Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA

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

T. Tony Cai and Ming Yuan

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 article, 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. Supplementary materials for this article are available online.

KEY WORDS: Block thresholding; Covariance matrix; Covariance operator; Minimax estimation; Operator norm; Optimal rate of convergence.



## 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—a standard task in imaging analysis, especially for the purpose of face recognition (see, e.g., Sirovich and Kirby 1987; Turk and Pentland 1991). Typically eigenimages are estimated directly from the sample covariance operator which does not account for the lattice structure of an image, but recent results suggest that consistency can only be achieved with a prohibitive sample size requirement; see, for example, Rudelson (1999) and Johnstone and Lu (2009). In this article, we develop new estimation procedures which are specifically designed to account for the lattice structure and show that, in doing so, drastically improved performance of covariance estimation can be achieved.

Let $\mathcal{G}(q_1, \ldots, q_d) = [q_1] \times [q_2] \times \cdots \times [q_d]$ be a $d$-dimensional lattice where $[q] = \{1, 2, \ldots, q\}$. Assume without loss of generality that $q_1 \le q_2 \le \cdots \le q_d$. 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 $\Sigma$ is defined over the Cartesian product space of $\mathcal{G}_d \times \mathcal{G}_d$, that is, $\Sigma \in \mathbb{R}^{\mathcal{G}_d \times \mathcal{G}_d}$. A particularly interesting case here is when the number of variables $p := q_1 q_2, \ldots, q_d$ 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 $\Sigma$ corresponds to a compact operator from $\ell_2(\mathcal{G}_d)$ to 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 nonincreasing sequence $a_k \downarrow 0$ and a constant $M > 0$ where

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

To fix ideas, in what follows, we shall take $D(\cdot,\cdot)$ to be the Manhattan or equivalently $\ell_1$ 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 $\mathcal{G}_d$.

T. Tony Cai is Professor, Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104. (E-mail: *tcai@wharton.upenn.edu*). Ming Yuan is Professor, Department of Statistics, University of Wisconsin-Madison, Madison, WI 53706 (E-mail: *myuan@stat.wisc.edu*). The research of Ming Yuan was supported in part by NSF Career Award DMS-1321692 and FRG Grant DMS-1265202. The research of Tony Cai was supported in part by NSF FRG Grant DMS-0854973.

Color versions of one or more of the figures in the article can be found online at *www.tandfonline.com/r/jasa*.



DOI: 10.1080/01621459.2014.1001067

We study in this article 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 distances whereas the latter consists of covariance operators where the covariances decay exponentially fast with their distances. We shall consider sub-Gaussian variables $X$ which satisfy, for some constant $\rho > 0$,

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

Denote by $\mathcal{P}_d(\alpha; M_0, M)$ the collection of sub-Gaussian 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 sub-Gaussian 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 $n$. We establish the following minimax rates of convergence for estimating $\Sigma$ 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 $q_1 \leq \cdots \leq q_d$. Given a random sample $X_1, \ldots, X_n$ from the distribution of $X$. The minimax risk for estimating the covariance operator $\Sigma$ under the operator norm $\|\cdot\|$ satisfies

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

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

$$\inf_{\tilde{\Sigma}(\text{data})} \sup_{\Sigma\in\mathcal{F}_2(\alpha;M_0,M)} \mathbb{E}\|\tilde{\Sigma} - \Sigma\|^2 \asymp \frac{\log(q_1q_2)}{n} + \min\left\{n^{-\frac{\alpha}{\alpha+1}}, \left(\frac{q_1}{n}\right)^{\frac{2\alpha}{2\alpha+1}}, \frac{q_1q_2}{n}\right\}. \tag{5}$$

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 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 2012). The setting here is, however, more complicated due to the lattice structure.

Our work relates to a fast growing literature on estimation of structured 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 (2012), 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 (2012) 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 article is organized as follows. After introducing basic notation and definitions, Section 2 establishes the minimax rates of covergence 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.

## 2. OPTIMAL RATES OF CONVERGENCE

In this section, we establish the optimal rates of convergence for estimating the covariance operator $\Sigma$. We begin by introducing some basic notation and definitions. Throughout the article, 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 $\Sigma$ 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 $\|\Sigma\|_{\ell_2 \to \ell_2}$.

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. More specifically, we have the following lower bounds for the minimax risk of estimating $\Sigma$ over $\mathcal{F}_d(\alpha; M_0, M)$ or $\mathcal{F}_d^*(\alpha_0, \alpha; M_0, M)$.

*Theorem 2.* Suppose that we observe a random sample $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(0, \Sigma)$ 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

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

and

$$\inf_{\tilde{\Sigma}(\text{data})} \sup_{\Sigma \in \mathcal{F}_d^*(\alpha_0, \alpha; M_0, M)} \mathbb{E}\|\tilde{\Sigma} - \Sigma\|^2 \geq C\left(\frac{\log p}{n} + \frac{1}{n}\prod_{k=1}^{d}(\min\{q_k, (\log n)^{1/\alpha}\})\right), \tag{9}$$

where $q_0 = 1$.

We now show the lower bounds given in Theorem 2 are indeed tight and the convergence rates are achievable. Without loss of generality, we shall assume in the rest of the article 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 sub-Gaussian random process $X$ defined over $\mathcal{G}_d$ with covariance operator $\Sigma$. 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 (9) for exponentially decaying covariance operator if $q \leq q_*$ where $q_*$ is defined as the solution to

$$\log n + d \log x = 2\alpha_0 x^{\alpha}. \tag{10}$$

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 $\Sigma$ and show that it can attain the rates from Theorem 2 under these settings.

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"

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

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

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

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

$$\hat{\Sigma}_{\mathbf{j}\mathbf{j}'} = \begin{cases} S_{\mathbf{j}\mathbf{j}'}, & \text{if } \|\mathbf{j} - \mathbf{j}'\|_\infty \le 1 \\ 0, & \text{otherwise.} \end{cases} \tag{11}$$

In other words, we estimate $\Sigma_{\mathbf{j}\mathbf{j}'}$ by its sample counterpart if and only if the two blocks $B_{\mathbf{j}}$ and $B_{\mathbf{j}'}$ are "neighbors," as illustrated in Figure 1.

We now show that with appropriate choice of $b$, the proposed estimator $\hat{\Sigma}$ can achieve the optimal rate of convergence. In particular, when $\Sigma \in \mathcal{F}_d(\alpha; M_0, M)$, we take

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

where

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

On the other hand, in the case of $\mathcal{F}_d^*(\alpha_0, \alpha; M_0, M)$, we shall take $b = \lceil q_* \rceil$ to be the block size where $q_*$ is the solution to (10). With these choices, we have

*Theorem 3.* Suppose that we observe a random sample $X_1, \ldots, X_n$ consisting of independent copies of a sub-Gaussian random process $X$ defined over $\mathcal{G}_d$ and wish to estimate its covariance operator $\Sigma \in \mathbb{R}^{\mathcal{G}_d \times \mathcal{G}_d}$. 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

$$\sup_{\Sigma \in \mathcal{F}_d(\alpha; M_0, M)} \mathbb{E} \|\hat{\Sigma} - \Sigma\|^2 \le C \left( \frac{\log p}{n} + \min \left\{ \left( n^{-1} \prod_{l=0}^{k} q_l \right)^{\frac{2\alpha}{2\alpha+d-k}} : 0 \le k \le d \right\} \right), \tag{12}$$

provided that $q > n^{1/(2\alpha+d)}$; and

$$\sup_{\Sigma \in \mathcal{F}_d^*(\alpha_0, \alpha; M_0, M)} \mathbb{E} \|\hat{\Sigma} - \Sigma\|^2 \le C \left( \frac{\log p}{n} + \frac{1}{n} \prod_{k=1}^{d} (\min\{q_k, (\log n)^{1/\alpha}\}) \right), \tag{13}$$

provided that $q > q_*$ where $q_*$ is the solution to (10).

Together with the lower bound given in Theorem 2, this etsablishes that the optimal rate of convergence for estimating $\Sigma \in \mathcal{F}_d^*(\alpha; M_0, M)$ and $\mathcal{F}_d^*(\alpha_0, \alpha; M_0, M)$ and the blockwise banding estimator is rate optimal.

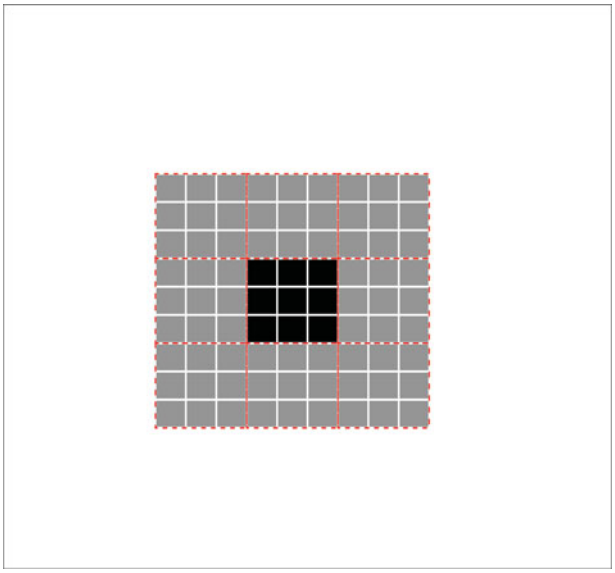

Figure 1. Blocks and their "neighbors"—A two-dimensional example of the blocking scheme. In this case, $k = 3$ and the blocks are represented with red dashed lines as boundary. The gray blocks are the "neighbors" of the solid black block.

## 3. ADAPTIVE BLOCK THRESHOLDING

Although the blockwise banding estimator proposed in the last section achieves the optimal rate of convergence, it is evident from its construction that it depends on the explicit knowledge of $\alpha$ which is typically unknown in practice. This makes the concept of adaptive estimation—a single estimator, not depending on the decay rate $\alpha$, that achieves the optimal rate of convergence simultaneously—of great practical importance. In this 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 $\alpha > 0$.

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 2012).

Recall that $\Sigma$ 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 (2012) for covariance matrices which corresponds to the case $d = 1$. Note that a covariance matrix is defined over the Cartesian product space of $[q] \times [q]$. The construction begins by dividing the two-dimensional lattice into blocks of size $s_0 \times s_0$ for some $s_0$, and the blocks are then consolidated systematically. More specifically, the blocks are created as follows:

- Start by constructing blocks of size $k \times k$ where $k = s_0$
  - Create blocks on the diagonal
  - Create more blocks successively toward right and toward bottom

* Two or one in an alternating fashion

- Double block size $k = 2k$ and create blocks successively
  - Three or two in an alternating fashion
- Continue until the whole matrix is covered

The final blocking of $[q] \times [q]$ is shown in Figure 2. Denote by $\mathcal{B}_l$ the collection of blocks with size $2^{l-1}s_0$. For example, $\mathcal{B}_1$ consists of the solid black blocks in Figure 2. A key property of the blocking scheme is that for any block $B \in \mathcal{B}_l$,

$$\min_{(i,j)\in B} |i - j| \geq 2^{l-1}s_0.$$

Interested readers are referred to Cai and Yuan (2012) for details.

Let

$$\mathcal{A}_l = \{(i, j) \in [q] \times [q] : (i, j) \in B \text{ for some } B \in \mathcal{B}_k \text{ and } k \leq l\}.$$

That is, $\mathcal{A}_l$ consists of the entries of a $q \times q$ symmetric matrix that are covered by blocks of size no greater than $2^{l-1}s_0$. The total number of blocks to cover $\mathcal{A}_l$ is therefore

$$|\mathcal{B}_1| + \cdots + |\mathcal{B}_l| \approx p/s_0 + p/(2s_0) + \cdots + p/(2^{l-1}s_0) \asymp p/s_0.$$

It turns out that for our purposes, this could be too many. To address this issue, we shall consider a reconfiguration of $\mathcal{A}_l$ so that they can be covered by a smaller number of blocks. To this end, consider a regular blocking at $\{((k-1)2^l + 1)s_0, (k2^l - 2)s_0 : k = 1, 2, \ldots\}$, that is, blocks of one of the following four configurations:

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

for some $k$ and $k'$. It is clear that, in general, the first three types of blocks are of size $(2^l - 3)s_0$ whereas the fourth type is of size $3s_0$. As an example, the reconfiguration for $\mathcal{A}_2$ is given in Figure 3. We denote by $\bar{\mathcal{B}}_l$ the collection of blocks that cover $\mathcal{A}_l$ after the reconfiguration. The main advantage of the reconfiguration is that now the number of blocks needed to cover $\mathcal{A}_l$ is of the order $p/(2^{l-1}s_0)$.

We are now in position to describe the blocking scheme for $\mathcal{G}_d \times \mathcal{G}_d$ when $d > 1$. To fix ideas, we focus on hypercubic lattices, and the discussion can be straightforwardly extended to accommodate the more general hyperrectangular lattices although the presentation is much more tedious. We shall adopt the following notation. Let $B_1, B_2, \ldots, B_k \subseteq \mathcal{G}_2$, write

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

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

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

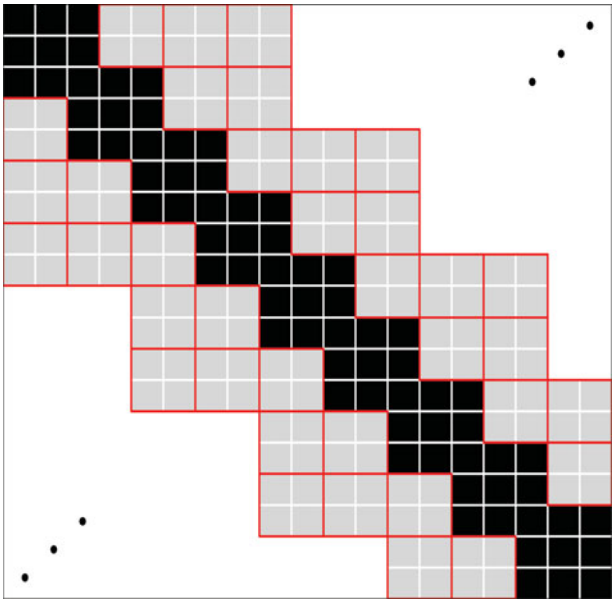

Figure 2. Blocking scheme for covariance matrices—Blocks are with increasing sizes away from the diagonal. The solid black blocks are of size $s_0 \times s_0$. The gray ones are consolidated to be of size $2s_0 \times 2s_0$.

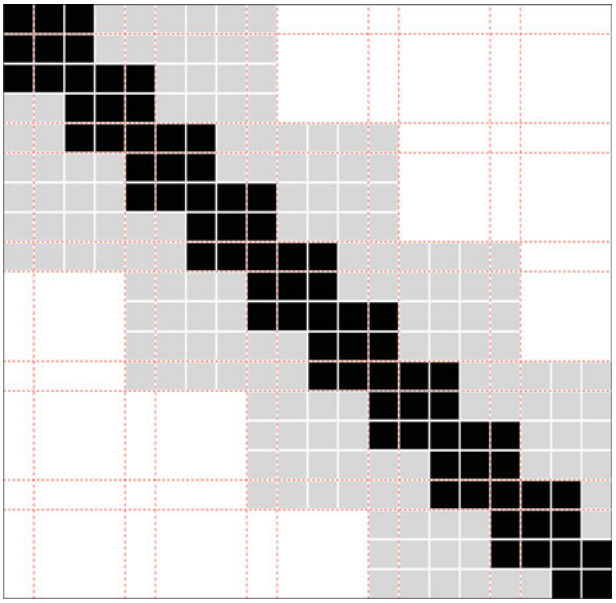

Figure 3. Reconfiguration of $\mathcal{A}_2$: Original blocks of size $s_0$ are represented as black whereas the area covered by blocks of original size $2s_0$ is in gray. Dashed lines show the reconfigured blocks.

We then consider a blocking of $\mathcal{G}_d \times \mathcal{G}_d$ as

$$\mathcal{C} := \bigcup_{1 \le l \le \lceil \log(p/s_0) \rceil} \mathcal{C}_l, \tag{14}$$

where

$$\mathcal{C}_l = \bigcup_{1 \le s \le d} \left( \bigcup_{1 \le i_1 \le \cdots \le i_s \le d} \bar{\mathcal{B}}_{l-1}^{\odot(i_1-1)} \odot \mathcal{B}_l \odot \bar{\mathcal{B}}_{l-1}^{\odot(i_2-i_1-1)} \odot \mathcal{B}_l \odot \cdots \odot \bar{\mathcal{B}}_{l-1}^{\odot(d-i_s)} \right). \tag{15}$$

For example, in the special case of $d = 2$,

$$\mathcal{C}_l = \left(\bar{\mathcal{B}}_{l-1} \odot \mathcal{B}_l\right) \bigcup \left(\mathcal{B}_l \odot \bar{\mathcal{B}}_{l-1}\right) \bigcup \left(\mathcal{B}_l \odot \mathcal{B}_l\right).$$

A few properties of the blocking scheme immediately follow.

*Lemma 2.* Let $\mathcal{C}$ and $\mathcal{C}_l$ be defined by (14) and (15). Then

- $\mathcal{C}$ consists of "blocks" that divide $\mathcal{G}_d \times \mathcal{G}_d$ in that, for any $B, B' \in \mathcal{C}$ and $B \ne B'$, $B \cap B = \emptyset$; and
$$\bigcup_{B \in \mathcal{C}} B = \mathcal{G}_d \times \mathcal{G}_d.$$
- $\mathcal{C}_l$ is the collection of blocks that has "size" no greater than $2^{l-1}s_0$ in that for any $B \in \mathcal{C}_l$, there exist $I_k = (a_k, b_k]$ for some $0 \le a_k < b_k \le q$ and $k = 1, \ldots, 2d$ such that
$$B = I_1 \times I_2 \times \cdots \times I_{2d}, \text{ and}$$
$$s(B) := \max_{1 \le k \le 2d} (b_k - a_k) \le 2^{l-1}s_0.$$
- There exists a constant $c > 0$ depending on $d$ only such that the number of elements in $\mathcal{C}_l$,
$$|\mathcal{C}_l| \le c(p/(2^{l-1}s_0))^d.$$

Once the blocking is defined, we then proceed to estimate the covariance operator $\Sigma$ block by block. By Lemma 2, 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, that is, $I_l = J_l$ for $l = 1, \ldots, d$, we shall estimate $\Sigma_B$ by its sample counterpart. If $B$ is large in that $s^d(B) > n/\log n$, we estimate $\Sigma_B$ simply by zero. For other blocks, we estimate $\Sigma_B$ 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 numerical constant. Similar to the covariance matrix case, our theoretical development indicates that the resulting block thresholding estimator is optimally rate adaptive whenever $\lambda_0$ is a sufficiently large. 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 $\lambda_0$ could potentially lead to further improved finite sample performance.

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 4.* 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

$$\sup_{\mathcal{P}(\alpha; M_0, M)} \mathbb{E}\|\hat{\Sigma} - \Sigma\|^2 \le C \min\left\{ n^{-\frac{2\alpha}{2\alpha+d}} + \frac{\log p}{n}, \frac{p}{n} \right\}, \tag{16}$$

and

$$\sup_{\mathcal{P}^*(\alpha_0, \alpha; M_0, M)} \mathbb{E}\|\hat{\Sigma} - \Sigma\|^2 \le C \min\left\{ \frac{(\log n)^{d/\alpha}}{n} + \frac{\log p}{n}, \frac{p}{n} \right\} \tag{17}$$

over all $\alpha > 0$.

## 4. NUMERICAL EXPERIMENTS

The proposed adaptive block thresholding procedure is easy to implement. We shall now present some numerical experiments to illustrate its merits.

We first conduct a set of simulation study following a Markov random field model. Because Markov random field models give rise to good, flexible, stochastic image models, they are commonly used in many areas of image processing. In particular, we simulated the stochastic process $X(t_1, t_2)$ $(t_1, t_2 \in \{1, \ldots, q\})$ such that

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

where $\epsilon(t_1, t_2) \overset{\text{iid}}{\sim} N(0, 1)$. We compare in particular the proposed adaptive block thresholding method with the sample covariance operator and a simple thresholding approach (see, e.g., Bickel and Levina 2008b). As indicated in the theoretical development, it is sufficient to take $\lambda_0 = 6$ for the adaptive block thresholding. The thresholding estimator of Bickel and Levina (2008b) does not take into account of the lattice graph structure and simply zero out those entries of the sample covariance operator that have small absolute values. The threshold level, as suggested by Bickel and Levina (2008b), is determined by cross-validation.

To illustrate the importance of accounting for the lattice graph structure, we first present results from a typical simulated dataset with $q = 25$. For a total of $n = 400$ simulated images, we computed the sample covariance operator, thresholding covariance operator, and the adaptive block thresholding estimator. To fix ideas, we consider estimating the leading eigenimage—a standard task in imaging analysis, especially for the purpose of face recognition (see, e.g., Sirovich and Kirby 1987; Turk and Pentland 1991). 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, for example, Turk and Pentland (1991). With a relatively small sample size, such an estimate may be unreliable. The absolute value of the loadings of the leading eigenimage for the true covariance operator and the three estimates are given in Figure 4. It is clear, visually, that the adaptive block thresholding method leads to a superior estimate of the eigenimage. More quantitatively, the correlation between the truth and the sample eigenimage and

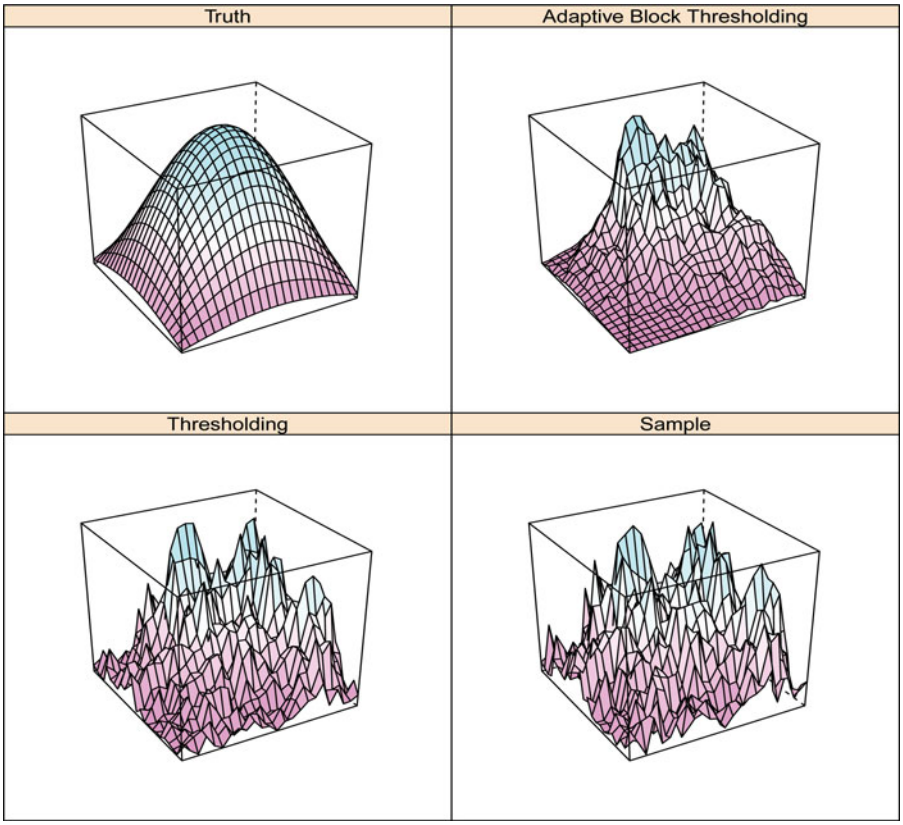


Figure 4. Importance of accounting for the lattice structure of images—loadings of the true leading eigenimage, along with the estimate derived from adaptive block thresholding, simple thresholding, and sample covariance operator.

the eigenimage estimated from simple thresholding is 41% and 43%, respectively, indicating their rather poor performance. As a comparison, the correlation between the true eigenimage and the eigenimage corresponding to the adaptive block thresholding estimate is 81%, which represents a significant improvement.

To further compare these different methods for estimating the covariance operator, we now consider different values of $q$: $q = 25, 35$, or 45. For each dataset, 400 realizations of $X$ were simulated, and the four estimators, sample, thresholding, adaptive block thresholding with $\lambda_0 = 6$ were evaluated. 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.

Next, for illustration purposes, we apply the block thresholding estimator to the AT&T database of faces, a benchmark database in image analysis, and face recognition (Samaria and Harter 1994). The dataset 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 ten images of size $46 \times 56$ pixels (coalesced from original pictures of size $92 \times 112$), with 256 gray levels per pixel. The readers are referred to Samaria and Harter (1994) for further details about the database. To visualize the resulting covariance operator estimate, the top panel of Figure 6 gives the first three 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 computed the scores corresponding to the leading eigenimages for each of the 400 images. The scores are given in the bottom panels of Figure 6. Images from the same subject correspond to points with the same color and symbol. It is noteworthy that the leading eigenimages appear to capture the main characteristics of the subject in that the images corresponding to a common subject tend to cluster together.

## 5. DISCUSSIONS

We studied in this article the minimax and adaptive estimation of covariance operators for random variables observed on a lattice graph in a general framework. 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 article 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 article are general and can be applied to other settings.

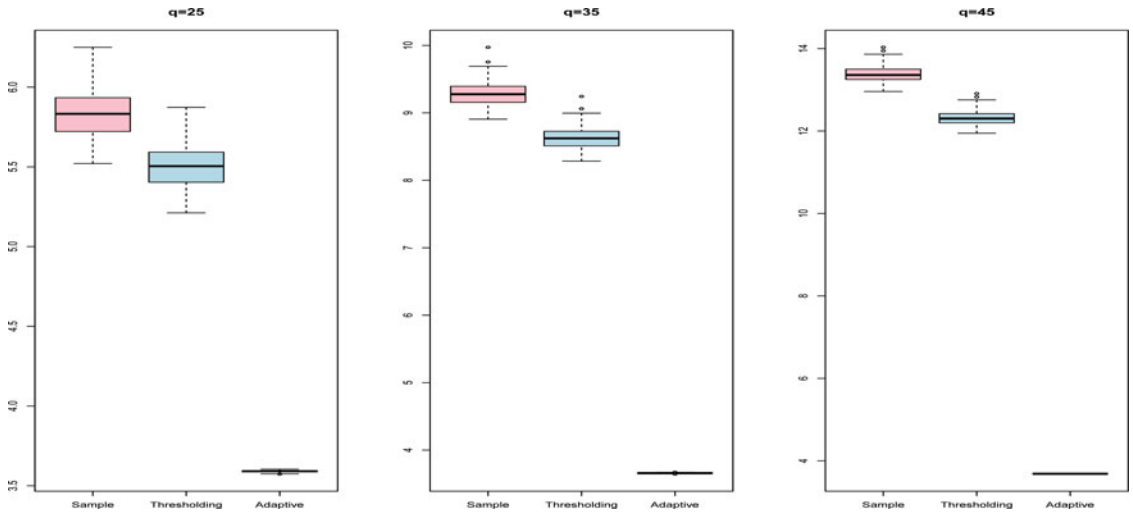


Figure 5. Comparison between different methods—each panel corresponds to a particular value of $q$. Reported here are the boxplots of the estimation error measured in operator norm for the sample covariance operator, thresholding, and adaptive block thresholding estimator with $\lambda_0 = 6$.

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 quantity $k(q)$ by

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

if the set on the right-hand side is nonempty, 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 article 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

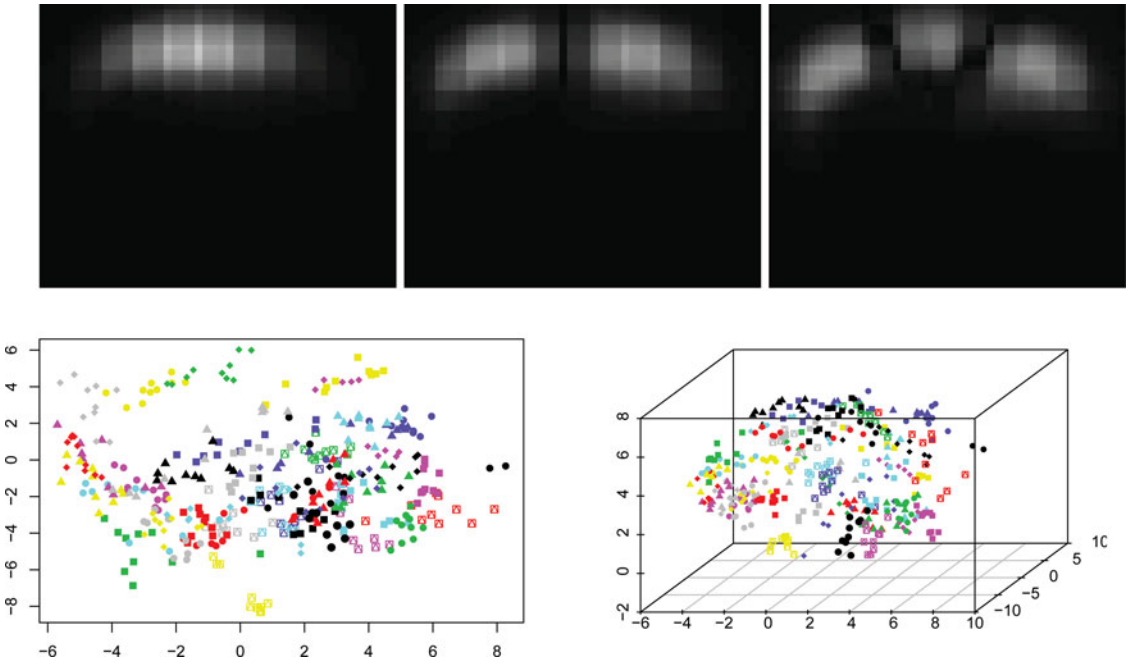

Figure 6. Estimated eigenimages—first three eigenimages corresponding to the adaptive block thresholding estimate, from left to right, are given in the top panel. The gray scale in each panel corresponds to the weight (absolute value) at each pixel, with the largest value represented by the brightest, and smallest value (0) represented by the darkest. The bottom panels gives the scores for each of the 400 images with images from the same subject plotted in the same color and symbol.

operator is important in spatial analysis. A feature of spatial data that is distinct from the setting of the present article is that the random variables are typically not observed on a regular lattice. For $r > 0$, define

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

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

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

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.

## APPENDIX

We present here the proofs to the main results.

### A.1 Proof of Theorem 2

We consider first the case of polynomially decaying covariances. Recall that $q_1 \le q_2 \le \cdots \le q_d$. Denote by

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

Recall that the lower bound for a special case yields a lower bound for the general case. It therefore suffices to show that

$$\inf_{\tilde\Sigma(\text{data})} \sup_{\Sigma\in\Theta_1} \mathbb{E}\|\tilde\Sigma - \Sigma\|^2 \ge \frac{C\log p}{n}, \tag{A.1}$$

and

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

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 $\Sigma_0$ be the identity operator, that is, $\sigma_0(s,t) = \delta_{st}$ where $\delta$ is the Kronecker's delta. Denote by

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

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 $\Sigma$. It is clear that for any $\Sigma \ne \Sigma_0 \in \Theta_1$, 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 $\log(1+x) \ge x - x^2/2$ for any $x \ge 0$. Therefore,

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

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

To prove (A.2), we consider separately the cases when (a) $k^* = 0$; (b) $k^* = d$; and (c) $1 \le k^* < d$. In each case, we appeal to the Varshamov–Gilbert bound (see, e.g., Tsybakov 2009) to construct $\Theta_2$. Consider first the case when $k^* = 0$. Simple calculation indicates that in this case,

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

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

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 $k^d \ge 8$, there exist a subset $\Omega := \{\omega_1, \ldots, \omega_N\}$ of $\{0,1\}^{\mathcal{G}(k,\ldots,k)}$ obeying $N \ge 2^{k^d/8}$ and

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

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

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

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

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

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 (A.2) 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,

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

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 $\Theta_2$ be defined in a similar fashion as before except that now $\omega_j$ 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,

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

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

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

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

$$\inf_{\tilde{\Sigma}(\text{data})} \sup_{\Sigma \in \Theta_2} \mathbb{E}\|\tilde{\Sigma} - \Sigma\|^2 \geq \frac{Cp}{n}. \tag{A.3}$$

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

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

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

We need to modify the construction of $\Theta_2$. 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) $N \geq 2^{q_1\cdots q_{k^*}k^{d-k^*}/8}$;
(c) for any $0 \leq j \neq j' \leq N$, $\|\omega_{j'} - \omega_j\|_1 \geq q_1 \ldots q_{k^*}k^{d-k^*}/8$, $\forall 0 \leq j \neq j' \leq N$, where $\omega_0 = 0$.

Take

$$k = \left\lceil \left(n/\prod_{l=0}^{k^*} q_l\right)^{\frac{1}{2\alpha+d-k^*}} \right\rceil. \tag{A.4}$$

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

$$\inf_{\tilde{\Sigma}(\text{data})} \sup_{\Sigma \in \Theta_2} \mathbb{E}\|\tilde{\Sigma} - \Sigma\|^2 \geq C\left(n^{-1}\prod_{l=0}^{k^*} q_l\right)^{\frac{2\alpha}{2\alpha+d-k^*}}. \tag{A.5}$$

The lower bound (8) for estimating $\Sigma \in \mathcal{F}_d(\alpha; M_0, M)$ then follows from (A.1) and (A.2).

The argument for exponentially decaying covariances is similar to that of the polynomially decaying ones. Let $q_* > 1$ be the solution to

$$\log n + d\log x = 2\alpha_0 x^\alpha. \tag{A.6}$$

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 $1 \leq k^* < d$.

## A.2 Proof of Theorem 3

The proof for $\mathcal{F}_d(\alpha; M_0, M)$ and $\mathcal{F}_d(\alpha_0, \alpha; M_0, M)$ is identical, and we shall focus on $\mathcal{F}_d(\alpha; M_0, M)$ for brevity.

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

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

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

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

Consequently

$$\begin{aligned}
\mathbb{E}\|\hat{\Sigma} - \Sigma\|^2 \leq\ & 2\cdot 3^{2d}\mathbb{E}\left(\max_{\|\mathbf{j}-\mathbf{j}'\|_\infty\leq 1} \|S_{\mathbf{j}\mathbf{j}'} - \Sigma_{\mathbf{j}\mathbf{j}'}\|\right)^2 \\
&+ 2M^2\left(n^{-1}\prod_{l=0}^{k^*} q_l\right)^{\frac{2\alpha}{2\alpha+d-k^*}}.
\end{aligned} \tag{A.7}$$

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

*Lemma A.1.* Let $I, J \subseteq \mathcal{G}_d$ with $\text{card}(I), \text{card}(J) \leq s$, then there exist constants $c_1, c_2 > 0$ such that

$$\mathbb{P}\{\|S_{I\times J} - \Sigma_{I\times J}\| \geq x\} \leq c_1 25^s \exp(-c_2 n\min\{x^2, x\}).$$

Recall that when $k^* = 0$,

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

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

$$N_1\cdots N_d \leq Cq_1\cdots q_d/b^d \leq Cpn^{-d/(2\alpha+d)}$$

for some constant $C > 0$. An application of Lemma A.1 and union bound yields

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

which implies that for any $x > 0$

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

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

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

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

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

for some constant $C > 0$. These two bounds together with (A.7) implies (12) 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 A.1 and union bound, we get

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

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

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

The claim (12) 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 A.1 and union bound, we now get

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

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

## A.3. Proof of Theorem 4

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

*A.3.1. Large blocks.* The following result is a consequence of the construction of $\mathcal{B}_d$ and the properties of $\mathcal{B}_1$.

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

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

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

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

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

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

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

provided that

$$2^{L-1}s_0 \ge (M/M_0)^{2/(2\alpha+d)} n^{1/(2\alpha+d)}. \tag{A.8}$$

Taking $\lambda_0 \ge 4C$ 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

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

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

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

Thus,

$$\mathbb{E}\|(\hat{\Sigma} - \Sigma) \circ W_L\|^2 = O(n^{-1}). \tag{A.9}$$

*A.3.2. 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,

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

Therefore,

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

Observe that

$$\left\|(\hat{\Sigma}-\Sigma)\circ\sum_{l=1}^{L-1}W_l\right\| \le \sum_{l=1}^{L-1}\left\|(\hat{\Sigma}-\Sigma)\circ W_l\right\|$$
$$\le \sum_{l=1}^{L-1}\left(\sum_{1\le k\le d}\left\|(\hat{\Sigma}-\Sigma)\circ W_{l,k}\right\|+\cdots+\left\|(\hat{\Sigma}-\Sigma)\circ W_{l,1\cdots d}\right\|\right),$$

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

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

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,

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

In other words,

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

Similarly, it can be shown that

$$\left\|(\hat{\Sigma}-\Sigma)\circ W_{l,k_1k_2}\right\|$$
$$\le\sup_{B\in\mathcal{A}'_{k_1k_2}(l)}\|\hat{\Sigma}_B-\Sigma_B\|\sum_{\substack{B_1,\ldots,B_d\in\tilde{\mathcal{B}}_1(l)\\ B_{k_1},B_{k_2}\in\mathcal{B}_1(l)}}\|u_{I_1\times\cdots\times I_d}\|^2$$

and so on. As a result,

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

We now appeal to the following result.

*Lemma A.3.* Let $u\in\ell_2(\mathcal{G}_d)$ such that $\|u\|=1$. Then

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

By Lemma A.3, 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 A.1, 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 $1-p^{-8}$,

$$\left\|(\hat{\Sigma}-\Sigma)\circ\sum_{l=1}^{L-1}W_l\right\|\le Cn^{-1/2}\sum_{l=1}^{L-1}2^{d(l-1)/2}s_0^{d/2}\le Cn^{-1/2}s_0^{d/2}2^{dL/2}.\tag{A.10}$$

*A.3.3. 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 blocks. By the bound (A.10) 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 A.4.* 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 A.4 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 blocks of size greater than $s_0$ will be treated as large blocks. In this case, following the small block bound (A.10), 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 A.4,

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

which implies that $\mathbb{E}\|(\hat{\Sigma}-\Sigma)\circ W_1\|^2=O\left(\frac{\log p}{n}\right)$. Together with (A.9), 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 (A.8), it follows from (A.10) that

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

Together with (A.9), this yields $\mathbb{E}\|\hat{\Sigma}-\Sigma\|^2=O(n^{-\frac{2\alpha}{2\alpha+d}})$.

*A.3.4. 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 $C > 0$.

## SUPPLEMENTARY MATERIALS

The supplementary materials for this article contain proofs for Lemmas 1, 3, 4, 5, and 6.

*[Received October 2010. Revised October 2011.]*

## REFERENCES


Bickel, P. and Levina, E. (2008a), "Regularized Estimation of Large Covariance Matrices," *The Annals of Statistics*, 36, 199–227. [255]

——— (2008b), "Covariance Regularization by Thresholding," *The Annals of Statistics*, 36, 2577–2604. [255]

Cai, T. T., and Liu, W. (2011), "Adaptive Thresholding for Sparse Covariance Matrix Estimation," *Journal of the American Statistical Association*, 494, 672–684. [255]

Cai, T. T., Liu, W., and Luo, X. (2011), "A Constrained $\ell_1$ Minimization Approach to Sparse Precision Matrix Estimation," *Journal of the American Statistical Association*, 494, 594–607. [255]

Cai, T. T., Liu, W., and Zhou, H. H. (2011), "Optimal Estimation of Large Sparse Precision Matrices," unpublished manuscript. [255]

Cai, T. T., and Yuan, M. (2012), "Adaptive Covariance Matrix Estimation Through Block Thresholding," *The Annals of Statistics*, 40, 2014–2042. [255,256,257]

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. [255]

Cai, T. T. and Zhou, H. (2012), "Optimal Rates of Convergence for Sparse Covariance Matrix Estimation," *The Annals of Statistics*, 40, 2389–2420. [255]

El Karoui, N. (2008), "Operator Norm Consistent Estimation of Large Dimensional Sparse Covariance Matrices," *The Annals of Statistics*, 36, 2717–2756. [255]

Fan, J., Fan, Y., and Lv, J. (2008), "High Dimensional Covariance Matrix Estimation Using a Factor Model" *Journal of Econometrics*, 147, 186–197. [255]

Friedman, J., Hastie, T., and Tibshirani, T. (2008), "Sparse Inverse Covariance Estimation With the Graphical Lasso," *Biostatistics*, 9, 432–441. [255]

Huang, J., Liu, N., Pourahmadi, M., and Liu, L. (2006), "Covariance Matrix Selection and Estimation via Penalised Normal Likelihood," *Biometrika*, 93, 85–98. [254]

Johnstone, I., and Lu, A. (2009), "On Consistency and Sparsity for Principal Components Analysis in High Dimensions," *Journal of the American Statistical Association*, 104, 682–693. [253]

Krause, E. (1987), *Taxicab Geometry*. New York: Dover. [253]

Lam, C. and Fan, J. (2009), "Sparsistency and Rates of Convergence in Large Covariance Matrix Estimation," *The Annals of Statistics*, 37, 4254–4278. [255]

Ledoit, O. and Wolf, M. (2004), "A Well-Conditioned Estimator for Large-dimensional Covariance Matrices," *Journal of Multivariate Analysis*, 88, 365–411. [254]

Rothman, A., Bickel, P., Levina, E., and Zhu, J. (2008), "Sparse Permutation Invariant Covariance Estimation," *Electronic Journal of Statistics*, 2, 494–515. [255]

Rothman, A., Levina, E., and Zhu, J. (2009), "Generalized Thresholding of Large Covariance Matrices," *Journal of the American Statistical Association*, 104, 177–186. [255]

Rudelson, M. (1999), "Random Vectors in the Isotropic Position,," *Journal of Functional Analysis*, 164, 60–72. [253]

Samaria, F., and Harter, A. (1994), "Parameterisation of a Stochastic Model for Human Face Identification," in *Proceedings of 2nd IEEE Workshop on Applications of Computer Vision*, Sarasota, FL, pp. 138–142. [259]

Sirovich, L., and Kirby, M. (1987), "Low-Dimensional Procedure for the Characterization of Human Faces," *Journal of the Optical Society of America A*, 4, 519–524. [253,258]

Tsybakov, A. (2009), *Introduction to Nonparametric Estimation*, New York: Springer. [254,261]

Turk, M., and Pentland, A. (1991), "Eigenfaces for Recognition," *Journal of Cognitive Neuroscience*, 3, 71–86. [253,258]

Yuan, M. (2010), "Sparse Inverse Covariance Matrix Estimation Via Linear Programming," *Journal of Machine Learning Research*, 11, 2261–2286. [255]

Yuan, M., and Lin, Y. (2007), "Model Selection and Estimation in the Gaussian Graphical Model," *Biometrika*, 94, 19–35. [255]