Minimax and Adaptive Estimation of Covariance Operator for Random Variables Observed on a Lattice Graph

T. Tony Cai* and Ming Yuan†
University of Pennsylvania and Georgia Institute of Technology

November 3, 2012

Abstract

Covariance structure plays an important role in high dimensional statistical inference. In a range of applications including imaging analysis and fMRI studies, random variables are observed on a lattice graph. In such a setting it is important to account for the lattice structure when estimating the covariance operator.

In this paper we consider both minimax and adaptive estimation of the covariance operator over collections of polynomially decaying and exponentially decaying parameter spaces. We first establish the minimax rates of convergence for estimating the covariance operator under the operator norm. The results show that the dimension of the lattice graph significantly affects the optimal rates convergence, often much more so than the dimension of the random variables. We then consider adaptive estimation of the covariance operator. A fully data driven block thresholding procedure is proposed and is shown to be adaptively rate optimal simultaneously over a wide range of polynomially decaying and exponentially decaying parameter spaces. The adaptive block thresholding procedure is easy to implement and numerical experiments are carried out to illustrate the merit of the procedure.

Keywords: Adaptive estimation, block thresholding, covariance matrix, covariance operator, lattice graph, minimax estimation, operator norm, optimal rate of convergence.
AMS 2000 Subject Classification: Primary 62H12; secondary 62F12, 62G09.

[1]Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104. The research of Tony Cai was supported in part by NSF FRG Grant DMS-0854973.

[2]H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332. The research of Ming Yuan was supported in part by NSF Career Award DMS-0846234.

1 Introduction

In many high-dimensional inference problems, random variables are observed on a lattice graph. For example, in imaging analysis the intensity values are observed on pixels that form a two dimensional lattice, and in fMRI studies the observations are made at voxels that can be described as a three dimensional lattice graph. In these applications, the covariance structure, which needs to be estimated from the data, often plays a critical role. For covariance estimation in such settings, it is important to account for the structural information because the covariance between two random variables often depends on where they are observed. Simply vectorizing the observations and estimating the covariance as a matrix typically does not lead to satisfactory results as the lattice structure is ignored.

Consider, for example, extracting eigenimages from a training set. This is a standard task in imaging analysis, especially for the purpose of face recognition (see, e.g., Sirovich and Kirby, 1987; Turk and Pentland, 1991). Figure 1 gives a simple illustration of the importance of accounting for the lattice structure of images. Eigenimages are the eigenvectors of the covariance operator of images. Typically eigenimages are estimated directly from the sample covariance operator which does not account for the lattice structure of an image. See, e.g., Turk and Pentland (1991). With a relatively small sample size, such an estimate may be unreliable. In this example, four hundred images of resolution 25×25 were simulated from a Markov random field model whose corresponding eigenimage is given in the left panel of Figure 1. The eigenimage estimated based on the sample covariance operator is given in the middle panel. The correlation between the truth and the sample eigenimage is 41% which indicates a rather poor estimate. As a comparison, we also applied the covariance operator estimation procedure developed in this paper which is specifically designed to account for the lattice structure. The right panel of Figure 1 provides the corresponding eigenimage. The correlation between this estimate and the true eigenimage is 81%, which represents a significant improvement over the one based on the sample covariance operator. More detailed discussion on this example is given in Section 4.

Let $\mathcal{G}(q_1, \ldots, q_d) = \{1, 2, \ldots, q_1\} \times \cdots \times \{1, 2, \ldots, q_d\}$ be a d-dimensional lattice. Assume without loss of generality that $q_1 \leq q_2 \leq \cdots \leq q_d$. Hereafter, we shall use $\mathcal{G}_d$ as a shorthand notation for the d-dimensional lattice $\mathcal{G}(q_1, \ldots, q_d)$ when no confusion occurs. Let $X = (X(t) : t \in \mathcal{G}_d)$ be a stochastic process defined on the lattice graph $\mathcal{G}_d$. Suppose we observe n independent realizations of X, denoted by $X_1, X_2, \ldots, X_n$. We are interested in estimating the covariance operator of X, $\Sigma = (\sigma(s,t))_{s,t \in \mathcal{G}_d}$ where $\sigma(s,t) = \mathrm{cov}(X(s), X(t))$, based on the random sample $\{X_1, X_2, \ldots, X_n\}$. Note that the covariance operator Σ is defined over the Cartesian product space of $\mathcal{G}_d \times \mathcal{G}_d$, i.e., $\Sigma \in \mathbb{R}^{\mathcal{G}_d \times \mathcal{G}_d}$. A particularly interesting case here is when the number of variables $p := q_1 q_2 \cdots q_d$ is moderate or large when compared with the sample size n. Estimating a covariance operator in the high-dimensional setting is difficult and it is crucial to take advantage of the special structure of the problem. In particular, it is often the case that the covariance between $X(s)$ and $X(t)$ diminishes as their distance $D(s,t)$ increases. Note that Σ corresponds to a compact operator from $\ell_2(\mathcal{G}_d)$ to

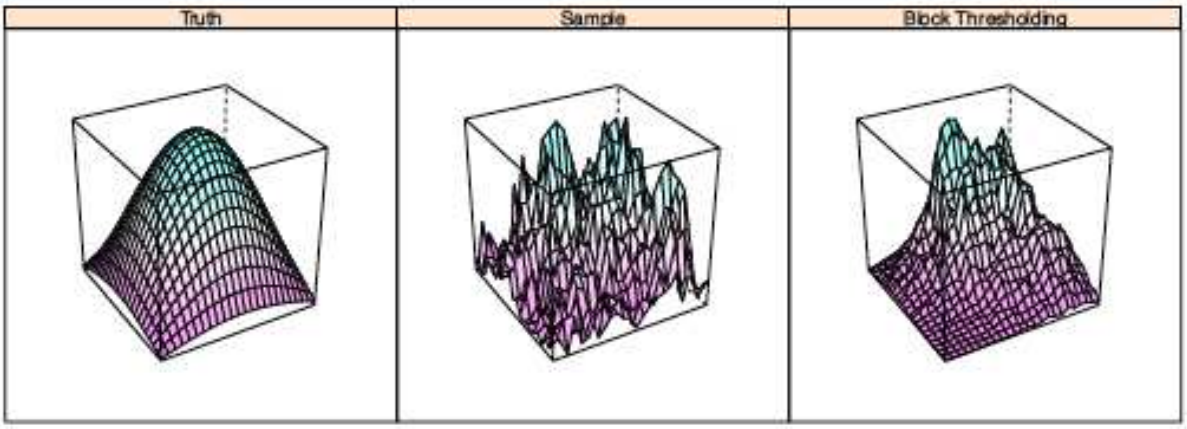


Figure 1: Importance of accounting for the lattice structure of images – the true eigenimage, along with the estimate derived from sample covariance operator, and a covariance operator that appropriately account for the lattice structure.

itself. Let $\|\Sigma\|$ be its operator norm. We shall consider the setting where the covariance operator $\Sigma \in \mathcal{F}_d(\{a_k\}; M)$ for some non-increasing sequence $a_k \downarrow 0$ and a constant $M > 0$ where

$$\mathcal{F}_d(\{a_k\}; M_0) = \left\{ \Sigma : \ \Sigma \succ 0, \ \|\Sigma\| \leq M_0, \ \sum_{s: D(s,t) \geq k} |\sigma(s,t)| \leq a_k, \quad \forall k > 0 \text{ and } t \in \mathcal{G}_d \right\}. \tag{1}$$

To fix ideas, in what follows, we shall take $D(\cdot, \cdot)$ to be the Manhattan or equivalently ℓ_1 distance on $\mathcal{G}_d$, a natural metric for lattice graph (Krause, 1987). Our development, however, can be easily generalized to deal with other distance measures on $\mathcal{G}_d$.

We study in this paper optimal and adaptive estimation of $\Sigma \in \mathcal{F}_d(\{a_k\}; M_0)$ under the operator norm $\|\cdot\|$. In particular we shall focus on two specific choices of $\{a_k : k \geq 1\}$, namely, $a_k = Mk^{-\alpha}$ and $a_k = M \exp(-\alpha_0 k^\alpha)$ for some constants $M, \alpha_0, \alpha > 0$. For brevity, in what follows, we denote by $\mathcal{F}_d(\alpha; M_0, M)$ the first class of covariance operators and $\mathcal{F}_d^*(\alpha_0, \alpha; M_0, M)$ the class that corresponds to $a_k = M \exp(-\alpha_0 k^\alpha)$. It is clear that the former describes a class of covariance operators where the covariance between two random variables decays polynomially in their distance whereas the latter consists of covariance operators where the covariances decay exponentially fast with their distances. We shall consider subgaussian variables X which satisfy, for some constant $\rho > 0$,

$$\mathbb{P}\left\{ \left| \sum_{t \in \mathcal{G}_d} u(t)(X(t) - \mathbb{E}X(t)) \right| > t \right\} \leq e^{-\rho t^2/2}, \qquad \text{for all } t > 0 \text{ and } \|u\| = 1. \tag{2}$$

Denote by $\mathcal{P}_d(\alpha; M_0, M)$ the collection of subgaussian distributions with the covariance operator $\Sigma \in \mathcal{F}_d(\alpha; M_0, M)$ and similarly, $\mathcal{P}_d^*(\alpha_0, \alpha; M_0, M)$ is the collection of subgaussian distributions with $\Sigma \in \mathcal{F}_d^*(\alpha_0, \alpha; M_0, M)$. We write $a_n \asymp b_n$ if there are constants $0 < c_1 \leq c_2$ such that $c_1 \leq a_n/b_n \leq c_2$ for all n. Combining the upper and lower bound results given in Section 2, we establish the following minimax rates of convergence for estimating Σ under the operator norm.

Theorem 1. *Let X be a random variable defined on a lattice graph $\mathcal{G}(q_1, \ldots, q_d)$ with $q_1 \le \cdots \le q_d$. Given a random sample $X_1, \ldots, X_n$ from the distribution of X. The minimax risk for estimating the covariance operator Σ under the operator norm $\|\cdot\|$ satisfies*

$$\inf_{\tilde{\Sigma}(\text{data})} \sup_{\mathcal{P}_d(\alpha; M_0, M)} \mathbb{E}\|\tilde{\Sigma} - \Sigma\|^2 \asymp \frac{\log p}{n} + \min\left\{ \left(n^{-1} \prod_{l=0}^{k} q_l \right)^{\frac{2\alpha}{2\alpha+d-k}} : 0 \le k \le d \right\}, \tag{3}$$

where $q_0 := 1$; and

$$\inf_{\tilde{\Sigma}(\text{data})} \sup_{\mathcal{P}_d^*(\alpha_0, \alpha; M_0, M)} \mathbb{E}\|\tilde{\Sigma} - \Sigma\|^2 \asymp \frac{\log p}{n} + \frac{1}{n} \prod_{k=1}^{d} (\min\{q_k, (\log n)^{1/\alpha}\}). \tag{4}$$

The minimax rates of convergence given in Theorem 1 quantify how well the covariance operators can be estimated. The optimal rates are established in two steps. We first obtain lower bounds for the minimax risk by applying Fano's lemma to a carefully constructed finite subset of the parameter spaces. A blockwise banding estimator is then proposed and is shown to attain the same rates of convergence as those of the minimax lower bounds and it is thus rate optimal.

Theorem 1 shows that the optimal rate of convergence for estimating the covariance operator depends not only on the total number p of variables but also on the individual dimensions $q_1, \ldots, q_d$ of the lattice. In the case of exponentially decaying covariance operators, the rate is determined jointly by p and those dimensions that are smaller than $(\log n)^{1/\alpha}$. The effect of dimensions on the optimal rate of convergence for polynomially decaying covariance operator is more profound. A revealing example is the case when $d = 2$. The optimal rate for estimating polynomially decaying covariance operator is given by

$$\inf_{\tilde{\Sigma}(\text{data})} \sup_{\Sigma \in \mathcal{F}_2(\alpha; M_0, M)} \mathbb{E}\|\tilde{\Sigma} - \Sigma\|^2 \asymp \frac{\log(q_1 q_2)}{n} + \min\left\{ n^{-\frac{\alpha}{\alpha+1}}, \left(\frac{q_1}{n}\right)^{\frac{2\alpha}{2\alpha+1}}, \frac{q_1 q_2}{n} \right\}. \tag{5}$$

We note an interesting phase transition behavior in the effect of the dimensionality of the lattice: the optimal rate of convergence does not depend on the specific value of q_2 whenever $q_2 \gg (n/q_1)^{1/(2\alpha+1)}$; and the rate does not depend on either q_1 or q_2 when $q_1 \gg n^{1/(2\alpha+2)}$.

It is also instructive to examine carefully the special case when $q_1 = \cdots = q_d =: q$ and hence $p = q^d$. In this case the minimax rates given in (3) and (4) can be more explicitly expressed as

$$\inf_{\tilde{\Sigma}(\text{data})} \sup_{\Sigma \in \mathcal{F}_d(\alpha; M_0, M)} \mathbb{E}\|\tilde{\Sigma} - \Sigma\|^2 \asymp \min\left\{ n^{-\frac{2\alpha}{2\alpha+d}} + \frac{d \log q}{n}, \frac{q^d}{n} \right\}, \tag{6}$$

and

$$\inf_{\tilde{\Sigma}(\text{data})} \sup_{\Sigma \in \mathcal{F}_d^*(\alpha_0, \alpha; M_0, M)} \mathbb{E}\|\tilde{\Sigma} - \Sigma\|^2 \asymp \min\left\{ \frac{(\log n)^{d/\alpha}}{n} + \frac{d \log q}{n}, \frac{q^d}{n} \right\}. \tag{7}$$

It is interesting to note the different roles played by the two measures of dimensionality d and p. Except for the case when the number p of variables is very small relative to the sample size n, the optimal rates depend on p only through its logarithm. Therefore, quality estimates can be obtained

with a relatively small sample size even if the number of variables is large. The dimension d of the lattice, on the other hand, has a much more severe impact on the optimal rate of convergence. For both classes of covariance operators, the rate of convergence quickly deteriorates when d increases, in a way reminiscent of the so-called "curse of dimensionality" often associated with the classical multivariate nonparametric regression (see, e.g., Tsybakov, 2009). As a result, a lot of more observations are needed to yield a good estimate as the dimension of the lattice increases.

In addition to the minimax optimality, we also study the problem of adaptive estimation of covariance operators for random variables observed on a lattice graph. A fully data driven block thresholding procedure is introduced in Section 3 and is shown to adaptively attain the optimal rate of convergence over $\mathcal{F}_d(\alpha; M_0, M)$ and $\mathcal{F}_d^*(\alpha_0, \alpha; M_0, M)$ simultaneously for all $\alpha_0, \alpha > 0$. The block thresholding procedure first carefully divides the sample covariance operator into blocks of varying sizes and then applies thresholding to each block depending on its size and operator norm. The idea of adaptive estimation through block thresholding can be traced back to nonparametric function estimation (see, e.g., Efromovich, 1985 and Cai, 1999), and has been recently applied to covariance matrix estimation (Cai and Yuan, 2011). The setting here is, however, more complicated due to the lattice structure.

Our work relates to a fast growing literature on estimation of sparse covariance and precision matrices. See, for example, Ledoit and Wolf (2004), Huang et al. (2006), Yuan and Lin (2007), Bickel and Levina (2008a, b), El Karoui (2008), Fan, Fan and Lv (2008), Friedman et al. (2008), Rothman et al. (2008), Lam and Fan (2009), Rothman, Levina and Zhu (2009), Yuan (2010), Cai and Liu (2011), Cai, Liu and Luo (2011), Cai and Yuan (2011), Cai, Liu and Zhou (2011), Cai and Zhou (2012), among many others. In particular, a commonly considered class of covariance matrices is the so-called bandable covariance matrices which amounts to a special case of $\mathcal{F}_d(\alpha; M_0, M)$ with $d = 1$. It can be easily deduced from (6) that the minimax rate of convergence for estimating bandable covariance matrices over $\mathcal{F}_1(\alpha; M_0, M)$ is

$$\inf_{\tilde{\Sigma}(\text{data})} \sup_{\Sigma \in \mathcal{F}_1(\alpha; M_0, M)} \mathbb{E}\|\tilde{\Sigma} - \Sigma\|^2 \asymp \min\left\{ n^{-\frac{2\alpha}{2\alpha+1}} + \frac{\log p}{n}, \frac{p}{n} \right\},$$

which was first established by Cai, Zhang and Zhou (2010). More recently, Cai and Yuan (2011) showed that a carefully devised block thresholding procedure can adaptively achieve the optimal rate of convergence over $\mathcal{F}_1(\alpha; M_0, M)$ simultaneously for all $\alpha > 0$. But unlike these earlier developments where the analysis techniques are specifically tailored for covariance matrices, our treatment here is more general and can handle not only higher dimensional lattices but also covariance operators with arbitrarily decaying rates.

The rest of the paper is organized as follows. After introducing basic notations and definitions, Section 2 establishes the minimax rates of convergence for estimating both polynomially decaying and exponentially decaying covariance operators. It is shown that a blockwise banding estimator attains the optimal rate of convergence. Section 3 considers adaptive estimation. A fully data-driven block thresholding estimator is constructed by first carefully dividing the sample covariance operator into blocks and then simultaneously estimating the entries in a block by thresholding.

This estimator is shown to attain the optimal rate of convergence adaptively over the collections of both polynomially decaying and exponentially decaying covariance operators. Section 4 considers the performance of the proposed method through numerical studies. Extensions to other related problems are discussed in Section 5. Section 6 contains the proofs of a main result and some technical lemmas.

2 Optimal Rates of Convergence

In this section, we establish the optimal rates of convergence for estimating the covariance operator Σ. We begin by introducing some basic notations and definitions. Throughout the paper, for $r \geq 1$ and $u \in \mathbb{R}^{\mathcal{G}_d}$, denote $\|u\|_r = (\sum_{t\in\mathcal{G}_d} |u(t)|^r)^{1/r}$. In the special case of $r = 2$ we denote $\|u\|$ for the usual Euclidean norm of u. For the covariance operator Σ of a random variable X defined on the lattice $\mathcal{G}_d$, we define $\|\Sigma\|_{\ell_r\to\ell_r} = \max_{\|u\|_r=1} \|\Sigma u\|_r$ for the operator norm from $\ell_r(\mathcal{G}_d)$ to $\ell_r(\mathcal{G}_d)$. When $r = 2$, we simply denote $\|\Sigma\|$ for the norm $\|\Sigma\|_{\ell_2\to\ell_2}$.

2.1 Minimax Lower Bounds

A key step in establishing the optimal rate of convergence is the derivation of the minimax lower bounds. We obtain separately the lower bounds for the collection of polynomially decaying covariance operators $\mathcal{F}_d(\alpha; M_0, M)$ and for the collection of exponential decaying covariance operators $\mathcal{F}_d^*(\alpha_0, \alpha; M_0, M)$. Note that any lower bound for a specific case yields immediately a lower bound for the general case. It therefore suffices to consider the case when X is normally distributed. The upper bounds derived in Section 2.2 show that these lower bounds are minimax rate optimal.

2.1.1 Polynomially decaying covariance operators

We have the following lower bound for the minimax risk of estimating Σ over the collection of polynomially decaying covariance operators $\mathcal{F}_d(\alpha; M_0, M)$.

Theorem 2. *Suppose that we observe a random sample $X_1, \ldots, X_n \sim_{\text{iid}} \mathcal{N}(0, \Sigma)$ and wish to estimate $\Sigma \in \mathbb{R}^{\mathcal{G}_d\times\mathcal{G}_d}$ under the operator norm $\|\cdot\|$. Then there exists a constant $C > 0$ not depending on p or n such that*

$$\inf_{\tilde{\Sigma}(\text{data})} \sup_{\Sigma\in\mathcal{F}_d(\alpha;M_0,M)} \mathbb{E}\|\tilde{\Sigma} - \Sigma\|^2 \geq C\left(\frac{\log p}{n} + \min\left\{\left(n^{-1}\prod_{l=0}^{k} q_l\right)^{\frac{2\alpha}{2\alpha+d-k}} : 0 \leq k \leq d\right\}\right), \qquad (8)$$

where $q_0 = 1$.

Proof. Recall that $q_1 \leq q_2 \leq \cdots \leq q_d$. Denote by

$$k^* = \operatorname*{argmin}_k \left\{\left(n^{-1}\prod_{l=0}^{k} q_l\right)^{\frac{2\alpha}{2\alpha+d-k}} : 0 \leq k \leq d\right\}.$$

It then suffices to show that

$$\inf_{\tilde{\Sigma}(\text{data})} \sup_{\Sigma \in \Theta_1} \mathbb{E}\|\tilde{\Sigma} - \Sigma\|^2 \geq \frac{C \log p}{n}, \tag{9}$$

and

$$\inf_{\tilde{\Sigma}(\text{data})} \sup_{\Sigma \in \Theta_2} \mathbb{E}\|\tilde{\Sigma} - \Sigma\|^2 \geq C \left(n^{-1} \prod_{l=0}^{k^*} q_l \right)^{\frac{2\alpha}{2\alpha+d-k^*}}, \tag{10}$$

for some carefully designed classes of covariance operators $\Theta_1, \Theta_2 \subset \mathcal{F}_d(\alpha; M_0, M)$.

Assume without loss of generality that $M_0 > 1$. Let Σ_0 be the identity operator, that is, $\sigma_0(s,t) = \delta_{st}$ where δ is the Kronecker's delta. Denote by

$$\Theta_1 = \{\Sigma_0\} \bigcup \left\{ \Sigma : \exists t_0 \in \mathcal{G}_d \text{ such that } \sigma(s,t) = \left\{ \begin{array}{ll} 1 + a\sqrt{n^{-1}\log p} & \text{if } s = t = t_0 \\ 1 & \text{if } s = t \neq t_0 \\ 0 & \text{otherwise} \end{array} \right\} \right.,$$

where $0 < a < 1/8$ is a small enough constant such that $\Theta_1 \subset \mathcal{F}_d(\alpha; M_0, M)$. Denote by $\mathbb{P}_\Sigma$ the joint distribution of n iid centered Gaussian processes $X_1, \ldots, X_n$ with covariance operator Σ. It is clear that for any $\Sigma \neq \Sigma_0 \in \Theta_1$, the Kullback-Leibler distance from $\mathbb{P}_\Sigma$ to $\mathbb{P}_{\Sigma_0}$ is given by

$$\mathcal{K}(\mathbb{P}_\Sigma | \mathbb{P}_{\Sigma_0}) = \frac{n}{2} \left[a\sqrt{n^{-1}\log p} - \log(1 + a\sqrt{n^{-1}\log p}) \right].$$

Note that $\log(1+x) \geq x - x^2/2$ for any $x \geq 0$. Therefore,

$$\mathcal{K}(\mathbb{P}_\Sigma | \mathbb{P}_{\Sigma_0}) \leq \frac{a^2 \log p}{4}.$$

Lower bound (9) then follows from Fano's lemma and the fact that $\|\Sigma_1 - \Sigma_2\| = a\sqrt{n^{-1}\log p}$, for any $\Sigma_1 \neq \Sigma_2 \in \Theta_1$.

To prove (10), we consider separately the cases when (a) $k^* = 0$; (b) $k^* = d$; and (c) $1 \leq k^* < d$. In each case, we appeal to the Varshamov-Gilbert bound (see, e.g., Tsybakov, 2009) to construct Θ_2. Consider first the case when $k^* = 0$. Simple calculation indicates that in this case,

$$n^{-\frac{2\alpha}{2\alpha+d}} \leq (q_1/n)^{-\frac{2\alpha}{2\alpha+d-1}},$$

which implies that $q_1 \geq n^{\frac{1}{2\alpha+d}}$.

Write $k = \lceil n^{1/(2\alpha+d)} \rceil$. Denote by $\{0,1\}^{\mathcal{G}(k,\ldots,k)}$ the collection of all functions that map from a d-dimensional lattice $\mathcal{G}(k,\ldots,k)$ to $\{0,1\}$. Then Varshamov-Gilbert bound indicates that for any k such that $k^d \geq 8$, there exist a subset $\Omega := \{\omega_1, \ldots, \omega_N\}$ of $\{0,1\}^{\mathcal{G}(k,\ldots,k)}$ obeying $N \geq 2^{k^d/8}$ and

$$\|\omega_{j'} - \omega_j\|_1 \geq k^d/8, \qquad \forall 0 \leq j \neq j' \leq N$$

where $\omega_0 = (0,\ldots,0)$. With slight abuse of notation, write $\omega_j : \mathcal{G}_d \mapsto \{0,1\}$ such that $\omega_j(s) = 0$ for any s such that $\|s\|_\infty > k$, and its restriction $\omega_j|_{\mathcal{G}(k,\ldots,k)} \in \Omega$. Denote by

$$\Sigma_j := \Sigma(\omega_j) = \delta_{st} + \left\{ \begin{array}{ll} an^{-1/2}k^{-d/2} & \text{if } \omega_j(s) = \omega_j(t) = 1 \\ 0 & \text{otherwise} \end{array} \right.,$$

where $0 < a < 1/4$ is a small enough constant such that $\Sigma_j \in \mathcal{F}_d(\alpha; M_0, M)$. It is not hard to see that for any $1 \le j \ne j' \le N$,

$$\max\left\{\|\mathbb{I}(\omega_j > \omega_{j'})\|_1, \|\mathbb{I}(\omega_j < \omega_{j'})\|_1\right\} \ge \frac{1}{2}\|\omega_{j'} - \omega_j\|_1 \ge k^d/16.$$

Thus,

$$\|\Sigma_{j'} - \Sigma_j\| \ge \max\left\{\|\Sigma(\mathbb{I}(\omega_j > \omega_{j'}))\|, \|\Sigma(\mathbb{I}(\omega_j < \omega_{j'}))\|\right\} \ge \frac{ak^{d/2}}{16n^{1/2}} \ge \frac{a}{16}n^{\alpha/(2\alpha+d)}.$$

Note that if the covariance operator of a Gaussian process X is Σ_j, then the covariance matrix of $\mathrm{vec}(X)$, the vectorized process, is given by $I_p + an^{-1/2}k^{-d/2}\mathrm{vec}(\omega_j)\mathrm{vec}(\omega_j)^\mathsf{T}$. It can then be computed that

$$\begin{aligned}
\mathcal{K}(\mathbb{P}_{\Sigma_j}|\mathbb{P}_{\Sigma_0}) &= \frac{n}{2}\Big[\mathrm{trace}(I_p + an^{-1/2}k^{-d/2}\mathrm{vec}(\omega_j)\mathrm{vec}(\omega_j)^\mathsf{T}) \\
&\qquad\qquad - \log\det(I_p + an^{-1/2}k^{-d/2}\mathrm{vec}(\omega_j)\mathrm{vec}(\omega_j)^\mathsf{T}) - p\Big] \\
&= \frac{n}{2}\left[an^{-1/2}k^{-d/2}\|\omega_j - \omega_0\|_1 - \log(1 + an^{-1/2}k^{-d/2}\|\omega_j - \omega_0\|_1)\right],
\end{aligned}$$

by the matrix determinant lemma. It follows from the fact $\log(1+x) \ge x - x^2/2$ for $x \ge 0$ that

$$\mathcal{K}(\mathbb{P}_{\Sigma_j}|\mathbb{P}_{\Sigma_0}) \le \frac{a^2}{4k^d}\|\omega_j - \omega_0\|_1^2 \le \frac{a^2k^d}{4} < \frac{\log N}{8}.$$

An application of Fano's lemma yields (10) by defining $\Theta_2 = \{\Sigma_j : 0 \le j \le N\}$.

Now consider the case $k^* = d$ where a similar argument can be used. Observe that in this case,

$$\left(n^{-1}q_1 \cdots q_{d-1}\right)^{\frac{2\alpha}{2\alpha+1}} \ge (n^{-1}q_1 \cdots q_d),$$

which, together with the fact that $q_1 \le \cdots \le q_d$, implies that $q_d \le n^{\frac{1}{2\alpha+d}}$.

Let Θ_2 be defined in a similar fashion as before except that now ω_j are defined over $\mathcal{G}_d$. More specifically let $\Omega := \{\omega_1, \ldots, \omega_N\}$ of $\{0,1\}^{\mathcal{G}_d}$ obeying $N \ge 2^{p/8}$ and

$$\|\omega_{j'} - \omega_j\|_1 \ge p/8, \qquad \forall 0 \le j \ne j' \le N,$$

which is possible thanks to another application of Varshamov-Gilbert bound. It can be calculated as before,

$$\|\Sigma_{j'} - \Sigma_j\| \ge \frac{a}{16}\sqrt{\frac{p}{n}},$$

for any $\Sigma_j \ne \Sigma_{j'} \in \Theta_2$; and

$$\mathcal{K}(\mathbb{P}_\Sigma|\mathbb{P}_{\Sigma_0}) \le \frac{\log N}{8},$$

for any $\Sigma \ne \Sigma_0 \in \Theta_2$. Fano's lemma then yields

$$\inf_{\tilde{\Sigma}(\mathrm{data})} \sup_{\Sigma \in \Theta_2} \mathbb{E}\|\tilde{\Sigma} - \Sigma\|^2 \ge \frac{Cp}{n}. \tag{11}$$

It remains to consider the case when $1 \le k^* < d$. Observe that in this case,

$$\left(n^{-1}\prod_{l=0}^{k^*} q_l\right)^{\frac{2\alpha}{2\alpha+d-k^*}} \le \left(n^{-1}\prod_{l=0}^{k^*+1} q_l\right)^{\frac{2\alpha}{2\alpha+d-k^*-1}},$$

which implies that $q_{k_*+1} \ge \left(n/\prod_{l=0}^{k^*} q_l\right)^{\frac{1}{2\alpha+d-k^*}}$.

We need to modify the construction of Θ_2. Similar to before, by Varshamov-Gilbert bound, there exists a subset $\Omega := \{\omega_1, \ldots, \omega_N\}$ of $\{0,1\}^{\mathcal{G}(q_1,\ldots,q_d)}$ such that

(a) $\omega_j(s) = 0$ for any s such that $\max\{s_{k^*+1}, \ldots, s_d\} > k$;

(b) $N \ge 2^{q_1\cdots q_{k^*}k^{d-k^*}/8}$;

(c) for any $0 \le j \ne j' \le N$, $\|\omega_{j'} - \omega_j\|_1 \ge q_1\cdots q_{k^*}k^{d-k^*}/8$, $\forall 0 \le j \ne j' \le N$, where $\omega_0 = 0$.

Take

$$k = \left\lceil \left(n/\prod_{l=0}^{k^*} q_l\right)^{\frac{1}{2\alpha+d-k^*}} \right\rceil. \tag{12}$$

Let $\Theta_2 = \{\Sigma_j : 0 \le j \le N\}$ where

$$\Sigma_j := \Sigma(\omega_j) = \delta_{st} + \begin{cases} an^{-1/2}k^{-d/2} & \text{if } \omega_j(s) = \omega_j(t) = 1 \\ 0 & \text{otherwise} \end{cases}.$$

Here $0 < a < 1/4$ is a small enough constant such that $\Sigma_j \in \mathcal{F}_d(\alpha; M_0, M)$. Then, by Fano's lemma, as before, it can be shown that

$$\inf_{\tilde{\Sigma}(\text{data})} \sup_{\Sigma\in\Theta_2} \mathbb{E}\|\tilde{\Sigma} - \Sigma\|^2 \ge C\left(n^{-1}\prod_{l=0}^{k^*} q_l\right)^{\frac{2\alpha}{2\alpha+d-k^*}}. \tag{13}$$

The lower bound (8) for estimating $\Sigma \in \mathcal{F}_d(\alpha; M_0, M)$ then follows from (9) and (10). ∎

2.1.2 Exponentially decaying covariance operators

We now turn to the exponentially decaying covariance operators. Similar to Theorem 2, we have the following lower bound for the minimax risk of estimating Σ over the collection of exponentially decaying covariance operators $\mathcal{F}_d^*(\alpha_0, \alpha; M_0, M)$.

Theorem 3. *Suppose that we observe a random sample $X_1, \ldots, X_n \sim_{\text{iid}} \mathcal{N}(0, \Sigma)$ and wish to estimate $\Sigma \in \mathbb{R}^{\mathcal{G}_d\times\mathcal{G}_d}$ under the operator norm $\|\cdot\|$. Then there exists a constant $C > 0$ not depending on p or n such that*

$$\inf_{\tilde{\Sigma}(\text{data})} \sup_{\Sigma\in\mathcal{F}_d^*(\alpha_0,\alpha;M_0,M)} \mathbb{E}\|\tilde{\Sigma} - \Sigma\|^2 \ge C\left(\frac{\log p}{n} + \frac{1}{n}\prod_{k=1}^{d}(\min\{q_k, (\log n)^{1/\alpha}\})\right). \tag{14}$$

Proof. The argument is similar to the polynomially decaying ones. Let $q_* > 1$ be the solution to

$$\log n + d \log x = 2\alpha_0 x^\alpha. \tag{15}$$

It is clear that $q_* \asymp (\log n)^{1/\alpha}$. More precisely,

$$\left(\frac{1}{2\alpha_0}\log n\right)^{1/\alpha} < q_* < \left(\left(\frac{1}{2\alpha_0} + \delta\right)\log n\right)^{1/\alpha}$$

for any $\delta > 0$. The case when $q_1 \geq q_*$ can be treated in the same fashion as the case when $k^* = 0$ for polynomially decaying covariance operators by taking $k = \lceil q_* \rceil$. Similarly the case when $q_d \leq q_*$ can be treated in the same fashion as the case when $k^* = d$; and the case when $q_1 < q_* < q_d$ can be treated in the same fashion as the case when $1 \leq k^* < d$. ∎

2.2 Upper Bounds

We now show the lower bounds given in Theorems 2 and 3 are indeed tight. Without loss of generality, we shall assume in the rest of the paper that X is centered, for the covariance operator is invariant to the mean. Recall that the sample covariance operator is given by

$$S = (S(s,t))_{s,t\in\mathcal{G}_d} := \left(\frac{1}{n}\sum_{i=1}^n X_i(s)X_i(t) - \bar{X}(s)\bar{X}(t)\right)_{s,t\in\mathcal{G}_d},$$

where $\bar{X}(s) = \frac{1}{n}\sum_{i=1}^n X_i(s)$. We first state the following result on the sample covariance operator.

Lemma 1. *Assume that $X_1, \ldots, X_n$ are independent copies of a subgaussian random process X defined over $\mathcal{G}_d$ with covariance operator Σ. Then there exists a constant $C > 0$ such that*

$$\mathbb{E}\|S - \Sigma\|^2 \leq \frac{Cp}{n}.$$

In the light of Lemma 1, the lower bound (8) for polynomially decaying covariance operators is attained by the sample covariance operator whenever $q \leq n^{1/(2\alpha+d)}$. Similarly, the sample covariance operator achieves the lower bound (14) for exponentially decaying covariance operator if $q \leq q^*$ where q_* is defined as the solution to (15). It therefore suffices to focus on the cases when $q > n^{1/(2\alpha+d)}$ for $\Sigma \in \mathcal{F}_d(\alpha; M_0, M)$; and when $q > q_*$ for $\Sigma \in \mathcal{F}_d^*(\alpha_0, \alpha; M_0, M)$. Our approach is constructive and in particular, we shall introduce a simple "blockwise banding" procedure for estimating Σ and show that it can attain the rates from Theorem 2 under these settings.

2.2.1 Blockwise Banding Estimator

We start by dividing the lattice $\mathcal{G}_d$ into blocks of size $b \times \cdots \times b$ for some b. More specifically, let $I_j^{(l)} = \{(j-1)b+1, (j-1)b+2, \ldots, jb\}$ for $j = 1, 2, \ldots, N_l - 1$ and $I_{N_l}^{(l)} = \{(N_l - 1)b + 1, \ldots, q_l\}$ where $N_l = \lceil q_l/b \rceil$ for $l = 1, \ldots, d$. Define a "block"

$$B_{\mathbf{j}} = I_{j_1}^{(1)} \times I_{j_2}^{(2)} \times \ldots \times I_{j_d}^{(d)},$$

for $\mathbf{j} = (j_1, j_2, \ldots, j_d) \in \mathcal{G}(N_1, \ldots, N_d)$. For a linear operator $A : \ell_2(\mathcal{G}_d) \mapsto \ell_2(\mathcal{G}_d)$, we shall define

$$A_{\mathbf{jj'}} := A_{B_\mathbf{j} \times B_\mathbf{j'}} = (a(s,t))_{s \in B_\mathbf{j}, t \in B_\mathbf{j'}}.$$

We then proceed to estimate all blocks $\Sigma_{\mathbf{jj'}}$ where $\mathbf{j}, \mathbf{j'} \in \mathcal{G}(N_1, \ldots, N_d)$ based upon their sample version. In particular, let

$$\hat{\Sigma}_{\mathbf{jj'}} = \begin{cases} S_{\mathbf{jj'}} & \text{if } \|\mathbf{j} - \mathbf{j'}\|_\infty \leq 1 \\ 0 & \text{otherwise} \end{cases}. \tag{16}$$

In other words, we estimate $\Sigma_{\mathbf{jj'}}$ by its sample counterpart if and only if the two blocks $B_\mathbf{j}$ and $B_\mathbf{j'}$ are "neighbors", as illustrated in Figure 2.

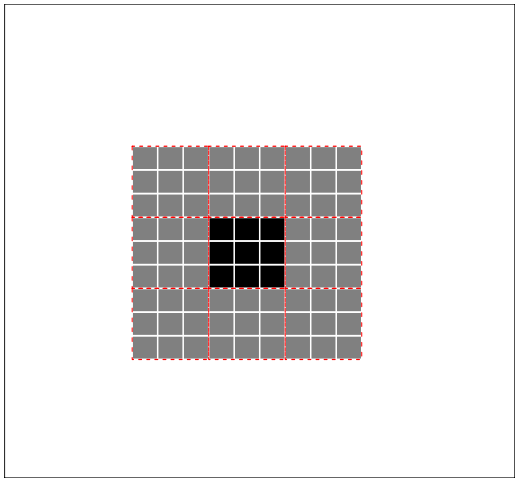


Figure 2: Blocks and their "neighbors" – A two dimensional example of the blocking scheme. In this case, $k = 3$ and the blocks are represented with red dashed lines as boundary. The grey blocks are the "neighbors" of the solid black block.

2.2.2 Polynomially decaying covariance operators

We now show that with appropriate choice of k, the proposed estimator $\hat{\Sigma}$ can achieve the optimal rate of convergence. Consider first the case when $\Sigma \in \mathcal{F}_d(\alpha; M_0, M)$. Recall that

$$k^* = \operatorname*{argmin}_k \left\{ \left(n^{-1} \prod_{l=0}^{k} q_l \right)^{\frac{2\alpha}{2\alpha+d-k}} : 0 \leq k \leq d \right\}.$$

Define the size of the block by

$$b = \left\lceil \left(n / \prod_{l=0}^{k^*} q_l \right)^{\frac{1}{2\alpha+d-k^*}} \right\rceil.$$

Theorem 4. *Suppose that we observe a random sample $X_1, \ldots, X_n$ consisting of independent copies of a subgaussian random process X defined over $\mathcal{G}_d$ and wish to estimate its covariance operator*

$\Sigma \in \mathbb{R}^{\mathcal{G}_d \times \mathcal{G}_d}$ *with* $q > n^{1/(2\alpha+d)}$. *Let* $\hat{\Sigma}$ *be the blockwise banding estimate defined as above. Then there exists a constant* $C > 0$ *not depending on* p *or* n *such that*

$$\sup_{\Sigma \in \mathcal{F}_d(\alpha; M_0, M)} \mathbb{E}\|\hat{\Sigma} - \Sigma\|^2 \leq C \left(\frac{\log p}{n} + \min \left\{ \left(n^{-1} \prod_{l=0}^{k} q_l \right)^{\frac{2\alpha}{2\alpha+d-k}} : 0 \leq k \leq d \right\} \right). \tag{17}$$

Proof. Define $\Sigma_1 = (\sigma_1(s,t))_{s,t \in \mathcal{G}_d}$ such that $\sigma_1(s,t) = \sigma(s,t)$ if $s \in B_{\mathbf{j}}$, $t \in B_{\mathbf{j}'}$ and $\|\mathbf{j} - \mathbf{j}'\|_\infty \leq 1$, and 0 otherwise. Let $\Sigma_2 = \Sigma - \Sigma_1$. Then

$$\|\hat{\Sigma} - \Sigma\| \leq \|\hat{\Sigma} - \Sigma_1\| + \|\Sigma_2\|.$$

It is easy to see that

$$\|\Sigma_2\| \leq \|\Sigma_2\|_{\ell_1 \to \ell_1} \leq \max_{s \in \mathcal{G}_d} \sum_{t: D(s,t) \geq b} |\sigma(s,t)| \leq M \left(n^{-1} \prod_{l=0}^{k^*} q_l \right)^{\frac{\alpha}{2\alpha+d-k^*}}.$$

To bound $\|\hat{\Sigma} - \Sigma_1\|$, note that

$$\|\hat{\Sigma} - \Sigma_1\| = \sup_{u \in \ell_2(\mathcal{G}_d): \|u\|=1} \left| \langle u, (\hat{\Sigma} - \Sigma_1) u \rangle \right|.$$

For any $u \in \ell_2(\mathcal{G}_d)$ with $\|u\| = 1$,

$$\begin{aligned}
\left| \langle u, (\hat{\Sigma} - \Sigma_1) u \rangle \right| &\leq \sum_{\|\mathbf{j}-\mathbf{j}'\|_\infty \leq 1} \left| \left\langle u_{B_{\mathbf{j}}}, \left(S_{\mathbf{j}\mathbf{j}'} - \Sigma_{\mathbf{j}\mathbf{j}'} \right) u_{B_{\mathbf{j}'}} \right\rangle \right| \\
&\leq \sum_{\|\mathbf{j}-\mathbf{j}'\|_\infty \leq 1} \|u_{B_{\mathbf{j}}}\| \|u_{B_{\mathbf{j}'}}\| \|S_{\mathbf{j}\mathbf{j}'} - \Sigma_{\mathbf{j}\mathbf{j}'}\| \\
&\leq \left(\sum_{\|\mathbf{j}-\mathbf{j}'\|_\infty \leq 1} \|u_{B_{\mathbf{j}}}\| \|u_{B_{\mathbf{j}'}}\| \right) \times \left(\max_{\|\mathbf{j}-\mathbf{j}'\|_\infty \leq 1} \|S_{\mathbf{j}\mathbf{j}'} - \Sigma_{\mathbf{j}\mathbf{j}'}\| \right)
\end{aligned}$$

where for any $a \in \ell_2(\mathcal{G}_d)$, $a_B = (a(t))_{t \in B}$. The Cauchy-Schwartz Inequality yields

$$\sum_{\|\mathbf{j}-\mathbf{j}'\|_\infty \leq 1} \|u_{B_{\mathbf{j}}}\| \|u_{B_{\mathbf{j}'}}\| \leq \frac{1}{2} \sum_{\|\mathbf{j}-\mathbf{j}'\|_\infty \leq 1} \left(\|u_{B_{\mathbf{j}}}\|^2 + \|u_{B_{\mathbf{j}'}}\|^2 \right) \leq 3^d \sum_{\mathbf{j} \in \mathcal{G}(N_1, \ldots, N_d)} \|u_{B_{\mathbf{j}}}\|^2 = 3^d.$$

Therefore $\|\hat{\Sigma} - \Sigma_1\| \leq 3^d \max_{\|\mathbf{j}-\mathbf{j}'\|_\infty \leq 1} \|S_{\mathbf{j}\mathbf{j}'} - \Sigma_{\mathbf{j}\mathbf{j}'}\|$ and hence

$$\|\hat{\Sigma} - \Sigma\| \leq \|\hat{\Sigma} - \Sigma_1\| + \|\Sigma_2\| \leq 3^d \max_{\|\mathbf{j}-\mathbf{j}'\|_\infty \leq 1} \|S_{\mathbf{j}\mathbf{j}'} - \Sigma_{\mathbf{j}\mathbf{j}'}\| + M \left(n^{-1} \prod_{l=0}^{k^*} q_l \right)^{\frac{\alpha}{2\alpha+d-k^*}}.$$

Consequently

$$\mathbb{E}\|\hat{\Sigma} - \Sigma\|^2 \leq 2 \cdot 3^{2d} \mathbb{E} \left(\max_{\|\mathbf{j}-\mathbf{j}'\|_\infty \leq 1} \|S_{\mathbf{j}\mathbf{j}'} - \Sigma_{\mathbf{j}\mathbf{j}'}\| \right)^2 + 2M^2 \left(n^{-1} \prod_{l=0}^{k^*} q_l \right)^{\frac{2\alpha}{2\alpha+d-k^*}}. \tag{18}$$

It remains to bound the expectation on the right hand side. We shall make use of the following result:

Lemma 2. *Let $I, J \subseteq \mathcal{G}_d$ with $\mathrm{card}(I), \mathrm{card}(J) \le s$, then there exist constants $c_1, c_2 > 0$ such that*

$$\mathbb{P}\left\{\|S_{I\times J} - \Sigma_{I\times J}\| \ge x\right\} \le c_1 25^s \exp(-c_2 n x^2).$$

Recall that when $k^* = 0$,

$$n^{-\frac{2\alpha}{2\alpha+d}} \le (q_1/n)^{-\frac{2\alpha}{2\alpha+d-1}},$$

and as a result $q_1 \ge b = \lceil n^{1/(2\alpha+d)} \rceil$. Then

$$N_1 \cdots N_d \le C q_1 \cdots q_d / b^d \le C p n^{-d/(2\alpha+d)}$$

for some constant $C > 0$. An application of Lemma 2 and union bound yields

$$\mathbb{P}\left\{\max_{\|\mathbf{j}-\mathbf{j}'\|_\infty \le 1} \|S_{\mathbf{jj}'} - \Sigma_{\mathbf{jj}'}\| \ge x\right\} \le c_1 C 3^d p n^{-d/(2\alpha+d)} 25^{b^d} \exp\left(-c_2 n x^2\right),$$

which implies that for any $x > 0$

$$\begin{aligned}
&\mathbb{E}\left(\max_{\|\mathbf{j}-\mathbf{j}'\|_\infty \le 1} \|S_{\mathbf{jj}'} - \Sigma_{\mathbf{jj}'}\|\right)^2 \\
\le\ & x^2 \mathbb{P}\left\{\max_{\|\mathbf{j}-\mathbf{j}'\|_\infty \le 1} \|S_{\mathbf{jj}'} - \Sigma_{\mathbf{jj}'}\| < x\right\} + \int_{x^2}^\infty \mathbb{P}\left\{\max_{\|\mathbf{j}-\mathbf{j}'\|_\infty \le 1} \|S_{\mathbf{jj}'} - \Sigma_{\mathbf{jj}'}\| \ge u\right\} du \\
\le\ & x^2 + c_1 C 3^d p n^{-d/(2\alpha+d)} 25^{b^d} \int_{x^2}^\infty \exp(-c_2 n u) du \\
\le\ & x^2 + c_1 C 3^d p n^{-d/(2\alpha+d)} 25^{b^d} (c_2 n)^{-1} \exp(-c_2 n x^2).
\end{aligned}$$

If $\log p \le n^{d/(2\alpha+d)}$, we take $x = c n^{-\alpha/(2\alpha+d)}$ for a sufficiently large constant $c > 0$ which yields

$$\mathbb{E}\left(\max_{\|\mathbf{j}-\mathbf{j}'\|_\infty \le 1} \|S_{\mathbf{jj}'} - \Sigma_{\mathbf{jj}'}\|\right)^2 \le C n^{-\frac{2\alpha}{2\alpha+d}}$$

for some constant $C > 0$. When $\log p > n^{d/(2\alpha+d)}$, it follows by taking $x = c\sqrt{\frac{\log p}{n}}$ for a sufficiently large constant $c > 0$ that

$$\mathbb{E}\left(\max_{\|\mathbf{j}-\mathbf{j}'\|_\infty \le 1} \|S_{\mathbf{jj}'} - \Sigma_{\mathbf{jj}'}\|\right)^2 \le C \frac{\log p}{n},$$

for some constant $C > 0$. These two bounds together with (18) implies (17) in this case.

When $k^* = d$, simple algebraic manipulation shows that

$$q_d \le n^{1/(2\alpha+d)} \le b.$$

Therefore, $N_1 = \cdots = N_d = 1$. By Lemma 2 and union bound, we get

$$\mathbb{P}\left\{\max_{\|\mathbf{j}-\mathbf{j}'\|_\infty \le 1} \|S_{\mathbf{jj}'} - \Sigma_{\mathbf{jj}'}\| \ge x\right\} \le c_1 C 3^d 25^p \exp\left(-c_2 n x^2\right),$$

which, following the same calculations as before, implies that for any $x > 0$,

$$\mathbb{E}\left(\max_{\|\mathbf{j}-\mathbf{j}'\|_\infty \le 1} \|S_{\mathbf{jj}'} - \Sigma_{\mathbf{jj}'}\|\right)^2 \le x^2 + c_1 3^d 25^p (c_2 n)^{-1} \exp(-c_2 n x^2).$$

The claim (17) follows by taking $x = cp/n$ for a large enough constant $c > 0$.

Finally when $1 \le k^* < d$, it can be shown that

$$q_{k^*} \le b \le q_{k_*+1}.$$

Therefore, $N_1 = \cdots = N_{k^*} = 1$. By Lemma 2 and union bound, we now get

$$\mathbb{P}\left\{\max_{\|\mathbf{j}-\mathbf{j}'\|_\infty \le 1} \|S_{\mathbf{jj}'} - \Sigma_{\mathbf{jj}'}\| \ge x\right\} \le c_1 C 3^d N_{k^*+1} \cdots N_d 25^{q_1 \cdots q_{k^*} b^{d-k^*}} \exp\left(-c_2 n x^2\right).$$

The desired result then follows from the same calculations as before. ∎

2.2.3 Exponentially decaying covariance operators

We turn to estimation of exponentially decaying covariance operators in $\mathcal{F}_d^*(\alpha_0, \alpha; M_0, M)$. We shall take $b = \lceil q_* \rceil$ to be the block size where q_* is the solution to (15). Similar to Theorem 4, we have the following upper bound.

Theorem 5. *Suppose we observe a random sample $X_1, \ldots, X_n$ consisting of independent copies of a subgaussian random process X defined over the lattice graph $\mathcal{G}_d$ and wish to estimate its covariance operator $\Sigma \in \mathbb{R}^{\mathcal{G}_d \times \mathcal{G}_d}$. Let $\hat{\Sigma}$ be the blockwise banding estimate defined as above. Then there exists a constant $C > 0$ not depending on p or n such that*

$$\sup_{\Sigma \in \mathcal{F}_d^*(\alpha_0, \alpha; M_0, M)} \mathbb{E}\|\hat{\Sigma} - \Sigma\|^2 \le C\left(\frac{\log p}{n} + \frac{1}{n}\prod_{k=1}^{d}(\min\{q_k, (\log n)^{1/\alpha}\})\right). \tag{19}$$

Together with the lower bound given in Theorem 3, this shows that the optimal rate of convergence for estimating $\Sigma \in \mathcal{F}_d^*(\alpha_0, \alpha; M_0, M)$ is $\frac{\log p}{n} + \frac{1}{n}\prod_{k=1}^{d}(\min\{q_k, (\log n)^{1/\alpha}\})$ and the blockwise banding estimator is rate optimal. The proof of Theorem 5 is identical to that of Theorem 4 by taking $b = \lceil q_* \rceil$, and is therefore omitted for brevity.

Although the blockwise banding estimator proposed here is capable of achieving the optimal rate of convergence, it is evident from its construction that doing so requires explicit knowledge of α which is typically unknown in practice. This makes the concept of adaptive estimation – a single estimator, not depending on the decay rate α, that achieves the optimal rate of convergence simultaneously – of great practical importance. In the next section, we shall introduce a fully data-driven adaptive estimator $\hat{\Sigma}$ and show that it is simultaneously rate optimal over the collection of the parameter spaces $\mathcal{F}_d(\alpha; M_0, M)$ and $\mathcal{F}_d^*(\alpha_0, \alpha; M_0, M)$ for all $\alpha > 0$.

3 Adaptive Estimation

The blockwise banding estimator constructed in Section 2.2 has been shown to attain the optimal rate of convergence. However, the estimator depends on the decaying rate α, which is typically unknown in practice, and the estimator is thus not adaptive. In this section we consider adaptive

estimation and construct an adaptive estimation procedure which does not require the knowledge of α, M_0 or M. The estimator will be shown to attain the optimal rate of convergence over $\mathcal{F}_d(\alpha; M_0, M)$ and $\mathcal{F}_d^*(\alpha_0, \alpha; M_0, M)$ simultaneously for all $\alpha_0, \alpha, M_0, M > 0$.

The main idea in our construction is block thresholding. We first carefully divide the sample covariance operator into blocks of varying sizes and then apply thresholding to each block depending on its size and operator norm. The idea of adaptive estimation through block thresholding can be traced back to nonparametric function estimation (see, e.g., Efromovich, 1985 and Cai, 1999), and has been recently applied to covariance matrix estimation (Cai and Yuan, 2011). To fix ideas, we first treat hypercubic lattices and follow with discussions on how to accommodate the more general hyperrectangular lattices.

We shall adopt the following notation. Let $B_1, B_2, \ldots, B_k \subseteq \mathcal{G}_2$, write

$$B_1 \odot B_2 \odot \cdots \odot B_k = \{((i_1, \ldots, i_k), (j_1, \ldots, j_k)) \in \mathcal{G}_k \times \mathcal{G}_k : (i_1, j_1) \in B_1, \ldots, (i_k, j_k) \in B_k\}$$

and

$$B_1^{\odot k} = \underbrace{B_1 \odot B_1 \odot \cdots \odot B_1}_{k \text{ times}}.$$

In addition, for two collections, $\mathcal{B}$ and $\mathcal{B}'$, of subsets from $\mathcal{G}_d$, we shall write

$$\mathcal{B} \odot \mathcal{B}' = \{B \odot B' : B \in \mathcal{B}, B' \in \mathcal{B}'\}.$$

3.1 Block Thresholding Estimator

Recall that Σ is defined over $\mathcal{G}_d \times \mathcal{G}_d$. A main challenge in adopting the strategy for our purpose is to fill the domain $\mathcal{G}_d \times \mathcal{G}_d$ by blocks of different sizes depending on the distance between the coordinates. The task becomes especially hard for $d > 1$ when it is no longer possible to visualize the blocking scheme.

To gain insights, let us first review the scheme developed by Cai and Yuan (2011) for covariance matrices which corresponds to the case $d = 1$. Note that a covariance matrix is defined over the Cartesian product space of $\{1, \ldots, q\} \times \{1, \ldots, q\}$. The construction begins by dividing the two dimensional lattice into blocks of size $s_0 \times s_0$ with $s_0 \sim \log q$. The choice of s_0 for our purpose will become clear later. The blocks are then consolidated systematically to yield a blocking of $\{1, \ldots, q\} \times \{1, \ldots, q\}$ as shown in Figure 3. Interested readers are referred to Cai and Yuan (2011) for details. Here we shall point out several key properties of the blocking. Denote by $\mathcal{B}_1 = \{B_1, B_2, \ldots\}$ the blocks constructed for $\{1, \ldots, q\} \times \{1, \ldots, q\}$, i.e.,

$$B_i \cap B_j = \emptyset, \qquad \text{and} \qquad \cup_{B \in \mathcal{B}_1} B = \{1, \ldots, q\} \times \{1, \ldots, q\}.$$

For a block $B \in \mathcal{B}_1$, there exist $I, J \in \{1, \ldots q\}$ such that $B = I \times J$. We refer to the maximum of the cardinality of I and J as the size of B, denoted by $s(B)$. It is clear that $s(B) = 2^{l-1} s_0$ for some $l \geq 1$. Denote by $\mathcal{B}_1(l)$ the subset of $\mathcal{B}_1$ consisting of all blocks of size $2^{l-1} s_0$. In particular, $\mathcal{B}_1(1)$ and $\mathcal{B}_2(2)$ are shown in Figure 3 as the solid back squares and the grey squares with red boundary

respectively. One of the most important property of this construction is so that blocks of large size are necessarily far away from the diagonal. More specifically, if $(i,j) \in B$ and $s(B) > 2s_0$, then $|i-j| \geq s(B)$.

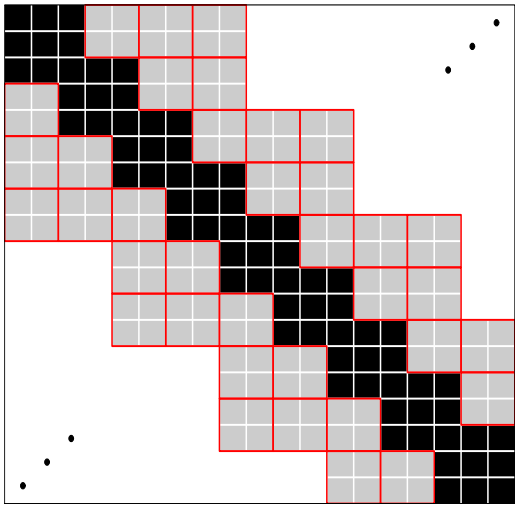


Figure 3: Blocking scheme for covariance matrices – Blocks are with increasing sizes away from the diagonal. The solid black blocks are of size $s_0 \times s_0$. The gray ones are of size $2s_0 \times 2s_0$.

Now consider dividing $\mathcal{G}_d \times \mathcal{G}_d$, the space suitable for covariance operator Σ, into blocks. It is tempting to simply use a product space $\mathcal{B}_1^{\odot d}$. Recall that

$$\mathcal{B}_1^{\odot d} = \{B := B_1 \odot \cdots \odot B_d : B_1, \ldots, B_d \in \mathcal{B}_1\}.$$

In other words, for any $\mathbf{i} = (i_1, \ldots, i_d), \mathbf{i}' = (i_1', \ldots, i_d') \in \mathcal{G}_d$, $(\mathbf{i}, \mathbf{i}') \in B$ if and only if $(i_l, i_l') \in B_l$ for $l = 1, \ldots, d$. Unfortunately, it turns out that there are too many blocks in $\mathcal{B}_1^{\odot d}$ and we need to consolidate these blocks.

For a block $B = B_1 \odot \cdots \odot B_d \in \mathcal{B}_1^{\odot d}$, write $s(B) = \max\{s(B_1), \ldots, s(B_d)\}$. Denote by $\mathcal{A}(l)$ the collection of blocks B from $\mathcal{B}_1^{\odot d}$ such that $s(B) = 2^{l-1}s_0$. It is clear that $\mathcal{B}_1^{\odot d} = \cup_{l \geq 1} \mathcal{A}(l)$. Our strategy is to consolidate the blocks within $\mathcal{A}(l)$ for a given $l > 0$.

To fix ideas, consider first the case of $d = 2$. Note that $\mathcal{A}(l) = \mathcal{A}_1(l) \cup \mathcal{A}_2(l) \cup \mathcal{A}_{12}(l)$ where $\mathcal{A}_{12}(l) = \mathcal{B}_1(l) \odot \mathcal{B}_1(l)$, and

$$\mathcal{A}_1(l) = \mathcal{B}_1(l) \odot \bar{\mathcal{B}}_1(l), \qquad \text{and} \qquad \mathcal{A}_2(l) = \bar{\mathcal{B}}_1(l) \odot \mathcal{B}_1(l).$$

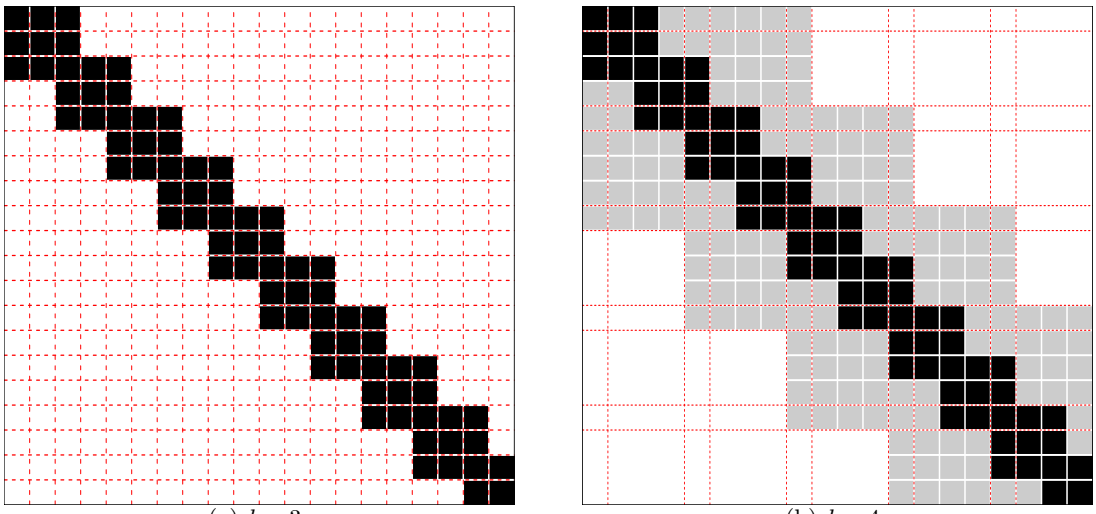
where $\bar{\mathcal{B}}_1(l) = \cup_{l' < l} \mathcal{B}_1(l')$ is the collection of blocks in $\mathcal{B}_1$ with size smaller than $2^{l-1}s_0$. To reduce the number of blocks in $\mathcal{A}_1(l)$ and $\mathcal{A}_2(l)$, a new blocking scheme is needed for the area covered by them. Due to symmetry, we consider only $\mathcal{A}_1(l)$. The main idea is to reconfigure the area from $\{1, \ldots, q\} \times \{1, \ldots, q\}$ covered by $\bar{\mathcal{B}}_1(l)$, denoted by

$$\mathcal{C}(l) = \{(i,j) \in \mathcal{G}_2 : (i,j) \in B \text{ for some } B \in \bar{\mathcal{B}}_1(l)\}.$$

To this end, consider a regular blocking at $\{((k-1)2^{l-1}+1)s_0, (k2^{l-1}-2)s_0 : k = 1, 2, \ldots\}$, i.e., blocks of one of the following four configurations:

$$\{((k-1)2^{l-1}+1)s_0, \ldots, (k2^{l-1}-3)s_0\} \times \{((k'-1)2^{l-1}+1)s_0, \ldots, (k'2^{l-1}-3)s_0\};$$
$$\{(k2^{l-1}-2)s_0, (k2^{l-1}-1)s_0, k2^{l-1}s_0\} \times \{((k'-1)2^{l-1}+1)s_0, \ldots, (k'2^{l-1}-3)s_0\};$$
$$\{((k-1)2^{l-1}+1)s_0, \ldots, (k2^{l-1}-3)s_0\} \times \{(k'2^{l-1}-2)s_0, (k'2^{l-1}-1)s_0, k'2^{l-1}s_0\};$$
$$\{(k2^{l-1}-2)s_0, (k2^{l-1}-1)s_0, k2^{l-1}s_0\} \times \{(k'2^{l-1}-2)s_0, (k'2^{l-1}-1)s_0, k'2^{l-1}s_0\}.$$

It is clear that the first three types of blocks are of size $(2^{l-1}-3)s_0$ whereas the fourth type is of size $3s_0$. The only exception occurs when $l < 3$ where the blocks are of smaller sizes. We shall neglect such a caveat in what follows for brevity. The reconfiguration for $l = 3$ and 4 are given in Figure 4 for illustration.



(a) $l = 3$ (b) $l = 4$

Figure 4: Reconfiguration of blocks of size smaller than $2^{l-1}s_0$: the left panel corresponds to the case when $l = 3$, no reconfiguration is necessary; the right panel represents the case when $l = 4$. Original blocks of size s_0 are represented as black whereas the area covered by blocks of original size $2s_0$ is in grey. Dashed lines show the reconfigured blocks.

It is clear that a subset of these blocks, denote by $\tilde{\mathcal{B}}_1(l)$, can cover $\mathcal{C}(l)$, i.e., $\bigcup_{B\in\tilde{\mathcal{B}}_1(l)} B = \mathcal{C}(l)$. The advantage of new blocking scheme $\tilde{\mathcal{B}}_1(l)$ over $\bar{\mathcal{B}}_1(l)$ is in the reduced number of blocks. For example, an inspection of the case of $l = 4$ as show in Panel (b) of Figure 4 suggests that the maximum number of blocks needed on a particular row or column is 11 for $\bar{\mathcal{B}}_1(l)$ and 9 for $\tilde{\mathcal{B}}_1(l)$. Such a reduction may not be striking. But in general, the number of blocks needed on a particular row or column increases linear in l for $\bar{\mathcal{B}}_1(l)$ and remains bounded for $\tilde{(\mathcal{B})}_1(l)$, a fact that will prove to be the key to achieve adaptation. We shall then define $\mathcal{A}'_1(l) = \mathcal{B}_1(l) \odot \tilde{\mathcal{B}}_1(l)$, $\mathcal{A}'_2(l) = \tilde{\mathcal{B}}_1(l) \odot \mathcal{B}_1(l)$, and $\mathcal{A}'(l) = \mathcal{A}'_1(l) \cup \mathcal{A}'_2(l) \cup \mathcal{A}_{12}(l)$. Let $\mathcal{B}_2 := \bigcup_{l>0} \mathcal{A}'(l)$. It is clear that $\mathcal{B}_2$ is a valid blocking of $\mathcal{G}_2 \times \mathcal{G}_2$ in that for any $B \neq B' \in \mathcal{B}_2$, $B \cap B' = \emptyset$, and $\cup_{B\in\mathcal{B}_2} B = \mathcal{G}_2 \times \mathcal{G}_2$.

The construction can be generalized to $d > 2$. With slight abuse of notation, define

$$\mathcal{A}(l) = (\cup_{1\le k\le d}\mathcal{A}_k(l))\bigcup(\cup_{1\le k_1<k_2\le d}\mathcal{A}_{k_1k_2}(l))\bigcup\cdots\bigcup\mathcal{A}_{12\ldots d}(l),$$

where

$$\begin{aligned}
\mathcal{A}_k(l) &= \bar{\mathcal{B}}_1(l)^{\odot(k-1)}\odot\mathcal{B}_1(l)\odot\bar{\mathcal{B}}_1(l)^{\odot(d-k)},\\
\mathcal{A}_{k_1k_2}(l) &= \bar{\mathcal{B}}_1(l)^{\odot(k_1-1)}\odot\mathcal{B}_1(l)\odot\bar{\mathcal{B}}_1(l)^{\odot(k_2-k_1-1)}\odot\mathcal{B}_1(l)\odot\bar{\mathcal{B}}_1(l)^{\odot(d-k_2)},\\
&\cdots\cdots\cdots\\
A_{12\cdots d}(l) &= \mathcal{B}_1(l)^{\odot d}.
\end{aligned}$$

We then replace $\mathcal{A}_k(l)$ with

$$\mathcal{A}'_k(l) = \left(\tilde{\mathcal{B}}_1(l)\right)^{\odot(k-1)}\odot\mathcal{B}_1(l)\odot\left(\tilde{\mathcal{B}}_1(l)\right)^{\odot(d-k)},$$

and $\mathcal{A}_{k_1k_2}(l)$ with

$$\mathcal{A}'_{k_1k_2}(l) = \left(\tilde{\mathcal{B}}_1(l)\right)^{\odot(k_1-1)}\odot\mathcal{B}_1(l)\odot\left(\tilde{\mathcal{B}}_1(l)\right)^{\odot(k_2-k_1-1)}\odot\mathcal{B}_1(l)\odot\left(\tilde{\mathcal{B}}_1(l)\right)^{\odot(d-k_2)},$$

and etc., leading to a blocking scheme for $\mathcal{G}_d\times\mathcal{G}_d$:

$$\mathcal{B}_d := \bigcup_{l>0}\left((\cup_{1\le k\le d}\mathcal{A}'_k(l))\bigcup(\cup_{1\le k_1<k_2\le d}\mathcal{A}'_{k_1k_2}(l))\bigcup\cdots\bigcup\mathcal{A}_{12\ldots d}(l)\right).$$

Once the blocking is defined, we then proceed to estimate the covariance operator Σ block by block. It is clear that for any $B\in\mathcal{B}_d$, there exist $I = I_1\times\cdots\times I_d, J = J_1\times\cdots\times J_d$ such that $I_1,\ldots,I_d,J_1,\ldots,J_d\subset\{1,\ldots,q\}$ and $B = I\times J$. Write $\Sigma_B = (\sigma(s,t))_{(s,t)\in B}$ for a block B, and let S_B be defined similarly. If B is a diagonal block, i.e., $I_l = J_l$ for $l = 1,\ldots,d$, we shall estimate Σ_B by its sample counterpart. If B is large in that $s^d(B) > n/\log n$, we estimate Σ_B simply by zero. For other blocks, we estimate Σ_B by S_B if

$$\|S_B\|/(\|S_{I\times I}\|\|S_{J\times J}\|)^{1/2}\ge\lambda_0 n^{-1/2}(s^d(B)+\log p)^{1/2},$$

and 0 otherwise where $\lambda_0 > 0$ is a turning parameter. Similar to the covariance matrix case, our theoretical development indicates that the resulting block thresholding estimator is optimally rate adaptive whenever λ_0 is a sufficiently large constant. In particular, it can be taken as fixed at $\lambda_0 = 6$ when X follows a multivariate normal distribution. In practice, a data-driven choice of λ_0 could potentially lead to further improved finite sample performance.

3.2 Adaptivity

It is clear from the construction, the proposed block thresholding estimator $\hat{\Sigma}$ does not rely on the knowledge of any particular parameter space. The following theorem shows that it simultaneously achieves the optimal rate of convergence over $\mathcal{F}(\alpha;M_0,M)$ and $\mathcal{F}^*(\alpha_0,\alpha;M_0,M)$ for all $\alpha_0,\alpha,M_0,M > 0$.

Theorem 6. *Let $\hat{\Sigma}$ be the block thresholding estimate defined above with $s_0 = \lceil (\log p)^{1/d} \rceil$. Then there exists a constant $C > 0$ such that*

$$\sup_{\mathcal{P}(\alpha; M_0, M)} \mathbb{E}\|\hat{\Sigma} - \Sigma\|^2 \leq C \min\left\{ n^{-\frac{2\alpha}{2\alpha+d}} + \frac{\log p}{n}, \frac{p}{n} \right\}, \tag{20}$$

and

$$\sup_{\mathcal{P}^*(\alpha_0, \alpha; M_0, M)} \mathbb{E}\|\hat{\Sigma} - \Sigma\|^2 \leq C \min\left\{ \frac{(\log n)^{d/\alpha}}{n} + \frac{\log p}{n}, \frac{p}{n} \right\} \tag{21}$$

over all $\alpha > 0$.

The main idea behind the proof of Theorem 6 is to consider separately "small" blocks and "large" blocks. Interestingly, for our purposes, whether a block is large or small is not determined by its volume, but by its maximum side length. We call a block B large if $s(B) > 2^{L-1} s_0$ where L is a natural number to be determined later. The detailed proof of Theorem 6 is postponed to Section 6.

3.3 Hyperrectangular Lattices

The block thresholding procedure introduced above can be applied to hyperrectangular lattices to achieve adaptivity as well. Recall that we start by dividing the lattice into blocks of size $(\log p)^{1/d} \times \cdots \times (\log p)^{1/d}$. When dealing with hyperrectangular blocks, we only need to avoid the case when some of the dimension qs that are smaller than $s_0 = (\log p)^{1/d}$. More specifically, if $q_1 \leq \cdots \leq q_{k^*} < s_0 \leq q_{k^*+1} \leq \cdots \leq q_d$, we shall begin by dividing the lattice $\{1, \ldots, q_1\} \times \cdots \times \{1, \ldots, q_d\}$ into blocks of size $q_1 \times \cdots \times q_{k^*} \times s_1 \times \cdots \times s_1$ where

$$s_1 = \left((\log p) / \prod_{l=1}^{k^*} q_l \right)^{1/(d-k^*)}.$$

We then follow a similar procedure to construct the final blocking scheme:

$$\mathcal{B}_d^* := \{\{1, \ldots, q_1\} \times \{1, \ldots, q_1\}\} \odot \cdots \odot \{\{1, \ldots, q_{k^*}\} \times \{1, \ldots, q_{k^*}\}\} \odot \mathcal{B}_{d-k^*}.$$

Similar to Theorem 6, it can be shown that the block thresholding estimator is adaptive under the hyperrectangular lattices.

Theorem 7. *Let $\hat{\Sigma}$ be the block thresholding estimate defined above with $s_0 = \lceil (\log p)^{1/d} \rceil$, and assume that $q_1 \leq q_2 \leq \cdots \leq q_d$. Then there exists a constant $C > 0$ such that*

$$\sup_{\Sigma \in \mathcal{F}(\alpha; M_0, M)} \mathbb{E}\|\hat{\Sigma} - \Sigma\|^2 \leq C \left(\frac{\log p}{n} + \min\left\{ \left(n^{-1} \prod_{l=0}^{k} q_l \right)^{\frac{2\alpha}{2\alpha+d-k}} : 0 \leq k \leq d \right\} \right),$$

where $q_0 = 1$, and

$$\sup_{\Sigma \in \mathcal{F}^*(\alpha_0, \alpha; M_0, M)} \mathbb{E}\|\hat{\Sigma} - \Sigma\|^2 \leq C \left(\frac{1}{n} \prod_{k=1}^{d} (\min\{q_k, (\log n)^{1/\alpha}\}) + \frac{\log p}{n} \right),$$

over all $\alpha > 0$.

The proof follows in a similar fashion as the hypercubic case by considering separately the "large" blocks and "small" blocks with the additional complication that it is not possible to make most blocks close to be cubic since the length along some directions may not be large enough. More precisely, we shall now refer to block $B \in \mathcal{B}_d^*$ "large" if

$$s(B) \geq C \left(n / \prod_{l=0}^{k^*} q_l \right)^{\frac{1}{2\alpha+d-k^*}}$$

for some constant $C > 0$. The details are omitted for brevity.

4 Numerical Experiments

The adaptive block thresholding procedure is easy to implement. We now conduct numerical experiments to illustrate the merits of the proposed adaptive block thresholding approach. We first consider a simple simulation study where the observations were generated from a Markov random field of order one. In particular, we simulate the stochastic process $X(t_1, t_2)$ $(t_1, t_2 \in \{1, \ldots, q\})$ such that

$$X(t_1, t_2) = 0.2 \left(X(t_1 - 1, t_2) + X(t_1, t_2 - 1) + X(t_1 + 1, t_2) + X(t_1, t_2 + 1) \right) + \epsilon(t_1, t_2)$$

where $\epsilon(t_1, t_2) \overset{iid}{\sim} N(0, 1)$. For each $q = 15, 25, 35$ or 45, 400 realizations of X were simulated. We computed both the sample covariance operator and the proposed block thresholding estimates with $\lambda = 1, 2$ or 6. One typical example when $q = 25$ was presented in Figure 1 in the introduction. For each choice of q, the experiment was repeated for 200 times, and for each run, the estimation error measured in terms of the operator norm is evaluated for each estimate. The results are summarized in Figure 5 where boxplots for each estimator are given.

It is evident that the block thresholding improves over the sample covariance operator. The improvement is particularly significant for large scale problem, that is when q is large. It is also interesting to note that in this simulation setting, $\lambda = 2$ appears to be a sensible choice.

Next, we apply the block thresholding estimator to the AT&T database of faces, a benchmark database in image analysis and face recognition. The data set contains a set of 400 face images taken between April 1992 and April 1994 at the AT&T laboratories in Cambridge, England. The images are taken for a total of 40 individuals. Each subject has 10 images of size 46×56 pixels (coalesced from original pictures of size 92×112), with 256 grey levels per pixel. The readers are referred to Samaria and Harter (1994) for further details about the database. Following the observations from the simulation study, we chose $\lambda = 2$ in this experiment. To visualize the resulting covariance operator estimate, Figure 6 gives the first 9 eigenimages corresponding to our estimate.

Several observations can be made from these eigenimages. First of all, it can be observed that most leading eigenimages pertain to local facial characteristics. In particular, most weights of the top three eigenimages are given to top portion of image, perhaps reflecting the different hairstyles or illumination on the forehead. To further appreciate the merits of our estimate, we compare

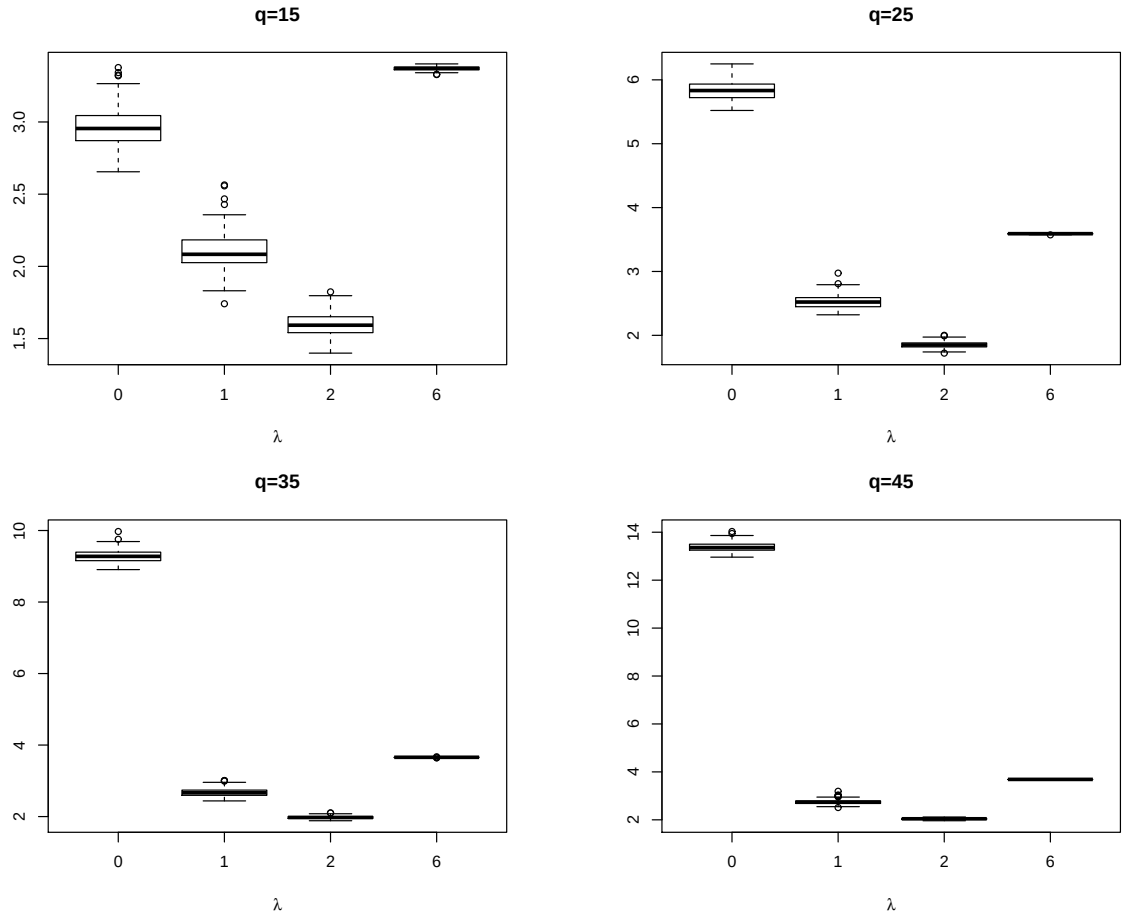


Figure 5: Each panel corresponds to a particular value of q. Reported here are the boxplots of the estimation error measured in operator norm for the block thresholding estimator with $\lambda = 1, 2$ or 6, along with the sample covariance operator corresponding to $\lambda = 0$.

the leading eigenimage from the block thresholding estimate with that from the sample covariance operator, hereafter referred as the sample eigenimage. Both eigenimages are given in the top panels of Figure 7: the left panel corresponds to the sample covariance operator whereas the right panel to the block thresholding estimator.

It is clear from Figure 7 that the sample eigenimage assigns weight over a broader area than the eigenimage estimated from the block thresholding estimate. The localization of loadings for our method could be more interpretable. We also remark that such localization does not come at the cost of capturing facial variation among individuals. The bottom panels of Figure 7 give the boxplots of the scores of the 10 images from each individual. It is clear that both estimates are fairly similar qualitatively. More precisely, ANOVA analysis shows that 90% of the variation of scores obtained from either method can be explained as the subject effect.

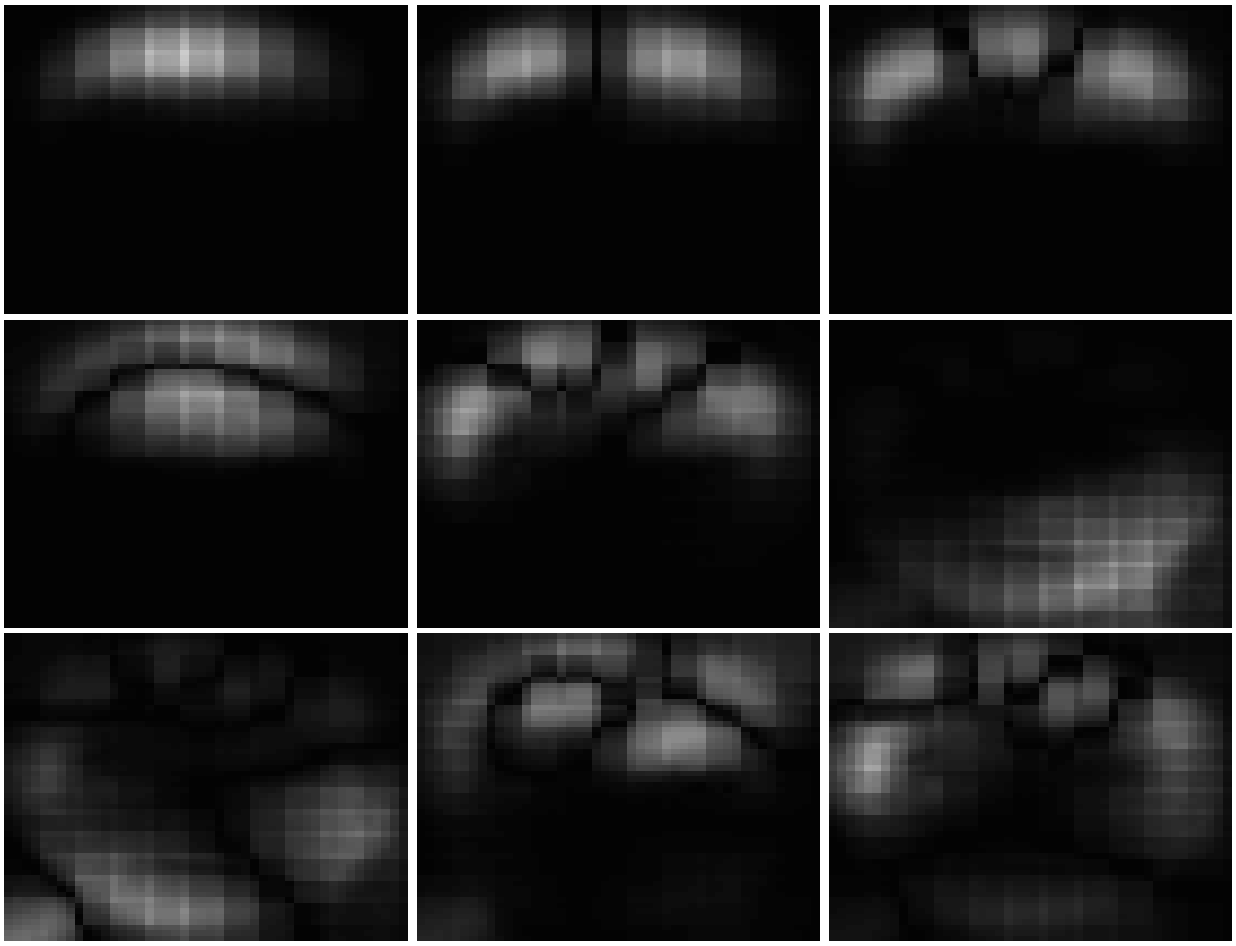


Figure 6: Estimated eigenimage – first 9 eigenimages corresponding to the block thresholding estimate, from left to right and top to bottom. The grey scale in each panel corresponds to the weight (absolute value) at each pixel, with the largest value represented by the brightest, and smallest value (0) represented by the darkest.

5 Discussions

In this article, we studied the minimax and adaptive estimation of covariance operators for random variables observed on a lattice graph. The framework is quite general. The more conventional covariance matrix estimation problem can be regarded as a special case where the random variables are observed on a one-dimensional lattice. To fix ideas, we focused in the present paper on two classes of covariance operators, those with polynomially decaying entries and those with exponentially decaying entries. We should note that the construction of the estimators and the technical tools developed in this paper are general and can be applied to other settings.

Consider for example the general parameter space $\mathcal{F}_d(\{a_k\}; M)$ defined in (1). Our results can be extended to other choices of $\{a_k : k \geq 1\}$. Let us focus on the hypercubic lattices. Define the

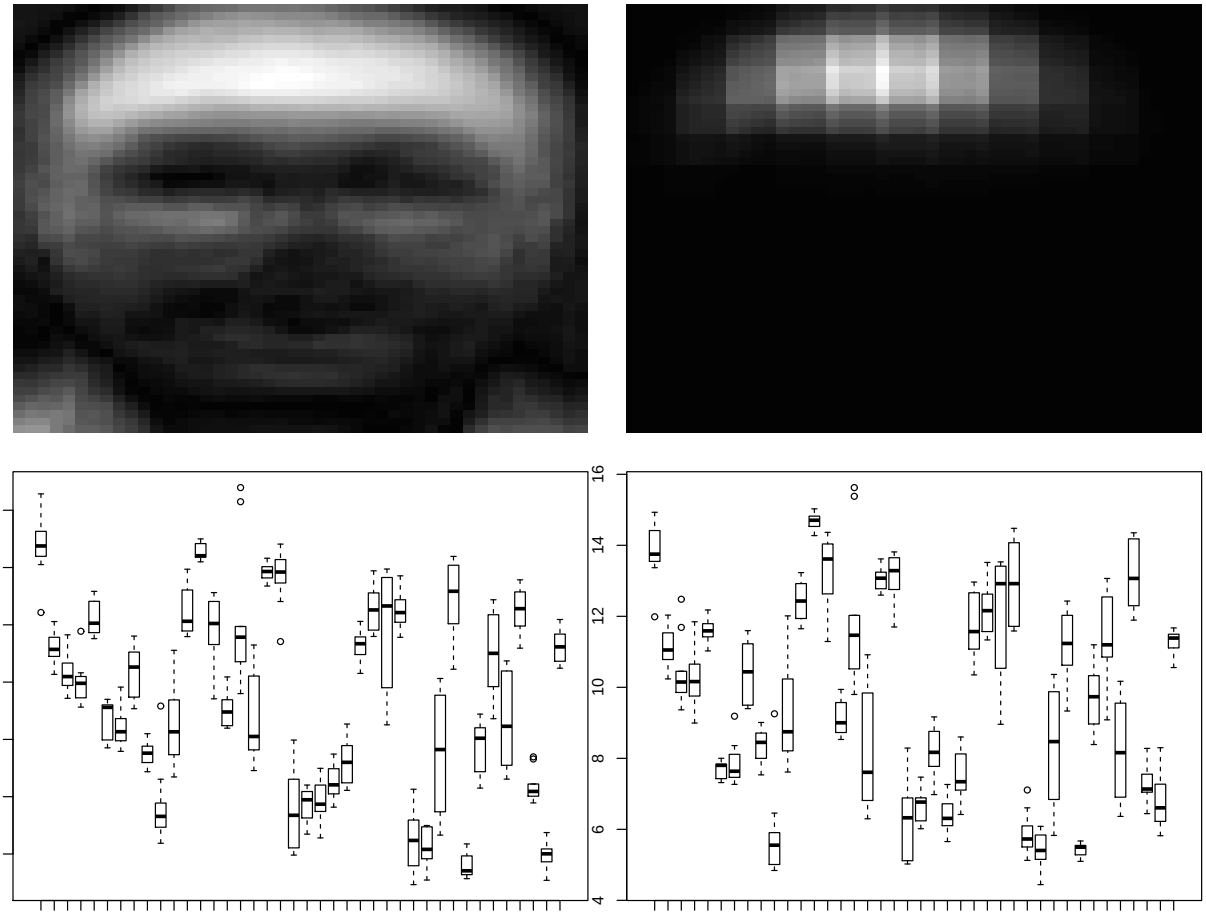


Figure 7: Leading eigenimage from the sample covariance operator and the block thresholding estimate – top panels show the weights (absolute value) assigned to each pixel. Bottom panels give the boxplots of the scores of the 10 images for each of the 40 subjects. Left panel corresponds to the sample covariance operator and right panel to the block thresholding estimate.

quantity $k(q)$ by

$$k(q) = \min\{1 \le k \le q : a_k \le n^{-1/2} k^{d/2-1}\}$$

if the set on the right hand side is non-empty, and $k(q) = q$ otherwise. Then following the same argument, it can be shown that the minimax rate of convergence is intimately related to the quantity $k(q)$. Under mild regularity conditions, the minimax risk for estimating the covariance operator over $\mathcal{F}_d(\{a_k\}; M)$ satisfies

$$\inf_{\tilde{\Sigma}(\text{data})} \sup_{\Sigma \in \mathcal{F}_d(\{a_k\};M)} \mathbb{E}\|\tilde{\Sigma} - \Sigma\|^2 \asymp \frac{[k(q)]^d + \log p}{n}.$$

Similar but more complicated rates can also be established for hyperrectangular lattices.

The techniques and results developed in this paper can also be used to solve other related problems. One such problem is the analysis of spatial data where X is a stochastic process defined

in a general metric space $(\mathcal{T}, D)$ with $\mathcal{T}$ of cardinality p. Taking into account the spatial structure when estimating the covariance operator is important in spatial analysis. A feature of spatial data that is distinct from the setting of the present paper is that the random variables are typically not observed on a regular lattice. For $r > 0$, define

$$N(r) = \max_{t \in \mathcal{T}} \text{card}\{s \in \mathcal{T} : D(s,t) \leq r\},$$

the largest number of elements of $\mathcal{T}$ contained in a ball of radius r. Assuming that

$$\max_{t \in \mathcal{T}} \sum_{s: D(s,t) \geq k} |\sigma(s,t)| \leq a_k,$$

then the minimax rate of convergence for estimating the covariance operator can also be established under certain regularity conditions. We shall report the details of the results elsewhere in the future as a significant amount of additional work is still needed.

6 Proofs

In this section we prove Theorem 6 and the technical lemmas used in the proof of the main results.

6.1 Proof of Theorem 6

We first consider polynomially decaying covariance operators $\mathcal{F}(\alpha; M_0, M)$.

6.1.1 Large blocks

The following result is a consequence of the construction of $\mathcal{B}_d$ and the properties of $\mathcal{B}_1$.

Lemma 3. *If* $(\mathbf{i}, \mathbf{j}) \in B \in \mathcal{B}_d$ *and* $s(B) \geq 2s_0$, *then* $D(\mathbf{i}, \mathbf{j}) > \|\mathbf{i} - \mathbf{j}\|_\infty \geq s(B)$.

Let $B = I \times J \in \mathcal{B}_d$. By Lemma 3, if $s(B) > 2^{L-1} s_0$ with $L > 1$, then

$$\|\Sigma_B\| \leq \|\Sigma_B\|_{\ell_1 \to \ell_1} \leq \max_{s \in \mathcal{G}_d} \sum_{t: D(s,t) \geq s(B)} |\sigma(s,t)| \leq M s^{-\alpha}(B).$$

On the other hand, by Lemma 2, there exists a constant $C > 1$ such that

$$\begin{aligned}
\|S_{I\times I} - \Sigma_{I\times I}\| &\leq C\|\Sigma_{I\times I}\| n^{-1/2} \left(s(B) + \log p\right)^{1/2}, \\
\|S_{J\times J} - \Sigma_{J\times J}\| &\leq C\|\Sigma_{J\times J}\| n^{-1/2} \left(s(B) + \log p\right)^{1/2}, \\
\|S_B - \Sigma_B\| &\leq C(\|\Sigma_{I\times I}\|\|\Sigma_{J\times J}\|)^{1/2} n^{-1/2} \left(s(B) + \log p\right)^{1/2},
\end{aligned}$$

with probability at least $1 - p^{-6}$. As a result,

$$\begin{aligned}
\|S_B\| &\leq M s^{-\alpha}(B) + C(\|\Sigma_{I\times I}\|\|\Sigma_{J\times J}\|)^{1/2} n^{-1/2} \left(s(B) + \log p\right)^{1/2} \\
&\leq 2C(\|\Sigma_{I\times I}\|\|\Sigma_{J\times J}\|)^{1/2} n^{-1/2} \left(s^d(B) + \log p\right)^{1/2} \\
&\leq 4C(\|S_{I\times I}\|\|S_{J\times J}\|)^{1/2} n^{-1/2} \left(s^d(B) + \log p\right)^{1/2}
\end{aligned}$$

provided that

$$2^{L-1}s_0 \geq (M/M_0)^{2/(2\alpha+d)} n^{1/(2\alpha+d)}. \tag{22}$$

Taking $\lambda_0 \geq 4C$ ensures $\hat{\Sigma}_B = 0$. By union bound, with probability at least $1 - p^{-4}$, $\hat{\Sigma}_B = 0$ for all $B \in \mathcal{B}_d$ such that $s(B) > 2^{L-1}s_0$. Let $W_L \in \{0,1\}^{\mathcal{G}_d \times \mathcal{G}_d}$ such that $w_L(s,t) = 1$ if and only if $(s,t) \in B \in \mathcal{B}_d$. Then

$$\begin{aligned}
\mathbb{E}\|(\hat{\Sigma} - \Sigma) \circ W_L\|^2 &= \mathbb{E}\left(\|(\hat{\Sigma} - \Sigma) \circ W_L\|^2 \mathbb{I}((\hat{\Sigma} - \Sigma) \circ W_L \neq 0)\right) \\
&\leq \left(\mathbb{E}\|(\hat{\Sigma} - \Sigma) \circ W_L\|^4\right)^{1/2} \mathbb{P}^{1/2}\{(\hat{\Sigma} - \Sigma) \circ W_L \neq 0\} \\
&\leq p^{-2}\left(\mathbb{E}\|(\hat{\Sigma} - \Sigma) \circ W_L\|^4\right)^{1/2} \\
&\leq p^{-2}\left(\mathbb{E}\|(\hat{\Sigma} - \Sigma) \circ W_L\|_{\mathrm{F}}^4\right)^{1/2},
\end{aligned}$$

where $\circ$ stands for the Schur product, i.e., elementwise product, and $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm. Observe that

$$\begin{aligned}
\mathbb{E}\|(\hat{\Sigma} - \Sigma) \circ W_L\|_{\mathrm{F}}^4 &= \mathbb{E}\left(\sum_{B \in \mathcal{B}_d : s(B) > 2^{L-1}s_0} \|\hat{\Sigma}_B - \Sigma_B\|_{\mathrm{F}}^2\right)^2 \\
&\leq 2\mathbb{E}\left(\sum_{B \in \mathcal{B}_d : s(B) > 2^{L-1}s_0} \|S_B - \Sigma_B\|_{\mathrm{F}}^2\right)^2 + 2\left(\sum_{B \in \mathcal{B}_d : s(B) > 2^{L-1}s_0} \|\Sigma_B\|_{\mathrm{F}}^2\right)^2 \\
&\leq 2M^4 p^4 n^{-2} + 2M^4 (2^{L-1}s_0)^{-4\alpha}.
\end{aligned}$$

Thus,

$$\mathbb{E}\|(\hat{\Sigma} - \Sigma) \circ W_L\|^2 = O\left(n^{-1}\right). \tag{23}$$

6.1.2 Small blocks

Now consider the smaller blocks. With slight abuse of notation, denote by $W_l \in \{0,1\}^{\mathcal{G}_d \times \mathcal{G}_d}$ where $w_l(s,t) = 1$ if and only if $(s,t) \in B \in \mathcal{B}_d$ such that $s(B) = l$. By triangular inequality,

$$\|\hat{\Sigma} - \Sigma\| \leq \left\|(\hat{\Sigma} - \Sigma) \circ \sum_{l=1}^{L-1} W_l\right\| + \|(\hat{\Sigma} - \Sigma) \circ W_L\|.$$

Therefore,

$$\mathbb{E}\|\hat{\Sigma} - \Sigma\|^2 \leq 2\mathbb{E}\left\|(\hat{\Sigma} - \Sigma) \circ \sum_{l=1}^{L-1} W_l\right\|^2 + 2\mathbb{E}\|(\hat{\Sigma} - \Sigma) \circ W_L\|^2.$$

Observe that

$$\begin{aligned}
\left\|(\hat{\Sigma} - \Sigma) \circ \sum_{l=1}^{L-1} W_l\right\| &\leq \sum_{l=1}^{L-1} \left\|(\hat{\Sigma} - \Sigma) \circ W_l\right\| \\
&\leq \sum_{l=1}^{L-1} \left(\sum_{1 \leq k \leq d} \left\|(\hat{\Sigma} - \Sigma) \circ W_{l,k}\right\| + \ldots + \left\|(\hat{\Sigma} - \Sigma) \circ W_{l,1\cdots d}\right\|\right),
\end{aligned}$$

where $w_{l,k}(s,t) = 1$ if and only if $(s,t) \in B \in \mathcal{B}_d$ for some $B \in \mathcal{A}'_k(l)$ and so on. The terms on the right hand side can be bounded in a similar fashion. We shall focus on $\|(\hat{\Sigma} - \Sigma) \circ W_{l,1}\|$ for brevity.

Recall that

$$\mathcal{A}'_1(l) = \mathcal{B}_1(l) \odot (\tilde{\mathcal{B}}_1(l))^{\odot(d-1)}.$$

Hence, for any $u \in \ell_2(\mathcal{G}_d)$,

$$\begin{aligned}
& \left\langle u, (\hat{\Sigma} - \Sigma) \circ W_{l,1} u \right\rangle \\
= & \sum_{B_1 = I_1 \times J_1 \in \mathcal{B}_1(l)} \sum_{B_2, \ldots, B_d = I_d \times J_d \in \tilde{\mathcal{B}}_1(l)} \left\langle u_{I_1 \times \cdots \times I_d}, (\hat{\Sigma}_{B_1 \odot \cdots \odot B_d} - \Sigma_{B_1 \odot \cdots \odot B_d}) u_{J_1 \times \cdots \times J_d} \right\rangle \\
\leq & \sum_{B_1 = I_1 \times J_1 \in \mathcal{B}_1(l)} \sum_{B_2, \ldots, B_d = I_d \times J_d \in \tilde{\mathcal{B}}_1(l)} \|\hat{\Sigma}_{B_1 \odot \cdots \odot B_d} - \Sigma_{B_1 \odot \cdots \odot B_d}\| \|u_{I_1 \times \cdots \times I_d}\| \|u_{J_1 \times \cdots \times J_d}\| \\
\leq & \frac{1}{2} \sup_{B \in \mathcal{A}'_1(l)} \|\hat{\Sigma}_B - \Sigma_B\| \sum_{B_1 = I_1 \times J_1 \in \mathcal{B}_1(l)} \sum_{B_2, \ldots, B_d = I_d \times J_d \in \tilde{\mathcal{B}}_1(l)} \left(\|u_{I_1 \times \cdots \times I_d}\|^2 + \|u_{J_1 \times \cdots \times J_d}\|^2 \right).
\end{aligned}$$

It is clear from the construction of $\mathcal{B}_d$ that if $(I_1 \times \cdots \times I_d) \times (J_1 \times \cdots \times J_d) \in \mathcal{B}_d$, then $(J_1 \times \cdots \times J_d) \times (I_1 \times \cdots \times I_d) \in \mathcal{B}_d$. Therefore,

$$\begin{aligned}
& \sum_{B_1 = I_1 \times J_1 \in \mathcal{B}_1(l)} \sum_{B_2, \ldots, B_d = I_d \times J_d \in \tilde{\mathcal{B}}_1(l)} \left(\|u_{I_1 \times \cdots \times I_d}\|^2 + \|u_{J_1 \times \cdots \times J_d}\|^2 \right) \\
& \qquad = 2 \sum_{B_1 = I_1 \times J_1 \in \mathcal{B}_1(l)} \sum_{B_2, \ldots, B_d = I_d \times J_d \in \tilde{\mathcal{B}}_1(l)} \|u_{I_1 \times \cdots \times I_d}\|^2.
\end{aligned}$$

In other words,

$$\left\| (\hat{\Sigma} - \Sigma) \circ W_{l,1} \right\| \leq \sup_{B \in \mathcal{A}'_1(l)} \|\hat{\Sigma}_B - \Sigma_B\| \sum_{B_1 = I_1 \times J_1 \in \mathcal{B}_1(l)} \sum_{B_2, \ldots, B_d = I_d \times J_d \in \tilde{\mathcal{B}}_1(l)} \|u_{I_1 \times \cdots \times I_d}\|^2.$$

Similarly, it can be shown that

$$\left\| (\hat{\Sigma} - \Sigma) \circ W_{l,k_1 k_2} \right\| \leq \sup_{B \in \mathcal{A}'_{k_1 k_2}(l)} \|\hat{\Sigma}_B - \Sigma_B\| \sum_{\substack{B_1, \ldots, B_d \in \tilde{\mathcal{B}}_1(l) \\ B_{k_1}, B_{k_2} \in \mathcal{B}_1(l)}} \|u_{I_1 \times \cdots \times I_d}\|^2$$

and so on. As a result,

$$\left\| (\hat{\Sigma} - \Sigma) \circ \sum_{l=1}^{L-1} W_l \right\| \leq \sup_{B \in \mathcal{B}_d : s(B) = l} \|\hat{\Sigma}_B - \Sigma_B\| \sum_{B_1 = I_1 \times J_1 \in \mathcal{B}_1(l)} \sum_{B_2, \ldots, B_d = I_d \times J_d \in \tilde{\mathcal{B}}_1(l)} \|u_{I_1 \times \cdots \times I_d}\|^2.$$

We now appeal to the following result.

Lemma 4. *Let $u \in \ell_2(\mathcal{G}_d)$ such that $\|u\| = 1$. Then*

$$\sum_{B_1 = I_1 \times J_1 \in \mathcal{B}_1(l)} \sum_{B_2, \ldots, B_d = I_d \times J_d \in \tilde{\mathcal{B}}_1(l)} \|u_{I_1 \times \cdots \times I_d}\|^2 \leq 13^d.$$

By Lemma 4, we get

$$\left\|(\hat{\Sigma}-\Sigma)\circ\sum_{l=1}^{L-1}W_l\right\| \le 13^d \sup_{B\in\mathcal{B}_d:s(B)<2^{L-1}s_0}\|\hat{\Sigma}_B-\Sigma_B\|.$$

Again by Lemma 2, there exists a constant $C>0$ such that

$$\|S_B-\Sigma_B\| \le CM_0n^{-1/2}(s^d(B)+\log p)^{1/2}$$

for all $B\in\mathcal{B}_d$ with probability at least $1-p^{-8}$. By the definition of $\hat{\Sigma}$, with the same probability,

$$\|\hat{\Sigma}_B-\Sigma_B\| \le CM_0n^{-1/2}(s^d(B)+\log p)^{1/2}$$

Therefore, with probability at least $1-p^{-8}$,

$$\left\|(\hat{\Sigma}-\Sigma)\circ\sum_{l=1}^{L-1}W_l\right\| \le Cn^{-1/2}\sum_{l=1}^{L-1}2^{d(l-1)/2}s_0^{d/2} \le Cn^{-1/2}s_0^{d/2}2^{dL/2}. \tag{24}$$

6.1.3 Adaptivity over $\mathcal{F}_d(\alpha;M_0,M)$

The adaptivity of the block thresholding follows from the bounds for large blocks and small blocks. More specifically, we call a block large if

$$s(B) \ge (M/M_0)^{2/(2\alpha+d)}n^{1/(2\alpha+d)}.$$

When $p<(M/M_0)^{2/(2\alpha+d)}n^{1/(2\alpha+d)}$, there are no large block. By the bound (24) for small blocks, we have $\|\hat{\Sigma}-\Sigma|\le C\sqrt{\frac{p}{n}}$ with probability at least $1-p^{-8}$. Denote $\mathcal{E}$ the even that the above inequality holds. Then

$$\mathbb{E}\left(\left\|\hat{\Sigma}-\Sigma\right\|^2\mathbb{I}(\mathcal{E})\right) \le \mathbb{E}^{1/2}\left(\left\|\hat{\Sigma}-\Sigma\right\|^4\right)\mathbb{P}^{1/2}(\mathcal{E}) \le \mathbb{E}^{1/2}\left(\sum_{B\in\mathcal{B}_d}\left\|\hat{\Sigma}_B-\Sigma_B\right\|\right)^4\mathbb{P}^{1/2}(\mathcal{E}).$$

We shall use the following lemma.

Lemma 5. *Let $\hat{\Sigma}$ be the block thresholding estimate defined above with $s_0=\lceil(\log p)^{1/d}\rceil$, then there exists a constant $C>0$ such that*

$$\mathbb{E}\left(\sum_{B\in\mathcal{B}_d}\left\|\hat{\Sigma}_B-\Sigma_B\right\|\right)^4 \le Cn^{-2}p^{10}.$$

Lemma 5 yields that $\mathbb{E}(\|\hat{\Sigma}-\Sigma\|^2\le\mathbb{E}(\|\hat{\Sigma}-\Sigma\|^2\mathbb{I}(\mathcal{E}))+Cp/n=O(p/n)$. When $s_0=\lceil(\log p)^{1/d}\rceil\ge(M/M_0)^{2/(2\alpha+d)}n^{1/(2\alpha+d)}$, only blocks of size s_0 will be preserved as small blocks and all block of size greater than s_0 will be treated as large blocks. In this case, following the small block bound (24), we have, with probability at least $1-p^{-8}$, $\|(\hat{\Sigma}-\Sigma)\circ W_1\|\le C\sqrt{\frac{\log p}{n}}$. Again denote by $\mathcal{E}$ the event that this inequality holds. Then by Lemma 5,

$$\mathbb{E}\left(\left\|(\hat{\Sigma}-\Sigma)\right\|^2\mathbb{I}(\mathcal{E})\right) \le \mathbb{E}^{1/2}\left(\left\|\hat{\Sigma}-\Sigma\right\|^4\right)\mathbb{P}^{1/2}(\mathcal{E})=O(p/n),$$

which implies that $\mathbb{E}\|(\hat{\Sigma}-\Sigma)\circ W_1\|^2 = O\left(\frac{\log p}{n}\right)$. Together with (23), we conclude that

$$\mathbb{E}\left\|\hat{\Sigma}-\Sigma\right\|^2 = O\left(\frac{\log p}{n}\right).$$

Similarly, when there are both large and small blocks by definition (22), it follows from (24) that

$$\mathbb{E}\left\|(\hat{\Sigma}-\Sigma)\circ\sum_{l=1}^{L-1}W_l\right\|^2 \leq Cn^{-\frac{2\alpha}{2\alpha+d}}.$$

Together with (23), this yields $\mathbb{E}\|\hat{\Sigma}-\Sigma\|^2 = O(n^{-\frac{2\alpha}{2\alpha+d}})$.

6.1.4 Adaptivity over $\mathcal{F}^*(\alpha_0,\alpha;M_0,M)$

This case can be proved in the exactly same way except that now a "large" block B satisfies $s(B)\geq 2s_0$ and $\exp(2\alpha_0 s^\alpha(B))s^{2d}(B)\geq Cn$ for some constant $C>0$. ∎

6.2 Proof of Auxiliary Results

6.2.1 Proof of Lemma 1

Similar to earlier work (see, e.g., Cai et al. 2010 and Yuan, 2010), the term $(\bar{X}(s)\bar{X}(t))_{s,t\in\mathcal{G}_d}$ is of higher order and negligible. For brevity, we shall ignore this term and focus on the first term of the sample covariance operator: $S_1 := \frac{1}{n}\sum_{i=1}^n X_i(s)X_i(t)$. It is well known that there exists a constant $c>0$ such that for any $u\in\ell_2(\mathcal{G}_d)$ obeying $\|u\|=1$,

$$\mathbb{P}\{\langle u,(S_1-\Sigma)u\rangle > x\} \leq \exp(-cnx^2).$$

See, e.g., Saulis and Statulevičius (1991). To make such a bound uniform over all u, note that for $u,u'\in\ell_2(\mathcal{G}_d)$ such that $\|u\|,\|u'\|=1$ and $\|u-u'\|\leq 1/4$,

$$\begin{aligned}\left|\langle u,(S_1-\Sigma)\,u\rangle - \langle u',(S_1-\Sigma)\,u'\rangle\right| &\leq \left|\langle u-u',(S_1-\Sigma)\,u\rangle\right| + \left|\langle u',(S_1-\Sigma)\,(u-u')\rangle\right| \\ &\leq 2\|u-u'\|\,\|S_1-\Sigma\| \\ &\leq \frac{1}{2}\|S_1-\Sigma\|.\end{aligned}$$

Let Q be the collection of centers of a $1/4$-cover set of the unit ball on $\ell_2(\mathcal{G}_d)$ such that $\mathrm{card}(Q)\leq C5^p$, for some constant $C>0$. By the union bound

$$\mathbb{P}\left\{\|S_1-\Sigma\|\geq x\right\} = \mathbb{P}\left\{\sup_{\|u\|=1}\langle u,(S_1-\Sigma)\,u\rangle \geq x\right\} \leq \mathbb{P}\left\{\sup_{u\in Q}\langle u,(S_1-\Sigma)\,u\rangle \geq x/2\right\} \leq C5^p\exp\left(-cnx^2\right),$$

for some constant $C>0$. Now observe that for any $x>0$

$$\begin{aligned}\mathbb{E}\,\|S_1-\Sigma\|^2 &\leq x^2\mathbb{P}\left\{\|S_1-\Sigma\|<x\right\} + \int_{x^2}^\infty \mathbb{P}\left\{\|S_1-\Sigma\|^2\geq u\right\}du \\ &\leq x^2 + C5^p\int_{x^2}^\infty \exp(-cnu)du \\ &\leq x^2 + C5^p(cn)^{-1}\exp(-cnx^2).\end{aligned}$$

Taking $x = c_1(p/n)^{1/2}$ for a sufficiently large constant $c_1 > 0$ yields $\mathbb{E}\left\|S_1 - \Sigma\right\|^2 \leq Cp/n$ for some constant $C > 0$. ∎

6.2.2 Proof of Lemma 2

It follows from a similar argument as that of Lemma 1 that for some constants $c_1, c_2 > 0$,

$$\mathbb{P}\left\{\left\|S_{(I\cup J)\times(I\cup J)} - \Sigma_{(I\cup J)\times(I\cup J)}\right\| \geq x\right\} \leq c_1 5^{2s} \exp\left(-c_2 n x^2\right).$$

The desired bound then follows from the fact that $\left\|S_{I\times J} - \Sigma_{I\times J}\right\| \leq \left\|S_{(I\cup J)\times(I\cup J)} - \Sigma_{(I\cup J)\times(I\cup J)}\right\|$.

6.2.3 Proof of Lemma 3

Recall that there exist $I_1, \ldots, I_d, J_1, \ldots, J_d \in \{1, \ldots, q\}$ such that $B = (I_1 \times \cdots \times I_d) \odot (I_1 \times \cdots \times I_d)$. Because $(\mathbf{i}, \mathbf{j}) \in B$, we know that $i_l \in I_l$ and $j_l \in J_l$ for $l = 1, \ldots, d$. Assume, without loss of generality, that $\mathrm{card}(I_1) = s(B)$. Then from the construction of $\mathcal{B}_d$, it is clear that $B_1 := I_1 \times J_1 \in \mathcal{B}_1(s(B))$. Note that $(i_1, j_1) \in B_1$ and $s(B_1) > 2s_0$. By the property of $\mathcal{B}_1$ (see, e.g., Cai and Yuan, 2011), we conclude that $|i_l - j_l \geq s(B_1)$. The proof is now completed because $d(\mathbf{i}, \mathbf{j}) \geq |i_l - j_l|$ and $s(B_1) = s(B)$.

6.2.4 Proof of Lemma 4

We proceed by induction, starting with $d = 1$. Observe that for large enough l, all blocks of $\tilde{\mathcal{B}}(l)$ can be expressed as

$$\begin{aligned}
&\{((k-1)2^{l-1}+1)s_0, \ldots, (k2^{l-1}-3)s_0\} \times \{((k'-1)2^{l-1}+1)s_0, \ldots, (k'2^{l-1}-3)s_0\};\\
&\{(k2^{l-1}-2)s_0, (k2^{l-1}-1)s_0, k2^{l-1}s_0\} \times \{((k'-1)2^{l-1}+1)s_0, \ldots, (k'2^{l-1}-3)s_0\};\\
&\{((k-1)2^{l-1}+1)s_0, \ldots, (k2^{l-1}-3)s_0\} \times \{(k'2^{l-1}-2)s_0, (k'2^{l-1}-1)s_0, k'2^{l-1}s_0\};\\
&\{(k2^{l-1}-2)s_0, (k2^{l-1}-1)s_0, k2^{l-1}s_0\} \times \{(k'2^{l-1}-2)s_0, (k'2^{l-1}-1)s_0, k'2^{l-1}s_0\}.
\end{aligned}$$

for some k, k' obeying $|k - k'| < 3$. As a result,

$$\sum_{B=I\odot J\in\tilde{\mathcal{B}}_1(l)} \|u_I\|_{\ell_2}^2 \leq 7\|u\|_{\ell_2}^2 \leq 7.$$

Together with the fact that $\sum_{B=I\odot J\in\mathcal{B}_1(l)} \|u_I\|_{\ell_2}^2 \leq 6\|u\|_{\ell_2}^2 \leq 6$, we conclude that

$$\sum_{B=I\odot J\in\mathcal{B}_1(l)\cup\tilde{\mathcal{B}}_1(l)} \|u_I\|_{\ell_2}^2 \leq 13.$$

Now assume that $\sum_{B_1,\ldots,B_d\in\mathcal{B}_1(l)\cup\tilde{\mathcal{B}}_1(l)} \|u_{I_1\times\cdots\times I_d}\|_{\ell_2}^2 \leq 13^d$, where $B_j = I_j \times J_j$. Then

$$\sum_{B_1,\ldots,B_d,B_{d+1}\in\mathcal{B}_1(l)\cup\tilde{\mathcal{B}}_1(l)} \|u_{I_1\times\cdots\times I_{d+1}}\|_{\ell_2}^2 \leq 13^d \sum_{B_{d+1}\in\mathcal{B}_1(l)\cup\tilde{\mathcal{B}}_1(l)} \|u_{\mathcal{G}_d\times I_{d+1}}\|_{\ell_2}^2 \leq 13^{d+1}.$$ ∎

6.2.5 Proof of Lemma 5

Observe that

$$\begin{aligned}
\mathbb{E}\left(\sum_{B\in\mathcal{B}_d}\left\|\hat{\Sigma}_B-\Sigma_B\right\|\right)^4 &\le p^6\mathbb{E}\left(\sum_{B\in\mathcal{B}_d}\left\|\hat{\Sigma}_B-\Sigma_B\right\|^4\right)\\
&\le p^6\mathbb{E}\left(\sum_{B\in\mathcal{B}_d}\left(\|S_B-\Sigma_B\|+\lambda_0\sqrt{\frac{s^d(B)+\log p}{n}}\right)^4\right)\\
&\le Cp^6\left(\mathbb{E}\sum_{B\in\mathcal{B}_d}\|S_B-\Sigma_B\|^4+\lambda_0^4\left(\frac{s^d(B)+\log p}{n}\right)^2\right).
\end{aligned}$$

Together with the fact that $\mathbb{E}\sum_{B\in\mathcal{B}_d}\|S_B-\Sigma_B\|^4 \le \mathbb{E}\sum_{B\in\mathcal{B}_d}\|S_B-\Sigma_B\|_{\mathrm{F}}^4 \le Cn^{-2}p^4$, we get $\mathbb{E}\left(\sum_{B\in\mathcal{B}_d}\left\|\hat{\Sigma}_B-\Sigma_B\right\|\right)^4 \le Cn^{-2}p^{10}$. ∎

References

[1] Bickel, P. and Levina, E. (2008a). Regularized estimation of large covariance matrices. *The Annals of Statistics* **36**, 199-227.

[2] Bickel, P. and Levina, E. (2008b). Covariance regularization by thresholding. *The Annals of Statistics* **36**, 2577-2604.

[3] Cai, T.T. and Liu, W.(2011). Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association* **494**, 672-684.

[4] Cai, T.T., Liu, W. and Luo, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association* **494**, 594-607.

[5] Cai, T. T., Liu, W. and Zhou, H. H. (2011). Optimal estimation of large sparse precision matrices. Manuscript.

[6] Cai, T. T. and Yuan, M. (2011). Adaptive covariance matrix estimation through block thresholding. *The Annals of Statistics*, to appear.

[7] Cai, T.T., Zhang, C.H. and Zhou, H. (2010). Optimal rates of convergence for covariance matrix estimation. *The Annals of Statistics* **38**, 2118-2144.

[8] Cai, T.T. and Zhou, H. (2012). Optimal rates of convergence for sparse covariance matrix estimation. *The Annals of Statistics*, to appear.

[9] El Karoui, N. (2008). Operator norm consistent estimation of large dimensional sparse covariance matrices. *The Annals of Statistics* **36**, 2717-2756.

[10] Fan, J., Fan, Y. and Lv, J. (2008). High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics* **147**, 186-197

[11] Friedman, J., Hastie, T. and Tibshirani, T. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9**, 432-441.

[12] Huang, J., Liu, N., Pourahmadi, M., and Liu, L. (2006). Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika* **93**, 85-98.

[13] Krause, E. (1987). *Taxicab Geometry*. Dover, New York.

[14] Lam, C. and Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *The Annals of Statistics* **37**, 4254-4278.

[15] Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis* **88**, 365-411.

[16] Muirhead, R. (2005). *Aspects of Multivariate Statistical Theory*. Wiley, London.

[17] Rothman, A., Bickel, P., Levina, E. and Zhu, J. (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics* **2**, 494-515.

[18] Rothman, A., Levina, E. and Zhu, J. (2009). Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association* **104**, 177-186.

[19] Samaria, F. and Harter, A. (1994). Parameterisation of a stochastic model for human face identification. *Proceedings of 2nd IEEE Workshop on Applications of Computer Vision*, Sarasota, FL.

[20] Saulis, L. and Statulevičius, V. A. (1991). *Limit Theorems for Large Deviations*. Springer, Berlin.

[21] Sirovich, L. and Kirby, M. (1987). Low-dimensional procedure for the characterization of human faces. *Journal of the Optical Society of America A* **4(3)**, 519-524.

[22] Tsybakov, A. (2009). *Introduction to Nonparametric Estimation*. Springer, New York.

[23] Turk, M. and Pentland, A. (1991). Eigenfaces for recognition. *Journal of Cognitive Neuroscience* **3(1)**, 71-86.

[24] Yuan, M. (2010). Sparse inverse covariance matrix estimation via linear programming. *Journal of Machine Learning Research* **11**, 2261-2286.

[25] Yuan, M. and Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika* **94**, 19-35.