p_1$. Here $\Sigma^{st} = \{\sigma_{ij}^{st}\}_{1\le i,j\le p_2}$ with σ_{ij}^{st} being the element at the i^{th} row and j^{th} column of the block sub-matrix. We then focus the theoretical study on the following class of block bandable matrices:

$$\begin{aligned}\mathscr{B}_{\alpha,\beta}(M,\varepsilon) =& \{\Sigma : |\sigma_{ij}^{st}| \le M(|i-j|^{-\alpha} \wedge |s-t|^{-\beta}), \forall 1 \le i \ne j \le p_2, 1 \le s \ne t \le p_1, \\ & \varepsilon \le \lambda_{\min}(\Sigma^{st}), \lambda_{\max}(\Sigma^{st}) \le \frac{1}{\varepsilon}, \forall s,t\},\end{aligned} \tag{1}$$

where $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ correspond to the minimal and maximal eigenvalue of a matrix A. The $\wedge$ denotes the minimum operator throughout the paper. The first condition reflects the bandable structure in both directions. The covariance decays, in the tail, at exponential rates bounded by α and β respectively along the latitude and along the longitude. The second condition imposes the boundedness of the eigenvalues to ensure non-singularity. This class is an extension of the classes studied by Cai et al. (2010); Bickel and Levina (2008a); Cai and Yuan (2012); Xue and Zou (2013) which are suited for covariances among locations in a equidistant one-dimensional grid, while our class $\mathscr{B}_{\alpha,\beta}(M,\varepsilon)$ is the corresponding class for the two-dimensional grid.

2.2. Double Tapering Estimation

We assume that the observed normalized data $\{X_i\}_{i=1}^n$ is drawn from a p dimensional normal distribution with zero mean and covariance matrix Σ. The (maximal likelihood estimator) sample

1. Such an ordering method is termed *default ordering* in Zhu and Liu (2009), which shows relatively better results than other ordering methods discussed therein.

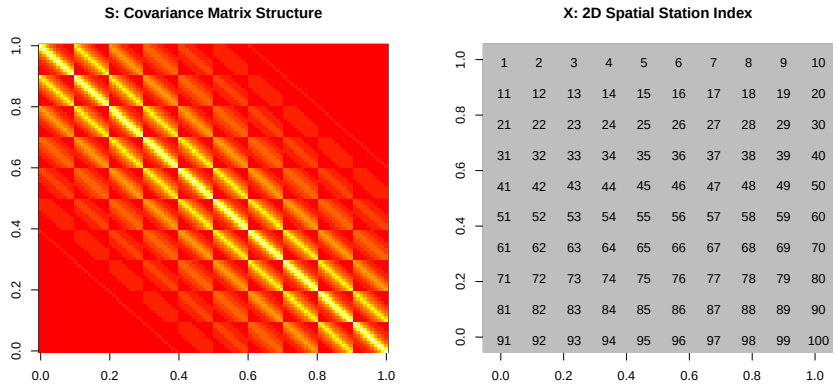

Figure 1: Left: Example of the spatial covariance matrix. Right: The two-dimensional stations are indexed in a raster scan order to build a one dimensional long vector.

covariance matrix is $\tilde{\Sigma}$. We define the double tapering covariance matrix $\hat{\Sigma}$, consisting of sub-block matrices $\hat{\Sigma}^{st}$, as

$$\hat{\Sigma}^{st} = \{\omega(\frac{|i-j|}{k})\tilde{\sigma}_{ij}^{st}\}_{1\le i,j\le p_2}, \qquad \hat{\Sigma} = \{\omega(\frac{|s-t|}{l})\hat{\Sigma}^{st}\}_{1\le s,t\le p_1}, \tag{2}$$

where $\omega(x)$ is a tapering function which is non-increasing and equals to one near zero, and zero when x is large [2]. The two-step tapering estimator corresponds to regularization in both dimensions, i.e. inside each sub-block matrix and across all sub-block matrices. Thus, we name it the *double tapering estimator* for the spatial block bandable matrix. For example, banding function could be defined as $\omega(x) = \mathbf{1}_{[0,\frac{3}{4}]}(x)$, while linear tapering function could be defined as follow,

$$\omega(x) = \begin{cases} 1, & \text{for} \quad x < 0.5, \\ 2-2x, & \text{for} \quad 0.5 \le x \le 1, \\ 0, & \text{for} \quad x > 1. \end{cases} \tag{3}$$

Other possible tapering and banding function $\omega(\cdot)$ are plotted in Figure 2.

3. Statistical Error

In this section, we first study the upper bound of the asymptotic minimax risk for the proposed double tapering estimator. Then the asymptotic minimax lower bound over the class of block bandable covariance matrices is derived, which shows the rate optimality of our proposed method. We consider the high-dimensional asymptotic case of all n, p_1 and p_2 increases to infinity.

3.1. Upper Bound

To derive the upper bound of the tapering estimator, we will adopt the so-called norm compression (in)equality (NCI) as a tool to deal with our block bandable matrix. NCI has been studied with

2. The tapering function $\omega(\cdot)$ is similar to the kernel function, but is not restricted to the unit integration.

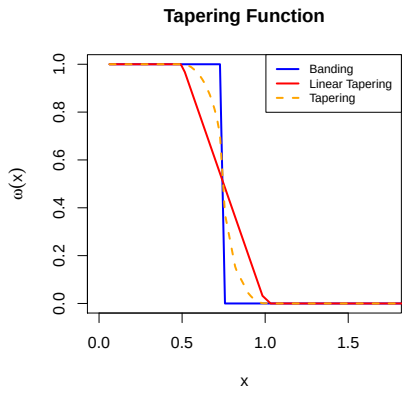

Figure 2: The tapering function $\omega(x)$ which is used in the double tapering estimator.

various matrix norms King (2003); Audenaert (2006), which is of independent interest. Here we will use the Frobenius norm, the square-root of the sum of squares of every element in the matrix, denoted as $\|\cdot\|_F$. Let matrix A be partitioned into $d_1 \times d_2$ blocks of sub-matrices A_{ij} as:

$$A = \begin{bmatrix} A_{11} & A_{12} & \dots & A_{1d_2} \\ A_{21} & A_{22} & \dots & A_{2d_2} \\ \dots & \dots & \dots & \dots \\ A_{d_1 1} & A_{d_1 2} & \dots & A_{d_1 d_2} \end{bmatrix}. \tag{4}$$

If we replace each A_{ij} with its (Frobenius) norm, then we get a "compressed" version of A, denoted by $N(A)$:

$$N(A) = \begin{bmatrix} \|A_{11}\| & \|A_{12}\| & \dots & \|A_{1d_2}\| \\ \|A_{21}\| & \|A_{22}\| & \dots & \|A_{2d_2}\| \\ \dots & \dots & \dots & \dots \\ \|A_{d_1 1}\| & \|A_{d_1 2}\| & \dots & \|A_{d_1 d_2}\| \end{bmatrix}. \tag{5}$$

The NCIs establish the relationship between $\|A\|$ and $\|N(A)\|$, which becomes an equality for the Frobenius norm.

Lemma 1 (NCI for Frobenius norm) *Let A be a matrix that could be partitioned as in (4), and $N(A)$ be its compressed matrix as defined in (5) with Frobenius norm. We have*

$$\|A\|_F = \|N(A)\|_F. \tag{6}$$

Proof Since $\|A\|_F^2$ is the sum of squares of every elements in the matrix while $\|A_{i_1 i_2}\|_F^2$ is the sum of squares of every elements in $A_{i_1 i_2}$, the i_1^{th} and i_2^{th} sub-matrix of the partitioned A, so obviously

$$\|A\|_F^2 = \sum_{i_1=1}^{d_1} \sum_{i_2=1}^{d_2} \|A_{i_1 i_2}\|_F^2 = \|N(A)\|_F^2. \tag{7}$$

■

This implies that the risk of the block bandable covariance matrix under Frobenius norm could be broken down into the sum of the risk corresponding to each individual small block. Based on this fact, we first derive the element-wise error and combine the results to get the risk upper bound under the Frobenius norm, which leads to the following theorem (proof in Appendix A).

Theorem 2 (Upper Bound) *Let the double tapering estimator $\hat{\Sigma}_k$ of the covariance matrix $\Sigma_{p\times p}$ be defined in (2). Then we have the following upper bound for the asymptotic minimax risk over the block bandable matrices class $\mathscr{B}_{\alpha,\beta}$ with $\alpha, \beta > 1$, as $n \to \infty$, $p1 \to \infty$ and $p_2 \to \infty$,*

$$\sup_{\mathscr{B}_{\alpha,\beta}} \mathbb{E}\frac{1}{p}\left\|\hat{\Sigma} - \Sigma\right\|_F^2 \leq Cn^{-\frac{(2\alpha-1)(2\beta-1)}{4\alpha\beta-1}} p_1^{\frac{2\beta-1}{4\alpha\beta-1}} p_2^{\frac{2\alpha-1}{4\alpha\beta-1}} \wedge \frac{p}{n}. \tag{8}$$

3.2. Lower bound

In this part, we discuss the optimality, i.e. the minimax lower bound for the estimation over the class $\mathscr{B}_{\alpha,\beta}$ of block bandable matrices. The proof applys a similar approach as in Cai et al. (2010). To show the above upper bound is optimal, we only need to show that the risk is also bounded below by Ckl/n for such

$$l = n^{\frac{2\alpha-1}{4\alpha\beta-1}} p_1^{-\frac{1}{4\alpha\beta-1}} p_2^{\frac{2\alpha}{4\alpha\beta-1}} \wedge \frac{p_1}{2}. \qquad k = n^{\frac{2\beta-1}{4\alpha\beta-1}} p_1^{\frac{2\beta}{4\alpha\beta-1}} p_2^{-\frac{1}{4\alpha\beta-1}} \wedge \frac{p_2}{2}, \tag{9}$$

In this paper, C denotes a general positive constant. Thus the C here can differ from the constant C in other places including in Theorem 2.

We first describe a version of Assouad's Lemma, a standard tool to get the minimax lower bound, whose proof is in Yu (1997); Tsybakov (2008); van der Vaart (1998). Let $\Omega = \{\omega = (\omega_1, \cdots, \omega_N) : \omega_j \in \{0, 1\}\}$ be the set of binary sequences of length N. Let $\mathscr{P} = \{\mathbb{P}_\omega : \omega \in \Omega\}$ be a set of 2^N distributions indexed by the elements of Ω. Let $H(\omega, \nu) = \sum_{j=1}^N \mathbf{1}_{\omega_j \neq \nu_j}$ be the *Hamming distance* between $\omega, \nu \in \Omega$.

Lemma 3 (Assouad van der Vaart (1998)) *Let $\mathscr{P} = \{\mathbb{P}_\omega : \omega \in \Omega\}$ be a set of distributions indexed by Ω and let $\theta(\mathbb{P})$ be a parameter. For any $m > 0$ and any metric d,*

$$\begin{aligned} &\max_{\omega\in\Omega} 2^m \mathbb{E}_\omega[d^m(\hat{\theta}, \theta(\mathbb{P}_\omega))] \\ \geq &\frac{N}{2} \min_{H(\omega,\nu)\geq 1} \frac{d^m(\theta(\mathbb{P}_\omega), \theta(\mathbb{P}_\nu))}{H(\omega, \nu)} \cdot \min_{H(\omega,\nu)=1} \|\mathbb{P}_\omega \wedge \mathbb{P}_\nu\|. \end{aligned} \tag{10}$$

Here, $\|\mathbb{P} \wedge \mathbb{Q}\| := \lambda(p \wedge q)$ is the affinity (or the total variational distance) of two probability measures $\mathbb{P}$ and $\mathbb{Q}$ with density p and q with respect to common dominating measure λ Pollard (2002).

To establish the minimax lower bound with Lemma 3, we construct a finite collection of normal distributions with the covariance matrices as a subset of $\mathscr{B}_{\alpha,\beta}$, and separately bound the Frobenius norm and the affinity on this subset. The σ_{ij}^{st} and σ_{ji}^{ts} in the covariance matrix represent the symmetric elements and have the same value. Therefore, we consider the following class of covariance matrices where we perturb each pair by a fixed amount τ.

$$\begin{aligned} \mathscr{B}_2 = &\{\Sigma(\theta) : \Sigma(\theta) = I_p + (\theta_{ij}^{st}\tau\frac{1}{\sqrt{n}}\mathbf{1}_{1\leq|i-j|\leq k, 1\leq|s-t|\leq l})_{p\times p} \\ &\forall \theta_{ij}^{st} = \theta_{ji}^{ts} = 0 \text{ or } 1, \forall 1 \leq s, t \leq p_1, 1 \leq k, l \leq p_2\}, \end{aligned} \tag{11}$$

where the binary θ_{ij}^{st} indicates whether the $(i,j)^{th}$ element in the $(s,t)^{th}$ sub-block matrix is perturbed, and $0 < \tau < M$. As $n \to \infty$ and for k, l in (9), it is easy to see that $\mathscr{B}_2 \subset \mathscr{B}_{\alpha,\beta}$, allowing us to establish the asymptotic minimax rate. Thus, in terms of Lemma 3, we have $\theta \in \Omega = \{0,1\}^N$ with $N \asymp klp_1p_2$. Here and in the following, the $\asymp$ denote that the asymptotic order is the same. That is, $\xi_n \asymp \eta_n$ if there exist a positive constant ϵ such that $\epsilon \leq \frac{\xi_n}{\eta_n} \leq \frac{1}{\epsilon}$. Since any two elements θ and θ' in Ω differs in exactly $H(\theta,\theta')$ components, we have

$$\left\|\Sigma(\theta) - \Sigma(\theta')\right\|_F^2 = \frac{\tau^2}{n} \sum_{s,t,i,j} |\theta_{ij}^{st} - (\theta')_{ij}^{st}|^2 = \frac{\tau^2}{n} H(\theta,\theta') \tag{12}$$

To bound the affinity, we first relate it to Kullback-Leibler (KL) divergence by the following Lemma (see Tsybakov (2008) for proof).

Lemma 4 *Let $KL(\mathbb{P},\mathbb{Q})$ be the KL divergence between two probability measures $\mathbb{P}$ and $\mathbb{Q}$, we have*

$$\|\mathbb{P} \wedge \mathbb{Q}\| \geq \frac{1}{2} e^{-KL(\mathbb{P},\mathbb{Q})}. \tag{13}$$

Using this, we have the following lower bound for the affinity (proof in Appendix B).

Lemma 5 *Let $\mathbb{P}_\theta \sim N(0,\Sigma(\theta))$ be the joint distribution of i.i.d sample $\{X_i\}_{i=1}^n$ with $\Sigma(\theta) \in \mathscr{B}_2$. Then*

$$\min_{H(\theta,\theta')=1} \left\|\mathbb{P}_\theta \wedge \mathbb{P}_{\theta'}\right\| \geq C_1. \tag{14}$$

Using Frobenius norm as the metric in Lemma 3 and $m = 2$, combining equations (12) and (14) with $C = C_1\tau^2$, we get the lower bound on risk.

Theorem 6 *The asymptotic minimax risk for estimating the covariance matrix Σ over $\mathscr{B}_2$ under the Frobenius norm satisfies, as $n \to \infty$ and for k, l in (9),*

$$\inf_{\hat{\Sigma}} \sup_{\mathscr{B}_2} \mathbb{E} \frac{1}{p} \left\|\hat{\Sigma} - \Sigma\right\|_F^2 \geq C \frac{kl}{n}, \tag{15}$$

for some $C > 0$.

Combining Theorem 2 and Theorem 6, we have the exact optimal minimax risk rate.

Corollary 7 *The minimax risk for estimating the covariance matrix Σ under Frobenius norm satisfies*

$$\inf_{\hat{\Sigma}} \sup_{\mathscr{B}_{\alpha,\beta}} \mathbb{E} \frac{1}{p} \left\|\hat{\Sigma} - \Sigma\right\|_F^2 \asymp n^{-\frac{(2\alpha-1)(2\beta-1)}{4\alpha\beta-1}} p_1^{\frac{2\beta-1}{4\alpha\beta-1}} p_2^{\frac{2\alpha-1}{4\alpha\beta-1}} \wedge \frac{p}{n}. \tag{16}$$

And this optimal minimax risk rate is achieved by the double tapering estimator.

The tapering in each direction should be adjusted according to the decay rate in that direction to achieve the optimal rate. The double tapering estimator is then rate optimal over the class $\mathscr{B}_{\alpha,\beta}$ when $\alpha, \beta > 1$. For the simple case of $\alpha = \beta$ (the decay rates are the same along the latitude as along the longitude), the optimal minimax risk rate is $O(n^{-\frac{(2\alpha+1)^2}{4(\alpha+1)^2-1}} p^{\frac{2\alpha+1}{4(\alpha+1)^2-1}})$, which only depends on the total size of the grid $p = p_1p_2$ but not on the specific values of p_1 and p_2.

4. Experimental Results

In this section, we study the numerical performance of our estimators. We first apply our method on synthetic data to demonstrate its performance under various decay rates, feature dimensions and the sample size. Then, we test it with real data in an application of image compression using principal component analysis (PCA) and a climate application with statistical downscaling.

4.1. Synthetic Data

The simulation of block bandable covariance matrices is based on the class $\mathscr{B}_{\alpha,\beta}$ in (1). Specifically, let the true covariance matrix Σ have the form

$$\sigma_{i,j}^{st} = \begin{cases} M(|i-j|^{-\alpha} \wedge |s-t|^{-\beta}), & i \neq j \text{ or } s \neq t \\ 2.5, & i = j, s = t \end{cases} \tag{17}$$

where $M = 1, \alpha = \beta = a$. The multivariate Gaussian random data is generated with mean zero and covariance matrix Σ. The rate of the parameters k and l are decided by the result in equation (9). Each of the simulation study is based on the average results from 100 replications.

In particular, we are interested in the relative performance of the regularized estimators (banding and tapering) under the Frobenius norm against the sample covariance matrix. Several factors are considered in our simulation study including the decay rate (a), dimension (p) and sample size (n). We focus on the square cases where $p_1 = p_2$, thus $p = p_1 p_2 = p_1^2$.

Decay Rate. We consider the decay rate $a = 1.1, 1.2, 1.3, 1.4$ and 1.5. The error for each setting is measured by the Frobenius norm of the difference between the (regularized) estimators and the true covariance matrix. The relative error, i.e. the ratio of the error for regularized estimators with respect to the sample covariance matrix is considered. Thus, smaller relative error (less than one) implies better performance than the sample covariance matrix.

Result of this comparison is presented in Figure 3. The sample size is set to be $n = 1000$, while the dimension is $p = 3600$. The red bar corresponds to the sample covariance matrix to itself, which is always one, while the blue and green bar correspond to the (double) banding and (double) linear tapering estimators, which reduce the relative error of the sample covariance matrix. Thus, both the banding and linear tapering estimator significantly improve the performance of the sample covariance matrix.

Dimension. We now compare cases where $p_1 = p_2 = 10, 20, 30, 40, 50, 60$, thus $p = 100, 400, ..., 3600$. Result of this comparison is presented in Figure 4. The sample size is set to be $n = 1000$, while the decay rate is $a = 1.5$. The red dashed line corresponds to the sample covariance matrix to itself, which is one, while the blue and green lines correspond to the banding and linear tapering estimators respectively. As we can see from the plot, both the banding and linear tapering estimator improve the performance of the sample covariance matrix more as the dimension increases.

Sample Size. Similar to the previous settings, we consider sample size n ranging from 250 to 3000. Result of this comparison is presented in Figure 4. The dimension is set to be $p = 3600$, while the decay rate is $a = 1.5$. The red dashed line corresponds to the sample covariance matrix to itself, while the blue and green lines correspond to the banding and linear tapering estimators respectively, which greatly reduce the relative error of the sample covariance matrix.

A more comprehensive simulation study results are reported in the supplemental materials. The overall result shows the effectiveness of our double tapering estimator.

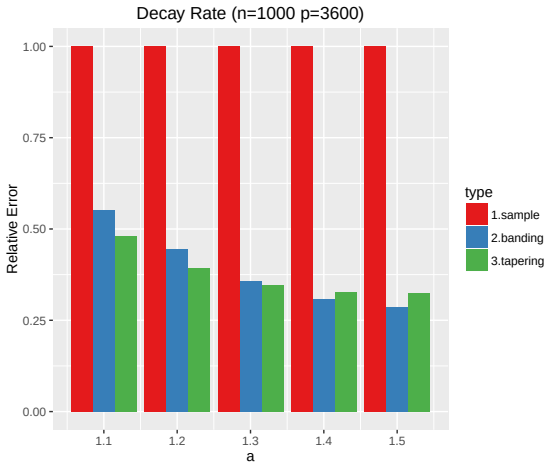

Figure 3: Relative error (w.r.t sample covariance matrix) for the banding and linear tapering estimator. The comparison is against the decay rate a.

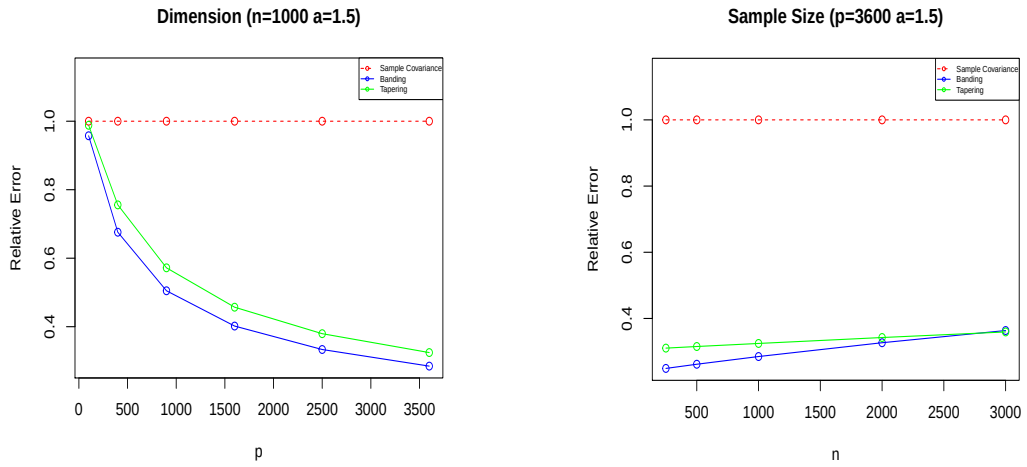

Figure 4: Relative error (w.r.t. sample covariance matrix) for the banding and linear tapering estimators. The comparison is on the left: against the dimension p; on the right: against the sample size n.

4.2. Application on Image Compression

Image compression aims to reduce the size of the image data in order to store or transmit data efficiently. Principal component analysis is a classical way to perform dimensionality reduction and to compress images. In this part, we compare the performance of the image compression and

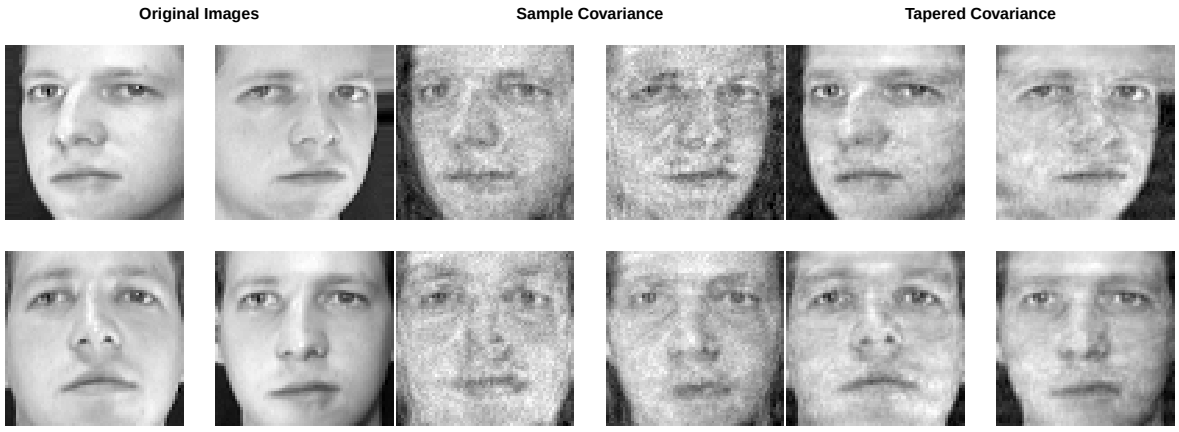

Figure 5: Examples of the face data set. The three groups (four images each) from left to right: original images, the reconstructed images by keeping $q = 1000$ principal components based on the sample covariance matrix and based on the double tapered covariance matrix.

reconstruction results using principal component analysis (PCA) based on the sample covariance matrix versus the results using PCA based on the proposed double tapered covariance estimator.

As an example, we consider the Olivetti face data[3] which contains 400 face images of 40 persons and each image is of size 64 by 64. We use 40% of the dataset for training, and the rest 60% dataset for testing.

In the training period, we estimate the mean of the features and the principal subspace spanned by the top q principal components based on the training data. For the proposed double tapering estimator, the parameters $k = 85$ and $l = 9$ are chosen via five-fold cross-validation within the training data. In the testing period, we compressed the testing images by projecting the data onto the trained principal subspace. The compressed images are then reconstructed with the top q principal components and the feature mean. The reconstruction error is based on the Frobenius norm between the reconstructed images and the true testing images.

The left four plots of Figure 5 shows four example original images. The reconstructed images based on sample covariance matrix and the double tapered covariance matrix are presented respectively in the middle four plots and the right four plots of Figure 5 under the same compression ratio, i.e. keeping the same number ($q = 1000$) of principal components. By keeping the same amount of information, our proposed method reconstructs clearer images compared with using the sample covariance matrix. This is confirmed with the reconstruction error plot in Figure 6.

4.3. Application on Climate Data

In this part, we apply our proposed spatial covariance estimation in a real climate task, i.e. statistical downscaling (Benestad et al., 2008, 2015). It is generally believed that the Global Climate Models

3. This dataset contains a set of face images taken at AT&T Laboratories Cambridge, which is available via the following link `scikit − learn.org/stable/datasets/olivetti_faces.html`

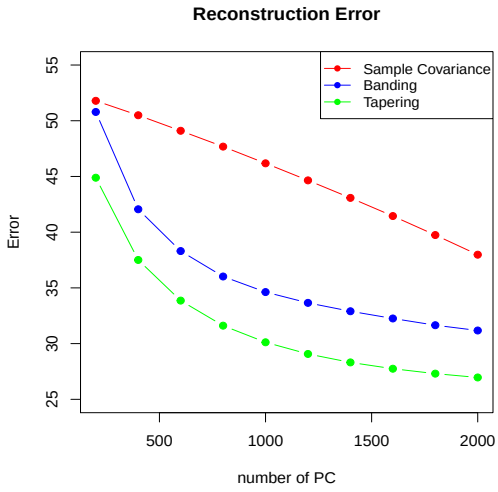

Figure 6: Reconstruction error on the testing set as a function of the number of principal component kept. The double tapered covariance matrix leads to lower reconstruction error compared with the sample covariance matrix.

(GCMs) provide coarse resolution outputs which preclude their application to accurately assess the effects of climate change on finer regional scale events. Statistical downscaling are methods that use statistical models to infer the local-scale climate information from coarsely resolved climate models. Figure 7 is a typical example of the data for this climate application.

In this experiment, we consider the reanalysis monthly mean temperature data from the National Oceanic & Atmospheric Administration (NOAA) [4] as the coarse resolution data, which includes grid points from the world atlas, while the finer resolution temperature data is from the University of Idaho Gridded Surface Meteorological Data (UofI METDATA) [5], which contain the data of contiguous united states.

We use the common time range, years 1979-2013, from the two data sets for training and testing. We choose years 1979-2008 data as the training set, and years 2009-2013 as the testing set. We normalize the data by removing the historical long range mean for each month. The spatial covariance matrix is obtained by the sample covariance between each grid point. Moreover, the double-tappered spatial covariance can be derived based on previous discussion. Since the features can be highly correlated in this task, we apply hierarchical clustering method to cluster the data and using the center of each cluster as the derived features for the benchmark predictive model (10-fold cross-validated LASSO). As we can see from Figure 8 and 9, result shows that the predict performance (testing MSE) based on the double-tappered spatial covariance matrix is better than the one with the sample covariance for both season-wise and overall comparison.

4. https://www.esrl.noaa.gov/psd/data/gridded/data.ncep.reanalysis.html
5. http://metdata.northwestknowledge.net/

Figure 7: Example of the statistical downscaling task. Left: the climate variable (monthly mean temperature in Kelvin) with a coarse resolution at a single time point (Jan. 2009). Right: the target climate variable with a relatively finer resolution. The goal of statistical downscaling is to learn the target climate variable in finer resolution based on the historical data of the fine and coarse resolution climate data (red is larger and black is smaller).

Figure 8: Downscaling result of the target climate variable in the testing set. The intensity (red is larger and black is smaller) is the average MSE in each grid point.

5. Conclusions

With the large amount of spatial data becoming available, high-dimensional spatial covariance matrix estimation is becoming more important for data analysis. Here we provide the first minimax rate optimality results of tapering estimators in this high-dimensional setting. To represent the spatial structure, we propose the block bandable matrices structure to incorporate spatial information. Accordingly, a double tapering estimator is proposed and is shown to achieve the optimal minimax rate over a class of block bandable covariance matrices. Numerical study confirms the performance of the proposed estimator.

Appendix A. Proof of Theorem 2

Let $X_m = (X_m^{(1)}, \cdots, X_m^{(p)})^\top$, $m = 1, 2, \cdots, n$. It is easy to see that $\tilde{\sigma}_{ij}^{st}$, the $(i,j)^{th}$ element of the $(s,t)^{th}$ sub-matrix corresponds to the $(i^*, j^*)^{th}$ element of the global matrix $\tilde{\Sigma}$, where $i^*(s,i) = (s-1)p_1 + i$ and $j^*(t,j) = (t-1)p_2 + j$. Therefore, we have $\tilde{\sigma}_{ij}^{st} = \frac{1}{n}\sum_{m=1}^{n} X_m^{(i^*(s,i))} X_m^{(j^*(t,j))}$. In the rest of the proof, we will write $i^*(s,i)$ and $j^*(t,j)$ more compactly as i^* and j^* if their

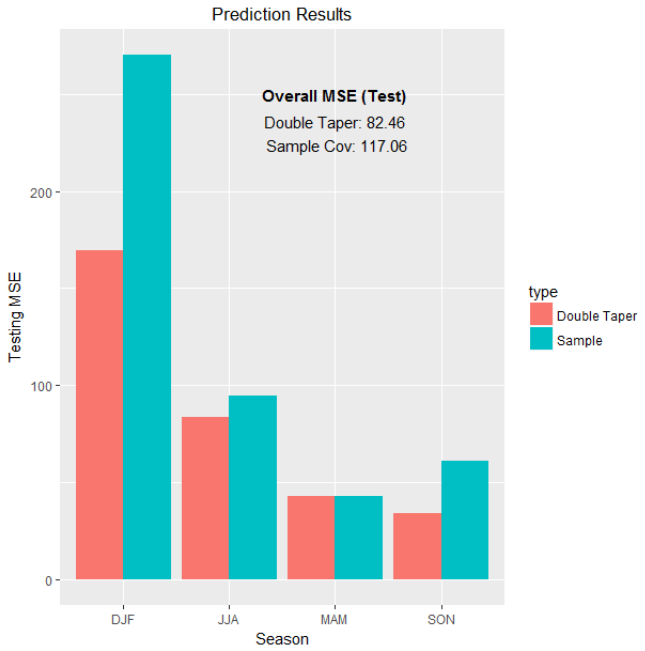

Figure 9: Testing MSE across four seasons: DJF (December, January, February), MAM (March, April, May), JJA (June, July, August), SON (September, October, November). The overall MSE is also presented.

meaning is clear from the text. Now we have $\mathbb{E}\tilde{\sigma}_{ij}^{st} = \sigma_{ij}^{st}$, and $Var(\tilde{\sigma}_{ij}^{st}) = \frac{1}{n}Var(X_m^{(i^*)}X_m^{(j^*)}) \leq \frac{1}{n}\mathbb{E}[(X_m^{(i^*)}X_m^{(j^*)})^2] \leq \frac{1}{n}\sqrt{\mathbb{E}(X_m^{(i^*)})^4}\sqrt{\mathbb{E}(X_m^{(j^*)})^4} \leq \frac{C}{n}$.

For the two-step definition of (2) in the original text, denote the coefficient (weight) of $\tilde{\sigma}_{ij}$ inside the sub-block matrix in the first step as $W = \{w_{ij}^{st}\}$, while denote the weight in the second step among sub-block matrices by $V = \{v_{ij}^{st}\}$. Then we have $\hat{\Sigma} = V \circ W \circ \tilde{\Sigma}$, where $\circ$ is the Hadamard product. Thus for the tapering estimator in (2) in the original text, the bias of $\hat{\sigma}_{ij}^{st}$ is $(1 - v_{ij}^{st}w_{ij}^{st})\sigma_{ij}^{st}$ and the variance of $\hat{\sigma}_{ij}^{st}$ is $(v_{ij}^{st}w_{ij}^{st})^2 Var(\tilde{\sigma}_{ij}^{st})$. Using the bound on the variance above, we have the element-wise risk bound

$$\mathbb{E}(\hat{\sigma}_{ij}^{st} - \sigma_{ij}^{st})^2 = \mathbb{E}(v_{ij}^{st}w_{ij}^{st}\tilde{\sigma}_{ij}^{st} - \sigma_{ij}^{st})^2 \leq (1 - v_{ij}^{st}w_{ij}^{st})^2(\sigma_{ij}^{st})^2 + (v_{ij}^{st}w_{ij}^{st})^2\frac{C}{n}. \tag{18}$$

Apply this element-wise risk bound with Lemma 1, we get

$$\frac{1}{p}\mathbb{E}\left\|\hat{\Sigma} - \Sigma\right\|_F^2 \leq \frac{1}{p}\sum_{1\leq s,t\leq p_1}\sum_{1\leq i,j\leq p_2}[(1 - v_{ij}^{st}w_{ij}^{st})^2(\sigma_{ij}^{st})^2 + (v_{ij}^{st}w_{ij}^{st})^2\frac{C}{n}]$$

Note that when both $|i - j| < \frac{k}{2}$ and $|s - t| < \frac{l}{2}$, $(1 - v_{ij}^{st}w_{ij}^{st})^2(\sigma_{ij}^{st})^2 = 0$ which can be dropped. Otherwise, it can be upper bounded by $(\sigma_{ij}^{st})^2$. Likewise, when either $|i - j| > k$ or $|s - t| > l$, $(v_{ij}^{st}w_{ij}^{st})^2\frac{C}{n} = 0$. Otherwise, it can be upper bounded by $\frac{C}{n}$. Taken together, the

above expression can be upper bounded by the sum of $R_1 = \frac{1}{p}\sum_{|s-t|\le l}\sum_{|i-j|\le k}\frac{C}{n}$ and $R^* = \frac{1}{p}\sum_{|s-t|\ge \frac{l}{2} \text{ or } |i-j|\ge\frac{k}{2}}(\sigma_{ij}^{st})^2$. Furthermore, the second term $R^* \le R_2 + R_3 + R_4$ where

$$R_2 = \frac{1}{p}\sum_{|s-t|<\frac{l}{2}}\sum_{|i-j|\ge\frac{k}{2}}(\sigma_{ij}^{st})^2, \qquad R_3 = \frac{1}{p}\sum_{|s-t|\ge\frac{l}{2}}\sum_{|i-j|<\frac{k}{2}}(\sigma_{ij}^{st})^2, \qquad R_4 = \frac{1}{p}\sum_{|s-t|\ge\frac{l}{2}}\sum_{|i-j|\ge\frac{k}{2}}(\sigma_{ij}^{st})^2.$$

Then we have

$$R_1 = \frac{1}{p}\sum_{|s-t|\le l}\sum_{|i-j|\le k}\frac{C}{n} \le \frac{1}{p}Clp_1kp_2\frac{1}{n} = C\frac{kl}{n}$$
$$R_2 = \frac{1}{p}\sum_{|s-t|<\frac{l}{2}}\sum_{|i-j|\ge\frac{k}{2}}(\sigma_{ij}^{st})^2 \le \frac{C}{p}lp_1p_2\frac{1}{k^{2\alpha-1}} = Clk^{-2\alpha+1}$$

Similarly, we have $R_3 \le Ckl^{-2\beta+1}$. Moreover,

$$R_4 = \frac{1}{p}\sum_{|s-t|\ge\frac{l}{2}}\sum_{|i-j|\ge\frac{k}{2}}(\sigma_{ij}^{st})^2 \le \frac{1}{p}C[p(p_1-\frac{l}{2})k^{-2\alpha+1} \wedge p(p_2-\frac{k}{2})l^{-2\beta+1}]$$

Summing over R_1, R_2, R_3 and R_4, the above inequalities imply that for some constant $C > 0$,

$$\begin{aligned} E\frac{1}{p}\left\|\hat{\Sigma}-\Sigma\right\|_F^2 &\le C\frac{kl}{n} + C[\frac{l}{2}k^{-2\alpha+1} + \frac{k}{2}l^{-2\beta+1} + (p_1-\frac{l}{2})k^{-2\alpha+1} + (p_2-\frac{k}{2})l^{-2\beta+1}] \\ &= C(\frac{kl}{n} + p_1k^{-2\alpha+1} + p_2l^{-2\beta+1}) \\ &\le Cn^{-\frac{(2\alpha-1)(2\beta-1)}{4\alpha\beta-1}}p_1^{\frac{2\beta-1}{4\alpha\beta-1}}p_2^{\frac{2\alpha-1}{4\alpha\beta-1}}. \end{aligned} \tag{19}$$

In the last step, the optimal k and l is chosen to make all three terms in the second to last expression to be of the same order. That is,

$$k^* = n^{\frac{2\beta-1}{4\alpha\beta-1}}p_1^{\frac{2\beta}{4\alpha\beta-1}}p_2^{-\frac{1}{4\alpha\beta-1}}, \qquad l^* = n^{\frac{2\alpha-1}{4\alpha\beta-1}}p_1^{-\frac{1}{4\alpha\beta-1}}p_2^{\frac{2\alpha}{4\alpha\beta-1}} \tag{20}$$

In the case of $k^* > p_2$ and $l^* > p_1$, we take $k = p_2$ and $l = p_1$. Thus

$$\mathbb{E}\frac{1}{p}\left\|\hat{\Sigma}-\Sigma\right\|_F^2 \le C\frac{p}{n}. \tag{21}$$

The equations (19) and (21) provides the upper bound in Theorem 2.

Appendix B. Proof of Lemma 5

Using the Lemma 4, we only need to bound from above the KL divergence between two perturbations of the distributions within the class.

$$KL(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) = n[\frac{1}{2}Tr(\Sigma(\theta')\Sigma^{-1}(\theta)) - \frac{1}{2}\log\det(\Sigma(\theta')\Sigma^{-1}(\theta)) - \frac{p}{2}] \tag{22}$$

Let $\Delta = \Sigma(\theta') - \Sigma(\theta)$. Then, we have

$$Tr(\Sigma(\theta^{'})\Sigma^{-1}(\theta)) - p = Tr(\Delta\Sigma^{-1}(\theta)). \tag{23}$$

Let the eigenvalues of $\Delta\Sigma^{-1}(\theta)$ be $\{\lambda_i\}_{i=1}^p$. Note that the eigenvalues of $I_p + \Delta\Sigma^{-1}(\theta)$ are $1 + \lambda_i$s, and with Taylor expansion we have

$$\log\det(\Sigma(\theta^{'})\Sigma^{-1}(\theta)) = \log\det(I + \Delta\Sigma^{-1}(\theta)) = Tr(\Delta\Sigma^{-1}(\theta)) - r_0, \tag{24}$$

with $r_0 \leq C_0 \sum_{i=1}^p \lambda_i^2$ for some constant C_0.

Using (23) and (24), the trace term cancels in (22), and we have $KL(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) \leq nr_0/2$. Since $\|\Delta\|_F^2$ is of order $1/n$, $KL(\mathbb{P}_\theta, \mathbb{P}_{\theta'})$ is bounded by a positive constant C. Then (13) implies $\min_{H(\theta,\theta')=1} \left\|\mathbb{P}_\theta \wedge \mathbb{P}_{\theta'}\right\| \geq C_1$, for some constant $C_1 = 0.5e^{-C} > 0$.

References

T.W. Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley Series in Probability and Statistics. 2003. ISBN 9780471360919.

Koenraad M.R. Audenaert. A norm compression inequality for block partitioned positive semidefinite matrices. *Linear Algebra and its Applications*, 413(1):155 – 176, 2006. ISSN 0024-3795. doi: http://dx.doi.org/10.1016/j.laa.2005.08.017.

Rasmus Benestad, Deliang Chen, Abdelkader Mezghani, Lijun Fan, and Kajsa Parding. On using principal components to represent stations in empirical-statistical downscaling. *Tellus A*, 67(0), 2015. ISSN 1600-0870.

R.E. Benestad, I. Hanssen-Bauer, and D. Chen. *Empirical-statistical Downscaling*. 2008. ISBN 9789812819123.

Peter J. Bickel and Elizaveta Levina. Regularized estimation of large covariance matrices. *Ann. Statist.*, (1):199–227, 02 2008a. doi: 10.1214/009053607000000758.

Peter J. Bickel and Elizaveta Levina. Covariance regularization by thresholding. *Ann. Statist.*, (6):2577–2604, 12 2008b. doi: 10.1214/08-AOS600.

T. Tony Cai and Ming Yuan. Adaptive covariance matrix estimation through block thresholding. *Ann. Statist.*, (4):2014–2042, 08 2012. doi: 10.1214/12-AOS999.

T. Tony Cai, Cun-Hui Zhang, and Harrison H. Zhou. Optimal rates of convergence for covariance matrix estimation. *Ann. Statist.*, (4):2118–2144, 08 2010. doi: 10.1214/09-AOS752.

T. Tony Cai, Zhao Ren, and Harrison H. Zhou. Optimal rates of convergence for estimating toeplitz covariance matrices. *Probability Theory and Related Fields*, 156(1):101–143, 2013. ISSN 1432-2064. doi: 10.1007/s00440-012-0422-7.

T. Tony Cai, Zhao Ren, and Harrison H. Zhou. Estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation. *Electron. J. Statist.*, (1):1–59, 2016. doi: 10.1214/15-EJS1081.

R. Christensen. *Plane Answers to Complex Questions: The Theory of Linear Models*. Springer texts in statistics. 2002. ISBN 9780387953618.

Noel A. C. Cressie. *Statistics for spatial data*. Wiley series in probability and mathematical statistics. J. Wiley & Sons, New York, Chichester, Toronto, 1993.

Jianqing Fan, Yuan Liao, and Han Liu. An overview of the estimation of large covariance and precision matrices. *The Econometrics Journal*, 19(1), 2016. ISSN 1368-423X.

Iain M. Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.*, (2):295–327, 04 2001. doi: 10.1214/aos/1009210544.

Cari Kaufman. Covariance tapering for likelihoodbased estimation in large spatial data sets. *Journal of the American Statistical Association*, pages 1545–1555, 2008.

Christopher King. Inequalities for trace norms of 2 by 2 block matrices. *Communications in Mathematical Physics*, 242 (3):531–545, 2003. ISSN 1432-0916. doi: 10.1007/s00220-003-0955-9.

Clifford Lam and Jianqing Fan. Sparsistency and rates of convergence in large covariance matrix estimation. *Ann. Statist.*, (6B):4254–4278, 12 2009. doi: 10.1214/09-AOS720.

J. LeSage and R.K. Pace. *Introduction to Spatial Econometrics*. Chapman-Hall, 2009.

Gabriel J. Lord, Catherine E. Powell, and Tony Shardlow. *An Introduction to Computational Stochastic PDEs:*. Cambridge University Press, 006 2014. doi: 10.1017/CBO9781139017329.

Harry Markowitz. Portfolio selection. *The Journal of Finance*, 7(1), 1952. ISSN 1540-6261.

J.M. Montero, G. Fernandez-Aviles, and J. Mateu. *Spatial and Spatio-Temporal Geostatistical Modeling and Kriging*. Wiley Series in Probability and Statistics. 2015. ISBN 9781118762431.

D. Pollard. *A User's Guide to Measure Theoretic Probability*. Cambridge Series in Statistical and Probabilistic Mathematics. 2002. ISBN 9780521002899.

Mohsen Pourahmadi. *High-dimensional Covariance Estimation: with High-Dimensional Data*. Wiley, 2014.

Pradeep Ravikumar, Martin J. Wainwright, Garvesh Raskutti, and Bin Yu. High-dimensional covariance estimation by minimizing l1-penalized log-determinant divergence. *Electron. J. Statist.*, pages 935–980, 2011. doi: 10.1214/11-EJS631.

Benjamin Shaby and David Ruppert. Tapered covariance: Bayesian estimation and asymptotics. *Journal of Computational and Graphical Statistics*, 21(2):433–452, 2012.

M. Sherman. *Spatial Statistics and Spatio-Temporal Data: Covariance Functions and Directional Properties*. Wiley Series in Probability and Statistics. 2011. ISBN 9780470974926.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 1st edition, 2008. ISBN 0387790519, 9780387790510.

A. W. van der Vaart. *Asymptotic statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998. ISBN 0-521-49603-9.

Wang, Lingxiao, Xiang Ren, and Quanquan Gu. Precision matrix estimation in high dimensional gaussian graphical models with faster rates. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Statistics*. Citeseer, 2016.

Han Xiao and Wei Biao Wu. Covariance matrix estimation for stationary time series. *Ann. Statist.*, (1):466–493, 02 2012. doi: 10.1214/11-AOS967.

Lingzhou Xue and Hui Zou. Minimax optimal estimation of general bandable covariance matrices. *Journal of Multivariate Analysis*, 116:45–51, 4 2013. ISSN 0047-259X. doi: 10.1016/j.jmva.2012.11.003.

Bin Yu. *Assouad, Fano, and Le Cam*, pages 423–435. Springer New York, New York, NY, 1997. ISBN 978-1-4612-1880-7. doi: 10.1007/978-1-4612-1880-7_29.

Zhengyuan Zhu and Yufeng Liu. Estimating spatial covariance using penalised likelihood with weighted l 1 penalty. *Journal of Nonparametric Statistics*, 21(7):925–942, 2009.