Submitted to the Annals of Statistics

LOCALIZATION ESTIMATOR FOR HIGH-DIMENSIONAL TENSOR COVARIANCE MATRICES

BY HAO-XUAN SUN[1,a], SONG XI CHEN[2,b] AND YUMOU QIU[3,c]

[1]*School of Mathematics and International Center for Interdisciplinary Statistics, Harbin Institute of Technology,* [a]*hxsun@hit.edu.cn*

[2]*Department of Statistics and Data Science, Tsinghua University,* [b]*songxichen@pku.edu.cn*

[3]*School of Mathematical Sciences and Center for Statistical Science, Peking University,* [c]*qiuyumou@math.pku.edu.cn*

This paper considers covariance matrix estimation for tensor data under high dimensionality. A multi-bandable covariance class is established to accommodate the need for underlying covariance structures of multi-layer lattices and general covariance decay patterns. We propose a high-dimensional covariance localization estimator for tensor data, which regulates the sample covariance matrix through a localization function. The statistical properties of the proposed estimator are studied by deriving the minimax rates of convergence under the spectral and the Frobenius norms. Numerical experiments and real data analysis on ocean eddy data are carried out to illustrate the utility of the proposed method in practice.

1. Introduction. Estimation of covariance matrices is a basic task in statistics as the covariance matrices and their estimates play a key role in many multivariate statistical procedures. For the fixed-dimensional case, the estimation quality of the sample covariance matrix can be assured by the conventional multivariate analysis. However, for high-dimensional problems, the consistency of the sample covariance matrix is no longer guaranteed (Bai, Silverstein and Yin, 1988; Bai and Yin, 1993; Johnstone, 2001). The last two decades have seen consistent high-dimensional covariance estimators being proposed, which include the banding and thresholding estimators (Bickel and Levina, 2008a,b), the tapering estimator (Cai, Zhang and Zhou, 2010) and the block thresholding estimator (Cai and Yuan, 2012).

The banding and tapering estimators (Bickel and Levina, 2008a) are designed for covariances with the so-called bandable covariance structure for vector data, where the components of the data vector follow a natural ordering with respect to their covariances. The separably bandable covariance class adopted in Zhang, Shen and Kong (2023) is an extension of the bandable class for vector data to matrix data, which models the joint covariance matrix as a Kronecker product of two marginal covariance matrices that govern the decay of covariances in each direction of the high-dimensional observations. McCormack and Hoff (2025) established the estimation theory for general separable covariances under the Frobenius norm. Despite their success, the aforementioned covariance estimators encounter limitations for tensor data. First, covariance estimators for high-dimensional vector data may not be applicable to tensor data. Second, separably bandable or general separable covariance structures may be overly restrictive for dependence patterns embedded in tensor data. Furthermore, theoretical results for tensor covariance estimation under the spectral norm have not been well studied.

Tensor data, represented as multi-layer arrays, can exhibit dependence arising from multi-source drivers, where heterogeneity exists across different directions or sources of the tensor. In this paper, we consider the covariance estimation problem for high-dimensional tensor

MSC2020 subject classifications. Primary 62H12; secondary 62C20.

Keywords and phrases. Covariance matrix estimation, high dimensionality, localization, minimax rate, tensor data.

data within a general multi-bandable class, where covariances between two tensor entries decay as their distance increases. The covariance matrices are allowed to be nonseparable and the decay rates can be heterogeneous across different slices of the tensor. High-dimensional tensor data are increasingly collected in many scientific studies. In the following, we provide several motivating examples of high-dimensional tensor data that exhibit general bandable structures. Consistent covariance estimation is urgently needed for those applications.

- *Geophysical data assimilation.* Data assimilation aims to combine forecast state variables from a complex dynamic system with observations, which is the key method to produce high-quality reanalysis data and predictions for weather forecasting (Sun et al., 2024) or other geoscience problems. Typical data assimilation methods, like the Ensemble Kalman Filter (EnKF), rely on estimating the forecast error covariance matrix. Geophysical state variables are naturally indexed by longitude, latitude, and height or depth, which form a three-order tensor. Accurate estimation of the high-dimensional tensor covariance matrices of state variables is fundamental for large-scale high-resolution data assimilation.
- *Classification for medical imaging.* In medical imaging studies, a key task is to classify patients or tissue samples into different categories based on imaging features, for example, classification of neurological diseases using functional magnetic resonance images and identification of cell types and cellular states using spatial transcriptomics data (Haviv et al., 2025). Such data can be represented as a high-dimensional tensor indexed by spatial coordinates, imaging modality, and by genes for transcriptomics. Linear and quadratic discriminant analysis requires an accurate estimation of the within-class covariance.
- *Covariance in attention mechanism.* In transformer-based large language models, a multi-layer attention mechanism is a key model structure, where each layer is a computational block consisting of a self-attention module and a feed-forward network (Vaswani et al., 2017). There are multiple layers stacked on top of each other, which take text representations from the previous layer and refine them. The collection of attention weights leads to a tensor indexed by layers and query/key positions. High-dimensional tensor covariance estimation helps to understand how token information passes through the attention network. See details in Section S9.3 of the supplementary material (SM).

A common thread across these applications is that the observations are high-dimensional and possess an intrinsic multi-layer tensor structure. Moreover, the dependence among components of the tensor cannot be modeled merely along a single direction or driver, and the covariance matrix may not be separable; rather, their covariances are shaped jointly by the spatial, temporal, biological, or architectural directions or drivers via a distance measure, and thus can be modeled by a multiple-layer version of the bandable covariance structure. Motivated by these applications, we propose consistent high-dimensional covariance estimation under a general bandable class for tensor data. Our main contributions are as follows.

- We propose a multi-bandable covariance class for tensor data with a general covariance decay function, and establish the minimax convergence rate for tensor covariance estimation under the spectral norm and this general covariance class. Theoretical analysis of tensor covariance estimation presents challenges beyond those encountered for high-dimensional covariance estimation for vector data, owing to the tensor's multi-way dependence structure. To address these challenges, we construct a new least favorable prior tailored to tensor covariances. The minimax result under the Frobenius norm is established as well.
- We develop a localization covariance estimator with general localization functions and establish sharp convergence bounds that match the minimax rates under the multi-bandable covariance class. This establishes the optimality of the proposed localization covariance estimator. The proof of the spectral-norm upper bound requires new technical tools tailored to the multi-way structure of tensor data, including a tensor-specific vectorization scheme and a novel block-partitioning argument.

- As a notable special case, our optimal minimax theory yields a new result for high-dimensional covariance estimation for vector data in that the banding estimator of Bickel and Levina (2008a) attains the same minimax optimal convergence rate as the tapering estimator of Cai, Zhang and Zhou (2010). This finding enriches the existing theory for the banding estimator.
- We propose a data-driven adaptive localization estimator which extends the method used in Liu and Ren (2020) to the tensor data setting using the Goldenshluger–Lepski method. We prove that the resulting adaptive estimator attains the same spectral-norm convergence rate and remains the minimax rate optimal.

The rest of the paper is organized as follows. We introduce the covariance estimation setting for high-dimensional tensor data and a motivating example in Section 2. The multi-bandable covariance class is proposed in Section 3.1, followed by the high-dimensional covariance localization estimator in Section 3.2. The consistency and the minimax optimal convergence rate of the localization estimator are established in Section 4. A data-driven adaptive estimator is also proposed. Sections 5 and 6 report the simulation studies and real data analysis, respectively. The discussion and conclusion are provided in Section 7. The technical details, proofs and additional results of simulations and the case study are presented in the supplementary material (SM).

2. Preliminaries. Throughout this paper, we use italic letters a, b for scalars and bold Roman letters $\mathbf{a}, \mathbf{b}$ for vectors and matrices. We use $\mathbf{0}_p$ and $\mathbf{1}_p$ to denote the p-dimensional vectors with all elements being 0 or 1, respectively. We write $a \asymp b$ if there are positive constants C_1 and C_2 such that $C_1 \le a/b \le C_2$. For a vector $\mathbf{a}$, $\|\mathbf{a}\|$ denotes the Euclidean norm. For a matrix $\mathbf{M}$, $\|\mathbf{M}\|$ and $\|\mathbf{M}\|_F$ denote the spectral and the Frobenius norm, respectively.

As we aim at developing consistent covariance estimators for tensor data, the following notations are introduced for the multi-bandable covariance class. Specifically, tensor data in this work are assumed to be organized over a multi-layer lattice. Let $\mathbb{X} := \{X(\mathbf{s})\}_{\mathbf{s}\in\mathcal{S}_d(\mathbf{p})}$ be random tensors sampled from a d-order lattice

$$\mathcal{S}_d(\mathbf{p}) = \{1, 2, \ldots, p_1\} \times \{1, 2, \ldots, p_2\} \times \cdots \times \{1, 2, \ldots, p_d\},$$

where $\mathbf{p} = (p_1, \ldots, p_d)^\top$ are the dimensions of the tensor entries in d directions. We consider in this work a regime where d is fixed but $p_\ell \to \infty$ for all $\ell = 1, 2, \ldots, d$. Denote by $\{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_p\}$ an arbitrary arrangement of the total $p = \prod_{\ell=1}^d p_\ell$ entries in $\mathcal{S}_d(\mathbf{p})$, where $\mathbf{s}_i = (s_{i1}, \ldots, s_{id})^\top$ denotes the coordinate of the ith entry. Then, a vectorization of a random tensor $\mathbb{X}$ according to the arrangement $\{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_p\}$ is $\mathbf{X} = (X(\mathbf{s}_1), \ldots, X(\mathbf{s}_p))^\top$. The lattice $\mathcal{S}_d(\mathbf{p})$ is introduced to represent a general class of data that can be mapped into some ordered tensor elements. For example, $d = 1$ represents high-dimensional vector data, which is the case studied in Bickel and Levina (2008a). The setting where $d > 1$ generalizes to tensors of arbitrary order, which are also common cases in many scientific studies. For instance, the oceanic data example which will be presented shortly corresponds to $d = 3$.

Suppose $\mathbb{X}_1, \ldots, \mathbb{X}_n$ are independent and identically distributed copies of a random tensor $\mathbb{X}$. Let $\mathbf{X}_m = (X_m(\mathbf{s}_1), \ldots, X_m(\mathbf{s}_p))^\top$ denote the vectorization of $\mathbb{X}_m$, the mth observation of the random tensor, where $X_m(\mathbf{s}_i)$ is the entry at grid point $\mathbf{s}_i$. Let $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma} = [\sigma_{ij}]_{p\times p}$ be the mean vector and covariance matrix of the vectorized data $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$. We are interested in estimating the covariance matrix $\boldsymbol{\Sigma}$. A natural estimator of the covariance matrix $\boldsymbol{\Sigma}$ is the sample covariance matrix

$$\mathbf{S}_n = [\hat{\sigma}_{ij}]_{p\times p} = \frac{1}{n-1} \sum_{m=1}^{n} \big(\mathbf{X}_m - \bar{\mathbf{X}}\big)\big(\mathbf{X}_m - \bar{\mathbf{X}}\big)^\top, \tag{2.1}$$

where $\hat{\sigma}_{ij}$ denotes the (i,j) entry of the sample covariance matrix and $\bar{\mathbf{X}} = n^{-1}\sum_{m=1}^{n}\mathbf{X}_m$. For the "large p small n" setting, the sample covariance matrix is no longer consistent for $\mathbf{\Sigma}$ (Bai, Silverstein and Yin, 1988; Bai and Yin, 1993; Johnstone, 2001) and a common solution is to take into account the underlying covariance structure.

Previous studies have constructed several consistent covariance matrix estimators under certain covariance classes to address the high dimensionality. Bickel and Levina (2008a) considered a banding estimator $\mathcal{B}_k(\mathbf{S}_n)$ with banding width k, with the (i,j) entry being

$$[\mathcal{B}_k(\mathbf{S}_n)]_{ij} = \hat{\sigma}_{ij}\mathbb{I}\{|i-j| \le k\}. \tag{2.2}$$

The banding estimator is designed for a bandable covariance class for vector data

$$\mathcal{U}_1(\alpha,\epsilon,C) = \Big\{\mathbf{\Sigma} = [\sigma_{ij}]_{p\times p} : \text{(i) } \max_j \sum_{|i-j|>k} |\sigma_{ij}| \le Ck^{-\alpha} \text{ for all } k>0, \\ \text{(ii) } 0 \le \lambda_{\min}(\mathbf{\Sigma}) \le \lambda_{\max}(\mathbf{\Sigma}) \le \epsilon^{-1}\Big\}, \tag{2.3}$$

where α, C and ϵ are some positive constants and $\lambda_{\min}(\mathbf{\Sigma})$ and $\lambda_{\max}(\mathbf{\Sigma})$ denote the minimum and maximum eigenvalues of $\mathbf{\Sigma}$, respectively. A minimum eigenvalue condition is imposed for estimating the precision matrix. However, this condition is not necessary for covariance estimation. They established that by choosing $k \asymp (n^{-1}\log p)^{-1/(2\alpha+2)}$ the estimation error of the banding estimator satisfies

$$\|\mathcal{B}_k(\mathbf{S}_n) - \mathbf{\Sigma}\|^2 = O_P\{(n^{-1}\log p)^{2\alpha/(2\alpha+2)}\}. \tag{2.4}$$

Cai, Zhang and Zhou (2010) introduced a tapering estimator $\mathcal{T}_k(\mathbf{S}_n)$ whose (i,j) entry is

$$[\mathcal{T}_k(\mathbf{S}_n)]_{ij} = \hat{\sigma}_{ij}\varphi(|i-j|; k/2, k), \tag{2.5}$$

where $\varphi(z;k_a,k_b) = \big\{1 - \{z-k_a\}_+/(k_b-k_a)\big\}_+$ is a tapering function and $\{z\}_+ = \max\{z, 0\}$. The estimator was designed for both $\mathcal{U}_1(\alpha,\epsilon,C)$ and the following covariance class

$$\mathcal{U}_2(\alpha,\epsilon,C) = \Big\{\mathbf{\Sigma} = [\sigma_{ij}]_{p\times p} : \text{(i) } |\sigma_{ij}| \le C|i-j|^{-\alpha-1} \text{ for all } i \ne j, \\ \text{(ii) } 0 \le \lambda_{\min}(\mathbf{\Sigma}) \le \lambda_{\max}(\mathbf{\Sigma}) \le \epsilon^{-1}\Big\}, \tag{2.6}$$

which is more restrictive on the off-diagonal elements. It is shown that the tapering estimator $\mathcal{T}_k(\mathbf{S}_n)$ with $k \asymp \min\{n^{1/(2\alpha+1)}, p\}$ can achieve the minimax optimal rate of convergence

$$\inf_{\hat{\mathbf{\Sigma}}} \sup_{\mathbf{\Sigma}\in\mathcal{U}_1(\alpha,\epsilon,C)} \mathbb{E}\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|^2 \asymp \min\big\{n^{-\frac{2\alpha}{2\alpha+1}} + \log p/n, p/n\big\}. \tag{2.7}$$

From another point of view, both bandable classes $\mathcal{U}_1(\alpha,\epsilon,C)$ and $\mathcal{U}_2(\alpha,\epsilon,C)$ regulate the magnitude of the covariance elements in terms of $|i-j|$, which were largely designed for vector data with $d=1$. Zhang, Shen and Kong (2023) extended the bandable covariance classes $\mathcal{U}_1(\alpha,\epsilon,C)$ and $\mathcal{U}_2(\alpha,\epsilon,C)$ to matrix data ($d=2$) by assuming separability in the covariance between the two coordinate directions. Specifically, denote by $\otimes$ the Kronecker product and $\mathrm{vec}(\cdot)$ the vectorization operator that sequentially stacks the columns of a matrix into a vector. Suppose that the matrix random variable $\mathbf{X} \in \mathbb{R}^{p_1\times p_2}$ follows a matrix normal distribution such that $\mathrm{vec}(\mathbf{X}) \sim \mathcal{N}(\mathrm{vec}(\mathbf{M}), \mathbf{\Sigma}_2\otimes\mathbf{\Sigma}_1)$ for a mean matrix $\mathbf{M}\in\mathbb{R}^{p_1\times p_2}$ and a covariance matrix $\mathbf{\Sigma}$ being within the separably bandable covariance class

$$\mathcal{U}_{3,q}(\alpha_1,\alpha_2,\epsilon,C) = \big\{\mathbf{\Sigma} = \mathbf{\Sigma}_2\otimes\mathbf{\Sigma}_1 : \mathbf{\Sigma}_1 \in \mathcal{U}_q(\alpha_1,\epsilon,C), \mathbf{\Sigma}_2 \in \mathcal{U}_q(\alpha_2,\epsilon,C) \\ \text{and } \min\{\lambda_{\min}(\mathbf{\Sigma}_1), \lambda_{\min}(\mathbf{\Sigma}_2)\} > \epsilon\big\} \tag{2.8}$$

for $q = 1$ or 2, where $\lambda_{\min}(\mathbf{\Sigma}_1)$ and $\lambda_{\min}(\mathbf{\Sigma}_2)$ are the minimum eigenvalues of $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$, respectively. They proposed a separably banding or tapering estimator $\hat{\mathbf{\Sigma}}_2(k_2) \otimes \hat{\mathbf{\Sigma}}_1(k_1)$ with

$$\big(\hat{\mathbf{\Sigma}}_1(k_1), \hat{\mathbf{\Sigma}}_2(k_2)\big) = \underset{\mathbf{\Sigma}_1, \mathbf{\Sigma}_2}{\arg\min} \|\tilde{\mathbf{\Sigma}}(k_1, k_2) - \mathbf{\Sigma}_2 \otimes \mathbf{\Sigma}_1\|_F^2, \tag{2.9}$$

where k_1 and k_2 are banding width parameters corresponding to the row and column directions of matrix data, $\tilde{\mathbf{\Sigma}}(k_1, k_2)$ is a doubly banding or tapering estimator that regularizes the sample covariance as $\mathbf{S}_n \circ \{\mathcal{B}_{k_2}(\mathbf{1}_{p_2}\mathbf{1}_{p_2}^\top) \otimes \mathcal{B}_{k_1}(\mathbf{1}_{p_1}\mathbf{1}_{p_1}^\top)\}$ or $\mathbf{S}_n \circ \{\mathcal{T}_{k_2}(\mathbf{1}_{p_2}\mathbf{1}_{p_2}^\top) \otimes \mathcal{T}_{k_1}(\mathbf{1}_{p_1}\mathbf{1}_{p_1}^\top)\}$, respectively, with $\circ$ being the elementwise product.

The aforementioned covariance classes can be too crude to reflect the underlying covariance structure encountered for general tensor data with $d \geq 3$. For data from a multi-order lattice with $d \geq 2$, the measurement $|i - j|$ in the bandable covariance classes $\mathcal{U}_1(\alpha, \epsilon, C)$ and $\mathcal{U}_2(\alpha, \epsilon, C)$ for vector data may be overly simple for tensor data. The separability assumption can be restrictive for situations where the detailed covariance information, especially the intricate cross-layer covariances, can be richer than what the separably bandable covariance class $\mathcal{U}_{3,q}(\alpha_1, \alpha_2, \epsilon, C)$ can offer. For example, Daley and Barker (2001) considered the nonseparable error correlation in an atmospheric variational data assimilation system, while Hakim (2005) revealed the nonseparable nature between the horizontal and vertical structures of forecast and analysis error covariance matrices for mid-latitude atmospheric data assimilation over the western north Pacific. Specific nonseparable spatio-temporal covariance structures can also be found in Cressie and Huang (1999) and Guttorp and Schmidt (2013).

Our motivating example is an ocean eddy dataset of salinity, which constitutes a three-order longitude-latitude-depth tensor. The data were the reanalysis salinity field associated with the eddy in the western Pacific, as shown in Figure 1a. We re-centered the eddy according to its daily center; see Section 6 for more description. The entries of the three-order tensor were vectorized with longitude as the fastest-varying index, followed by latitude, and depth as the slowest-varying index. Specifically, depth was ordered from shallow to deep. Within each depth level, grid points were ordered first by latitude from south to north, and within each latitude, by longitude from west to east. Figure 1b displays the sample correlation matrix of daily salinity changes in the target area, with two insets corresponding to the correlation at the sea surface and a zoomed-in region, respectively.

Although such a correlation matrix displayed a resemblance to the bandable structure defined in Bickel and Levina (2008a), a closer examination in Section 4 reveals its deviation from the bandable class $\mathcal{U}_1(\alpha, \epsilon, C)$ in (2.3). Indeed, it possesses a finer *bandable-in-bandable* form, which offers a richer covariance structure than that offered by the bandable class. Furthermore, the three-order ocean tensor data may not be adequately captured by the separably bandable covariance class in the spirit of $\mathcal{U}_{3,q}(\alpha_1, \alpha_2, \epsilon, C)$ for matrix data. These motivate us to study covariance estimation for high-dimensional data on d-order lattice $\mathcal{S}_d(\mathbf{p})$ with a multi-bandable structure.

3. Methodology.

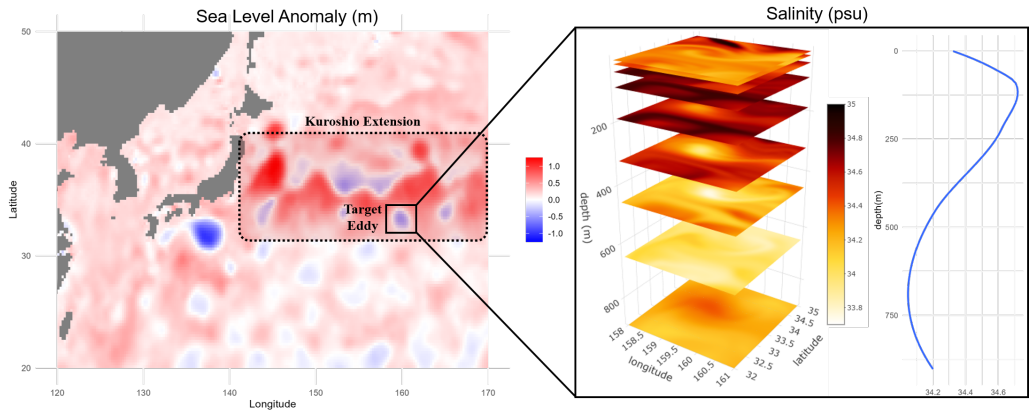
We introduce a multi-bandable covariance class with a general covariance decay function for tensor data, followed by a localization estimator for covariance matrix estimation.

3.1. *Multi-bandable covariance class.*

We start by introducing some notations to describe the relative position between any two grid points in a d-order lattice $\mathcal{S}_d(\mathbf{p})$, which aims to extend the exclusion $|i - j| > k$ used in the bandable covariance class $\mathcal{U}_1(\alpha, \epsilon, C)$. Specifically, for tensor data $\mathbb{X}$ sampled from $\mathcal{S}_d(\mathbf{p})$ with the vectorizations being $\mathbf{X} = (X(\mathbf{s}_1), X(\mathbf{s}_2), \ldots, X(\mathbf{s}_p))^\top$, we define an absolute difference between two coordinates $\mathbf{s}_i$ and $\mathbf{s}_j$ as

$$\boldsymbol{\delta}_{ij} := (\delta_{ij1}, \delta_{ij2}, \ldots, \delta_{ijd})^\top := (|s_{i1} - s_{j1}|, |s_{i2} - s_{j2}|, \ldots, |s_{id} - s_{jd}|)^\top. \tag{3.1}$$

(a) Sea level anomaly (left) and salinity field (right) of an eddy



(b) Correlation matrix and insets of daily salinity changes

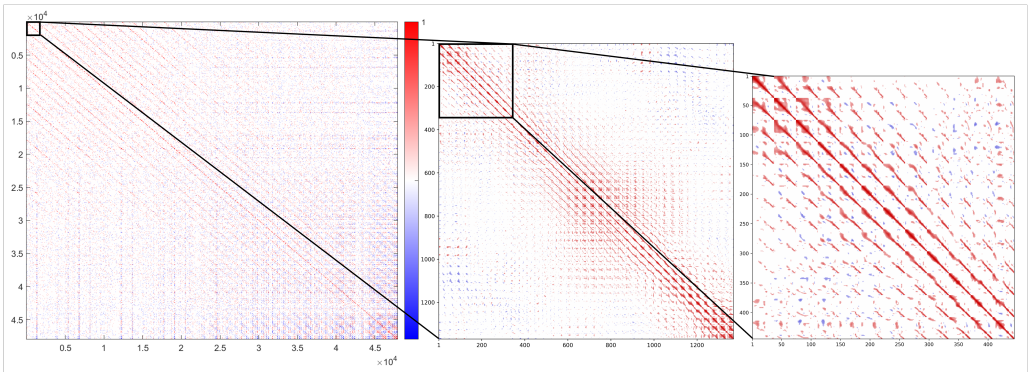


Fig 1: (a) Sea level anomaly on September 1, 2024 that displays an ocean eddy (left), and the salinity field in the practical salinity unit (psu) in 10 depth layers and the average salinity as a function of the depth (right); (b) the sample correlation matrix of daily salinity changes among all tensor entries (left), the entries of the first depth layer (middle) and a zoomed-in region spanning 32°N to 33°N and 158°E to 161°E (right).

Such a definition allows for different scales and dimensions p_ℓ in different coordinate directions of $\mathcal{S}_d(\mathbf{p})$. Specifically, the distance between the two coordinates $\mathbf{s}_i$ and $\mathbf{s}_j$ can be represented by some function of $\boldsymbol{\delta}_{ij}$, for example, by $\|\boldsymbol{\delta}_{ij}\|$ for the L_2 distance.

To impose restrictions on the covariance between two grid points that are separated by some d-dimensional vector $\mathbf{k} = (k_1, k_2, \ldots, k_d)$ in the lattice $\mathcal{S}_d(\mathbf{p})$, we define two sets

$$\mathcal{H}_d(\mathbf{k}) = \{\mathbf{k}' = (k'_1, k'_2, \ldots, k'_d) : 1 \le k'_\ell \le k_\ell \text{ for all } \ell = 1, 2, \ldots, d\} \text{ and}$$

$$\mathcal{H}^c_d(\mathbf{k}) = \{\mathbf{k}' = (k'_1, k'_2, \ldots, k'_d) : k'_\ell > k_\ell \text{ for at least one } \ell = 1, 2, \ldots, d\}$$

as a d-order hyper-rectangle and a hyper-rectangle-excluded region with respect to the scaling vector $\mathbf{k}$, respectively. The set $\mathcal{H}_d(\mathbf{k})$ is called the preserved $\mathbf{k}$-zone and it contains all the non-zero absolute coordinate differences $\boldsymbol{\delta}_{ij}$ with $\delta_{ij\ell}$ less than k_ℓ for all ℓ, whereas $\mathcal{H}^c_d(\mathbf{k})$ collects the excluded $\boldsymbol{\delta}_{ij}$ as they exceed the prescribed range by $\mathbf{k}$ in at least one direction. It is worth noting that $\mathcal{H}^c_d(\mathbf{k})$ does not denote the complement set of $\mathcal{H}_d(\mathbf{k})$ as it does not include all zeros. When $d = 1$, this reduces to the usual exclusion $|i - j| > k$ in the bandable

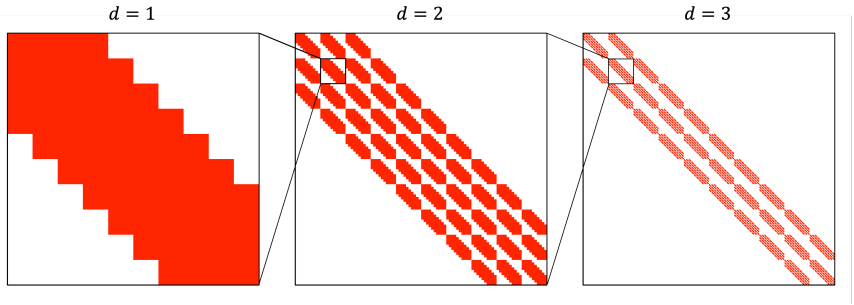


Fig 2: Illustration for pairs of variables (red) with distance within the preserved $\mathbf{k}$-zones $\mathcal{H}_d(k)$ for one-order lattice $\mathcal{S}_1(10)$ (left), two-order lattice $\mathcal{S}_2(10,10)$ (middle) and three-order lattice $\mathcal{S}_3(10,10,10)$ (right) under a column-major vectorization scheme.

covariance class $\mathcal{U}_1(\alpha,\epsilon,C)$. Figure 2 illustrates pairs of variables with distances within the preserved $\mathbf{k}$-zones for $d=1$, 2 and 3 in the column-major vectorization, which shows a "bandable-in-bandable" structure.

We consider a more general form of the upper bound than $Ck^{-\alpha}$ used in Condition (i) of the bandable covariance class $\mathcal{U}_1(\alpha,\epsilon,C)$ in (2.3). Specifically, we define a general d-variate *covariance decay function* $\tau(\mathbf{k})$, which offers patterns of covariance decay and is regulated by the following assumption.

ASSUMPTION 1. *The covariance decay function $\tau(\mathbf{k})$ is a d-variate function defined on $\mathcal{H}_d(\mathbf{p})$ and is non-negative, uniformly bounded, marginally non-increasing with respect to each direction of the lattice and satisfies $\tau(\mathbf{k})\to 0$ if and only if $\min_\ell k_\ell \to \infty$.*

Assumption 1 is a condition on the non-increasing covariance of the two entries as they become gradually apart. The covariance decay function $\tau(\mathbf{k})$ includes the polynomial decay

$$\tau(\mathbf{k}) = C\sum_{\ell=1}^{d} k_\ell^{-\alpha_\ell}\mathbb{I}\{1\le k_\ell < p_\ell\} \tag{3.2}$$

for $\{\alpha_\ell\}_{\ell=1}^d$ and C being positive numbers, and the exponential decay

$$\tau(\mathbf{k}) = C\sum_{\ell=1}^{d} \beta_\ell^{-k_\ell}\mathbb{I}\{1\le k_\ell < p_\ell\} \tag{3.3}$$

for $\{\beta_\ell\}_{\ell=1}^d$ all larger than 1. These functions illustrate the manner in which the sum of covariances diminishes for all grid point pairs that are spatially separated by a d-order hyper-rectangular region $\mathcal{H}_d(\mathbf{p})$. The covariance decay function $\tau(\mathbf{k})$ allows discontinuity at the boundary $k_\ell = p_\ell$ for some ℓ. For example, for the bandable covariance class $\mathcal{U}_1(\alpha,\epsilon,C)$, $\tau(k) = Ck^{-\alpha}\mathbb{I}\{1\le k<p\}$ which yields $\tau(p)=0$. For (3.2) and (3.3), the term $\mathbb{I}\{1\le k_\ell < p_\ell\}$ in $\tau(\mathbf{k})$ reflects the fact that the ℓth direction no longer contributes to the covariance decay form when k_ℓ exceeds p_ℓ; see discussion in Section S7.1 of the SM for general cases.

With the above preparation, we propose a covariance class for tensor data. Specifically, given a covariance decay function $\tau(\mathbf{k})$ and a positive constant ϵ, we define a d-order *multi-bandable covariance class*

$$\mathcal{U}(d,\tau,\epsilon) = \Big\{\mathbf{\Sigma}=[\sigma_{ij}]_{p\times p}:\ \text{(i)}\ \max_j \sum_{\{i:\boldsymbol{\delta}_{ij}\in\mathcal{H}_d^c(\mathbf{k})\}} |\sigma_{ij}| \le \tau(\mathbf{k}) \text{ for all } \mathbf{k}\in\mathcal{H}_d(\mathbf{p}),$$

$$\text{(ii)}\ 0\le \lambda_{\min}(\mathbf{\Sigma}) \le \lambda_{\max}(\mathbf{\Sigma}) \le \epsilon^{-1}\Big\}. \tag{3.4}$$

The proposed covariance class regularizes the covariance by the preserved $\mathbf{k}$-zone $\mathcal{H}_d(\mathbf{k})$ and the covariance decay function $\tau(\mathbf{k})$ as illustrated by the following examples.

EXAMPLE 1 (Polynomially decayed covariances). *Consider tensor data* $\mathbb{X} = \{X(\mathbf{s}) : \mathbf{s} \in \mathcal{S}_d(\mathbf{p})\}$ *with covariance matrix* $\mathbf{\Sigma} = [\sigma_{ij}]_{p\times p}$ *where*

$$\sigma_{ij} = \prod_{\ell=1}^{d} (1 + |s_{i\ell} - s_{j\ell}|)^{-\alpha_\ell - 1}, \tag{3.5}$$

and $p = \prod_{\ell=1}^{d} p_\ell$ *and* $\{\alpha_\ell\}_{\ell=1}^{d}$ *are positive constants. The product form in* (3.5) *is chosen for clarity, while the definition* (3.4) *does not require separability. Then, for each scaling vector* $\mathbf{k} \in \mathcal{H}_d(\mathbf{p})$*, the condition* $\boldsymbol{\delta}_{ij} \in \mathcal{H}_d^c(\mathbf{k})$ *excludes entries with* $|s_{i\ell} - s_{j\ell}| > k_\ell$ *for at least one* ℓ *and thus naturally retains pairs of entries that are close in all the coordinate directions. One can verify that* $\mathbf{\Sigma} \in \mathcal{U}(d, \tau, \epsilon)$ *by taking* $\tau(\mathbf{k})$ *in* (3.2) *and* $\epsilon^{-1} = \prod_{\ell=1}^{d}(1 + 2\alpha_\ell^{-1})$.

EXAMPLE 2 (Exponentially decayed covariances). *For tensor* $\mathbb{X} = \{X(\mathbf{s}) : \mathbf{s} \in \mathcal{S}_d(\mathbf{p})\}$ *with covariance matrix* $\mathbf{\Sigma} = [\sigma_{ij}]_{p\times p}$ *such that* $\sigma_{ij} = \prod_{\ell=1}^{d} \beta_\ell^{-|s_{i\ell} - s_{j\ell}|}$ *for* $\beta_\ell > 1$ *for all* ℓ*. One can see that* $\mathbf{\Sigma} \in \mathcal{U}(d, \tau, \epsilon)$ *with* $\tau(\mathbf{k})$ *given in* (3.3) *and* $\epsilon^{-1} = \prod_{\ell=1}^{d}\{(\beta_\ell + 1)/(\beta_\ell - 1)\}$.

It is worth noting that $\mathcal{U}(1, \tau, \epsilon)$ contains the bandable covariance class $\mathcal{U}_1(\alpha, \epsilon, C)$ in (2.3), with the polynomial decay $Ck^{-\alpha}$ in Condition (i) replaced with a general covariance decay function $\tau(\mathbf{k})$. In addition, the separably bandable covariance class $\mathcal{U}_{3,1}(\alpha_1, \alpha_2, \epsilon, C)$ in (2.8) is a special case of $\mathcal{U}(2, \tau, \epsilon)$, as it prescribes the polynomial decay only and assumes additional separability in the columns and rows of matrix data. In contrast, the proposed covariance class for $d = 2$ permits general covariance decay patterns and prescribes the bandable-in-bandable covariance structure in matrix data without the separability assumption.

3.2. *Localization estimator.* Let $\mathbf{S}_n = [\hat{\sigma}_{ij}]_{p\times p}$ be the sample covariance and $\mathbf{a}/\mathbf{b}$ denote the elementwise division of two vectors $\mathbf{a}$ and $\mathbf{b}$. We propose a *localization* estimator as

$$\mathcal{L}_h(\mathbf{S}_n; \mathbf{k}_h) = \left[\hat{\sigma}_{ij} h(\boldsymbol{\delta}_{ij}/\mathbf{k}_h)\right]_{p\times p}, \tag{3.6}$$

where $h(\boldsymbol{\delta}_{ij}/\mathbf{k}_h)$ is a weight taking values in $[0, 1]$ for the (i, j)-entry and is determined by the absolute coordinate difference $\boldsymbol{\delta}_{ij}$, a vector of scaling parameters $\mathbf{k}_h = (k_{h1}, k_{h2}, \ldots, k_{hd})^\top$ and a d-variate *localization function* h.

The estimator $\mathcal{L}_h(\mathbf{S}_n; \mathbf{k}_h)$ is motivated by the localization technique in the data assimilation area, where, to cope with spurious correlations, only the covariance in a local area of each assimilated grid point is preserved (Hamill, Whitaker and Snyder, 2001; Furrer and Bengtsson, 2007). Drawing on this, the proposed estimator provides regularization on the underlying covariance structure through two ingredients, the localization function h and a scaling vector $\mathbf{k}_h$. The localization function h is usually chosen with the goal of accounting for the diminishing covariance of two tensor entries as they are gradually separated.

ASSUMPTION 2. *(i) For* $\mathbf{z} = (z_1, z_2, \ldots, z_d)$ *with* $z_\ell \geq 0$ *for* $\ell = 1, 2, \ldots, d$*, the* d*-variate localization function* $h(\mathbf{z})$ *satisfies* $h(\mathbf{0}_d) = 1$*,* $h(\mathbf{z}) = 0$ *if* $z_\ell \geq 1$ *for an* $\ell \in \{1, 2, \ldots, d\}$*, and* $h(\mathbf{z})$ *is marginally non-increasing with respect to each* z_ℓ*. (ii) There exist constants* $c_1, c_2, \ldots, c_d \in (0, 1)$ *such that* $h(\mathbf{z}) = 1$ *for* $0 \leq z_\ell \leq c_\ell$ *and all* $\ell = 1, 2, \ldots, d$.

Assumption 2 (i) prescribes that h should be non-negative, compactly supported on $[0, 1]^d$ and marginally non-increasing. Part (ii) is to control the bias of using the localization function

h, which ensures that $\hat{\sigma}_{ij}$ would be fully preserved when the standardized distance between the ith and jth entries in the ℓth direction is no larger than c_ℓ.

To account for the variation of the estimation error under the spectral norm, we require another restriction on the variation of the localization function h, which requires the notion of Vitali variation. The Vitali variation (Vitali, 1908) is a generalization of the total variation to multi-dimensional spaces. Specifically, given two distinct points $\mathbf{a} = (a_1, a_2, \ldots, a_d)^\top$ and $\mathbf{b} = (b_1, b_2, \ldots, b_d)^\top$ that satisfy $a_\ell \le b_\ell$ for $\ell = 1, 2, \ldots, d$, suppose $[\mathbf{a}, \mathbf{b}] = [a_1, b_1] \times [a_2, b_2] \times \cdots \times [a_d, b_d]$, we define the quasi-volume

$$\text{qVol}(h; [\mathbf{a}, \mathbf{b}]) = \sum_{j_1=0}^{J_1} \cdots \sum_{j_d=0}^{J_d} (-1)^{j_1+\cdots+j_d} h\{b_1 + j_1(a_1 - b_1), \ldots, b_d + j_d(a_d - b_d)\}, \tag{3.7}$$

where $J_\ell = \mathbb{I}\{a_\ell \neq b_\ell\}$ for each $\ell = 1, 2, \ldots, d$. Then, given d univariate partitions $0 = a_{\ell,0} < a_{\ell,1} < \cdots < a_{\ell,N_\ell} = 1$ for $\ell = 1, 2, \ldots, d$ and some positive integers N_ℓ, let $\mathcal{P}$ be a collection of all sets of the form $\mathcal{A} = A_1 \times A_2 \times \cdots \times A_d$ where $A_\ell = [a_{\ell,n_\ell}, a_{\ell,n_\ell+1}]$ for each $\ell = 1, 2, \ldots, d$ and $0 \le n_\ell \le N_\ell - 1$, the Vitali variation of $h(\mathbf{z})$ is defined as

$$\text{ViV}(h) := \sup_{\mathcal{P}} \sum_{\mathcal{A} \in \mathcal{P}} |\text{qVol}(h; \mathcal{A})|. \tag{3.8}$$

In particular, if h is d-times continuously differentiable on $[0, 1]^d$, namely, the mixed partial derivative $\partial^d h / (\partial z_1 \partial z_2 \cdots \partial z_d)$ exists and is continuous, then

$$\text{ViV}(h) = \int_0^1 \int_0^1 \cdots \int_0^1 \Big| \frac{\partial^d h(\mathbf{z})}{\partial z_1 \partial z_2 \ldots \partial z_d} \Big| \mathrm{d}z_1 \mathrm{d}z_2 \ldots \mathrm{d}z_d.$$

Note that this simplified version of Vitali variation does not work for discontinuous localization functions. For instance, the commonly used doubly banding function $\prod_{\ell=1}^2 \mathbb{I}\{z_\ell < 1\}$ is discontinuous along z_1 or $z_2 = 1$, while its Vitali variation is defined by (3.8).

ASSUMPTION 3. *The localization function h satisfies ViV$(h) < C$ for a constant $C > 0$.*

Assumption 3 is valid for a wide range of functions. Two sufficient conditions for the bounded Vitali variation are as follows. One is that $h(\mathbf{z})$ is continuous in $[0, 1)^d$ except for some isolated discontinuous points and there exists an axis-aligned rectangular partition of $[0, 1]^d$ denoted by $\{\mathcal{D}_b\}$ such that $h(\mathbf{z})$ is smooth in each $\mathcal{D}_b$ and satisfies $\int_{\mathbf{z} \in \mathcal{D}_b} \big| \frac{\partial^d h(\mathbf{z})}{\partial \mathbf{z}_1 \partial \mathbf{z}_2 \ldots \partial \mathbf{z}_d} \big| \mathrm{d}\mathbf{z} < C$ for some constant C. The other sufficient condition is that $h(\mathbf{z})$ is a multiplicative function $\prod_{\ell=1}^d h_\ell(z_\ell)$ with the total variation of each $h_\ell(z_\ell)$ being finite. One may refer to Fang, Guntuboyina and Sen (2021) for the detailed discussion.

For $d = 1$, the banding function (2.2) and the tapering function (2.5) are two special cases of the localization function h. For $d > 1$, the localization function h can extend beyond the banding and the tapering functions by adapting to tensor data in a multi-order lattice through the absolute coordinate difference $\boldsymbol{\delta}_{ij}$. For example, $\tilde{\boldsymbol{\Sigma}}(k_1, k_2)$, the separably banding or tapering estimator (2.9) with $d = 2$, effectively employs $h(z) = \prod_{\ell=1}^2 \mathbb{I}\{z_\ell < 1\}$ or $\prod_{\ell=1}^2 \varphi(z_\ell; 0.5, 1)$, respectively, with the local weight assigned to $\hat{\sigma}_{ij}$ obtained via letting $z_\ell = \delta_{ij\ell}/k_\ell$. One can see that Assumptions 2 and 3 hold for both functions. Furthermore, the two functions are specific examples of a class of multiplicative localization functions $h(\mathbf{z}) = \prod_{\ell=1}^d h_\ell(z_\ell)$ with each $h_\ell(z_\ell)$ satisfying Assumptions 2 and 3. The merit of such a design is that the heterogeneity across different directions of the lattice can be addressed by different covariance decay patterns offered by h_ℓ and k_ℓ, respectively.

Specifically, we define a *multi-banding* estimator associated with a set of banding widths or scaling vector $\mathbf{k} = (k_1, k_2, \ldots, k_d)^\top$ as

$$\hat{\mathbf{\Sigma}}_{\mathbf{k}} := \big[\hat{\sigma}_{ij}\mathbb{I}\big\{\delta_{ij\ell} < k_\ell \text{ for all } \ell = 1, 2, \ldots, d\big\}\big]_{p\times p}. \tag{3.9}$$

The multi-banding estimator is a specialized localization estimator where the localization function is $h(\mathbf{z}) = \prod_{\ell=1}^{d}\mathbb{I}\{z_\ell < 1\}$ and the scaling vector is $\mathbf{k}$, which is analogous to $\mathbf{k}_h$ in the general localization estimators (3.6).

In the data assimilation field, a commonly used localization function h is the Gaspari-Cohn (GC) function (Gaspari and Cohn, 1999)

$$\mathrm{GC}(z) = \begin{cases} 1 - \frac{5}{3}z^2 + \frac{5}{8}z^3 + \frac{1}{2}z^4 - \frac{1}{4}z^5, & 0 \le z \le 1; \\ -\frac{2}{3}z^{-1} + 4 - 5z + \frac{5}{3}z^2 + \frac{5}{8}z^3 - \frac{1}{2}z^4 + \frac{1}{12}z^5, & 1 < z \le 2; \\ 0, & z \ge 2. \end{cases} \tag{3.10}$$

An illustration of the multiplicative banding and tapering functions for $d = 2$, and the GC function for $d = 1$ and 2 is given in Figure S5 in Section S7.3 of the SM. In practice, the GC function can impose a weight $\mathrm{GC}(2\|\boldsymbol{\delta}_{ij}/\mathbf{k}_{\mathrm{GC}}\|)$ on the (i, j)th entry of the sample covariance $\mathbf{S}_n$, which decays with respect to the L_2 distance and is homogeneous among the d directions of the lattice after scaling by a scaling vector $\mathbf{k}_{\mathrm{GC}}$. Although the GC function does not satisfy Assumption 2 (ii), the consistency of the localization estimator using the GC function can be established by a separate bias analysis, which will be discussed in Section 4.2.

The scaling vector $\mathbf{k}_h$ determines the maximum allowable separation distance between two tensor entries beyond which their covariance is set to be zero. A choice of $\mathbf{k}_h$ is based on a cross-validation measure on the covariance estimation with respect to the scaling vectors via the data splitting in Bickel and Levina (2008a) and Zhang, Shen and Kong (2023), whose details are outlined in Section 5. For $d = 1$, Qiu and Chen (2015) provided a ratio consistent selection for the banding width of the banding or tapering estimator. For high-dimensional inverse covariance estimation, Liu and Ren (2020) selected the banding widths via a data-driven Lepski's method (Lepski, 1992) and showed that the resulting estimator attains the minimax optimal spectral-norm convergence rate.

It is noted that there is no guarantee that the localization estimator (3.6) is non-negative definite. One may address this issue via the spectral decomposition of the localization estimator $\mathcal{L}_h(\mathbf{S}_n; \mathbf{k}_h)$ and replace the negative eigenvalues with a small positive constant.

4. Theoretical results. We establish the consistency of the localization estimator under the spectral and the Frobenius norms, which requires the following assumption.

ASSUMPTION 4. *The vectorized data* $\mathbf{X}_1$ *follows a sub-Gaussian distribution such that*

$$\mathbb{P}\big\{\big|\mathbf{v}^\top(\mathbf{X}_1 - \mathbb{E}\mathbf{X}_1)\big| > t\big\} \le \exp(-\rho t^2/2) \text{ for all } t > 0 \text{ and } \|\mathbf{v}\| = 1 \tag{4.1}$$

for a constant $\rho > 0$.

4.1. *Covariance estimation under the spectral norm.* We first investigate the consistency of the localization estimator under the spectral norm. Specifically, we define $V(\mathbf{k}) = \prod_{\ell=1}^{d} k_\ell$ for a scaling vector $\mathbf{k}$ and denote by $\mathcal{I}_d(\mathbf{k})$ the collection of all the positive integer points in $\mathcal{H}_d(\mathbf{k})$. Let C be a general positive constant that may vary in different contexts. The following lemma represents the localization estimator $\mathcal{L}_h(\mathbf{S}_n; \mathbf{k}_h)$ by a set of the multi-banding estimators $\{\hat{\mathbf{\Sigma}}_{\mathbf{k}}\}_{\mathbf{k}}$.

LEMMA 1. *Under Assumptions 2 and 3, the localization estimator can be written as*

$$\mathcal{L}_h(\mathbf{S}_n;\mathbf{k}_h)=\sum_{\mathbf{k}\in\mathcal{I}_d(\mathbf{k}_h)} w(\mathbf{k})\hat{\mathbf{\Sigma}}_{\mathbf{k}}, \tag{4.2}$$

where $\{\hat{\mathbf{\Sigma}}_{\mathbf{k}}\}_{\mathbf{k}}$ *are the multi-banding estimators in* (3.9) *and* $\{w(\mathbf{k})\}_{\mathbf{k}}$ *are weights satisfying*

$$\begin{aligned} w(\mathbf{k}) &:= (-1)^d qVol\{h;[(\mathbf{k}-\mathbf{1}_d)/\mathbf{k}_h,\mathbf{k}/\mathbf{k}_h]\} \\ &= \sum_{\mathbf{r}=(r_1,r_2,\ldots,r_d)\in\{0,1\}^d} (-1)^{r_1+r_2+\cdots+r_d} h\{(\mathbf{k}+\mathbf{r}-\mathbf{1}_d)/\mathbf{k}_h\}. \end{aligned} \tag{4.3}$$

Lemma 1 demonstrates that any localization estimator can be decomposed into a weighted sum of a larger number of multi-banding estimators $\{\hat{\mathbf{\Sigma}}_{\mathbf{k}}\}_{\mathbf{k}}$, with the weights $\{w(\mathbf{k})\}_{\mathbf{k}}$ corresponding to the quasi-volume defined in (3.7). The decomposition serves as a new technique to analyze general localization estimators for tensor covariances. Once the estimation upper bound for the multi-banding estimator is derived, the upper bound of general localization estimators can be obtained by applying this lemma.

LEMMA 2. *Under Assumptions 1 and 4, for* $\log p = o(n)$ *and* $V(\mathbf{k}) = o(n)$*, the multi-banding estimator in* (3.9) *satisfies*

$$\sup_{\mathbf{\Sigma}\in\mathcal{U}(d,\tau,\epsilon)} \mathbb{E}\|\hat{\mathbf{\Sigma}}_{\mathbf{k}}-\mathbb{E}\hat{\mathbf{\Sigma}}_{\mathbf{k}}\|^2 \le C\frac{\log p+V(\mathbf{k})}{n}. \tag{4.4}$$

Lemma 2 establishes that $\mathbb{E}\|\hat{\mathbf{\Sigma}}_{\mathbf{k}}-\mathbb{E}\hat{\mathbf{\Sigma}}_{\mathbf{k}}\|^2$, as a measure of variation of the multi-banding estimator, is bounded by $\log p/n$ and $V(\mathbf{k})/n$ up to a constant factor. For $d=1$, the lemma gives the bound $(\log p+k)/n$ for the banding estimator $\mathcal{B}_k(\mathbf{S}_n)$, which improves the $k\log p/n$ established in Bickel and Levina (2008a). This sharper bound also does not follow directly from the proof of Cai, Zhang and Zhou (2010) as they decomposed the tapering estimator into an average of matrices formed by the sum of disjoint blocks and then controlled the variation of each block. The proof, when applied to the banding estimator, however, would introduce an additional factor of k in the upper bound, which prevented attaining the optimal rate of convergence $n^{-2\alpha/(2\alpha+1)}$ in (2.7) when $p>n^{1/(2\alpha+1)}$ for the banding estimator. Our proof avoids this loss for $d=1$ by partitioning the banding estimator into covariance and cross-covariance matrices of size $k\times k$, with the number of non-zero blocks in each row or column no larger than 3. This idea is then extended to tensor data, which leads to a new technical tool tailored to the multi-way structure of tensor data by a tensor-specific vectorization scheme and a novel block-partitioning argument. Consequently, the variation control of $\hat{\mathbf{\Sigma}}_{\mathbf{k}}$ is reduced to handle the random matrices of size $V(\mathbf{k})\times V(\mathbf{k})$, leading to the upper bound $C\{\log p+V(\mathbf{k})\}/n$.

The following theorem provides the consistency of the localization estimator.

THEOREM 1. *Under Assumptions 1, 2, 3 and 4, for* $\log p = o(n)$ *and* $V(\mathbf{k}_h) = o(n)$*, the localization estimator* (3.6) *satisfies*

$$\sup_{\mathbf{\Sigma}\in\mathcal{U}(d,\tau,\epsilon)} \mathbb{E}\big\|\mathcal{L}_h(\mathbf{S}_n;\mathbf{k}_h)-\mathbf{\Sigma}\big\|^2 \le C\tau^2(\mathbf{k}_h\circ\mathbf{c})+C\frac{\log p+V(\mathbf{k}_h)}{n}, \tag{4.5}$$

where $\mathbf{c}=(c_1,c_2,\ldots,c_d)^\top$ *is defined in Assumption 2 (ii). Specifically, the localization estimator with the scaling vector* $\mathbf{k}_h=\arg\min_{\mathbf{k}\in\mathcal{H}_d(\mathbf{p})}\{\tau^2(\mathbf{k})+V(\mathbf{k})/n\}$ *satisfies*

$$\sup_{\mathbf{\Sigma}\in\mathcal{U}(d,\tau,\epsilon)} \mathbb{E}\big\|\mathcal{L}_h(\mathbf{S}_n;\mathbf{k}_h)-\mathbf{\Sigma}\big\|^2 \le C\varepsilon_{n,\mathbf{p}}+C\frac{\log p}{n}, \tag{4.6}$$

where $\varepsilon_{n,\mathbf{p}}=\min_{\mathbf{k}\in\mathcal{H}_d(\mathbf{p})}\{\tau^2(\mathbf{k})+V(\mathbf{k})/n\}$.

Theorem 1 demonstrates the consistency of the localization estimator (3.6) for estimating the covariance matrix $\mathbf{\Sigma}$ if the dimension satisfies $\log p = o(n)$ and the scaling vector satisfies $V(\mathbf{k}_h) = o(n)$. Specifically, the estimation error of the localization estimator in (4.5) consists of two terms: $\tau^2(\mathbf{k}_h \circ \mathbf{c})$ and $n^{-1}\{\log p + V(\mathbf{k}_h)\}$. The $\tau^2(\mathbf{k}_h \circ \mathbf{c})$ term represents the bias introduced by the entries $\{\hat{\sigma}_{ij} h(\boldsymbol{\delta}_{ij}/\mathbf{k}_h)\}$ in $\mathcal{L}_h(\mathbf{S}_n; \mathbf{k}_h)$ due to the weight $h(\boldsymbol{\delta}_{ij}/\mathbf{k}_h)$ being less than 1, which converges to 0 according to Assumption 1 as long as the scaling parameters $k_{h\ell} \to \infty$ as the sample size n and dimensions $p_\ell \to \infty$. In addition, the latter term represents a form of variation under the spectral norm.

We now turn our attention to $\varepsilon_{n,\mathbf{p}}$ in (4.6), which provides a general form of the trade-off between the covariance decay function $\tau(\mathbf{k})$ and $V(\mathbf{k})/n$, the "volume" of the preserved $\mathbf{k}$-zone divided by n. Hence, $\varepsilon_{n,\mathbf{p}}$ can be guaranteed to diminish to 0 under Assumption 1 as the dimensions $\{p_\ell\}$ and the sample size n increase. Specifically, the rate $\varepsilon_{n,\mathbf{p}}$ can be attained by properly choosing the localization scaling vector $\mathbf{k}_h$ for the localization function h, which is determined by solving a bounded optimization problem. Specifically, if $n = o(p_\ell)$ for $\ell = 1, 2, \ldots, d$, the rate $\varepsilon_{n,\mathbf{p}}$ is only relevant to the sample size n.

When $d = 1$ and $\tau(k) = Ck^{-\alpha}\mathbb{I}\{1 \le k < p\}$,

$$\varepsilon_{n,\mathbf{p}} \asymp \min_{1 \le k \le p} \left(k^{-2\alpha}\mathbb{I}\{1 \le k < p\} + k/n\right) \asymp \min\{n^{-2\alpha/(2\alpha+1)}, p/n\} \tag{4.7}$$

by choosing $k_h \asymp \min\{n^{1/(2\alpha+1)}, p\}$. The result generalizes the convergence rate (2.7) in Cai, Zhang and Zhou (2010) to the localization estimators equipped with general localization functions that satisfy Assumptions 2 and 3, which covers the banding estimator $\mathcal{B}_k(\mathbf{S}_n)$ in Bickel and Levina (2008a) as a special case. Combined with the minimax lower bound established in Cai, Zhang and Zhou (2010), this implies that the rate given in (4.7) is actually the minimax optimal rate for the banding estimator under the spectral norm.

Another finding is that the banding estimator $\mathcal{B}_k(\mathbf{S}_n)$ of Bickel and Levina (2008a), which was designed for $d = 1$, may not guarantee consistency for $d \ge 2$. We illustrate this issue using the example in (3.5) with $d = 2$, where $\tau(\mathbf{k}) = C\sum_{\ell=1}^{2} k_\ell^{-\alpha_\ell}\mathbb{I}\{1 \le k_\ell < p_\ell\}$ for $\alpha_\ell > 0$ and $\ell = 1, 2$. Suppose the matrix data $\mathbf{X} = \{X(s_1, s_2) : (s_1, s_2) \in \mathcal{S}_2(p_1, p_2)\}$ is vectorized in the column-major order, namely, a grid point (s_1, s_2) is mapped to the $\{s_1 + (s_2 - 1)p_1\}$th dimension after vectorization. The heatmap of the corresponding covariance matrix is shown in Figure 3, which does not satisfy the bandable condition in (2.3) for vector data, since

$$\sum_{|i-j| \ge p_1} |\sigma_{ij}| \ge |\sigma_{i\,i+p_1}| \not\to 0 \ \text{ as } p_1 \to \infty, \tag{4.8}$$

for each given i. The reason for $|\sigma_{i\,i+p_1}| \not\to 0$ is that $\sigma_{i\,i+p_1}$ is the covariance between the variables at the locations $\mathbf{s}_i = (s_1, s_2)^\top$ and $\mathbf{s}_{i+p_1} = (s_1, s_2 + 1)^\top$, which are adjacent in a row of $\mathcal{S}_2(p_1, p_2)$. This shows that the covariance matrices in the multi-bandable class in (3.4) may not satisfy the bandable class (2.3), which implies that the banding estimator $\mathcal{B}_k(\mathbf{S}_n)$ of Bickel and Levina (2008a) may not be consistent for $\mathbf{\Sigma}$ under the multi-bandable class.

In fact, to make the mean squared error (MSE) of $\mathcal{B}_k(\mathbf{S}_n)$ converge to 0, it is required that both the bias term $\|\mathcal{B}_k(\mathbf{\Sigma}) - \mathbf{\Sigma}\|$ and the variation term $\mathbb{E}\|\mathcal{B}_k(\mathbf{S}_n) - \mathbb{E}\mathcal{B}_k(\mathbf{S}_n)\|^2$ converge to 0 as $n, p_1, p_2 \to \infty$. For the example above, the bias term satisfies $\|\mathcal{B}_k(\mathbf{\Sigma}) - \mathbf{\Sigma}\| \ge \max_i |\sigma_{1\,p_1+1}|$ for $k < p_1$, while the variation is bounded by $\mathbb{E}\|\mathcal{B}_k(\mathbf{S}_n) - \mathbb{E}\mathcal{B}_k(\mathbf{S}_n)\|^2 = O\big(\{k + \log(p_1 p_2)\}/n\big)$ from (S.13) in the SM. From the illustration in Figure 3, the banded area $|i - j| \le k$ with $k < p_1$ would exclude row-adjacent covariances $\{\sigma_{i\,i+p_1}\}$. This exclusion would lead to a non-ignorable bias of $\mathcal{B}_k(\mathbf{S}_n)$. However, choosing $k > p_1$ would make the variation not diminish if $p_1 > n$. Therefore, the MSE of $\mathcal{B}_k(\mathbf{S}_n)$ would not converge to 0 if $p_1 > n$.

Beyond the full covariance matrix, one may consider estimating the covariance matrix of a slice of tensor data obtained by fixing one coordinate of a tensor direction. For $d \ge 2$, given a

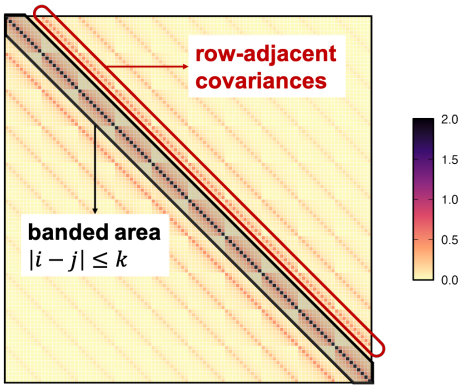


Fig 3: Heatmap of $\mathbf{\Sigma} = [\sigma_{ij}]_{p\times p} \in \mathcal{U}(2,\tau,\epsilon)$ for a 10×10 matrix data vectorized in the column-major order with the entries $\sigma_{ij} = \prod_{\ell=1}^{2}(1+\delta_{ij\ell})^{-\alpha_\ell - 1}$ for $\alpha_1 = 0.1$ and $\alpha_2 = 0.2$. The gray region is a banded area $|i-j| \le k$ for the banding estimator with a banding width $k < p_1$, and the red colored bands illustrate the covariances between adjacent grid points along rows, which decay rather slowly and prevent the bandable condition in (2.3).

fixed $s_d \in \{1,\dots,p_d\}$, we let $\mathbb{X}_{m,s_d}^{(-d)} = \{X_m(s_1,\dots,s_{d-1},s_d) : 1 \le s_1 \le p_1, \dots, 1 \le s_{d-1} \le p_{d-1}\}$ be the s_dth slice of the tensor $\mathbb{X}_m$ on the dth direction. Let $\mathbf{X}_{m,s_d}^{(-d)}$ be the vectorization of $\mathbb{X}_{m,s_d}^{(-d)}$ and $\mathbf{\Sigma}_{s_d}^{(-d)}$ be its covariance matrix. Denote by $\mathbf{S}_{n,s_d}^{(-d)}$ the corresponding sample covariance matrix. Similar to (3.6), we let $\mathcal{L}_{h^{(-d)}}\big(\mathbf{S}_{n,s_d}^{(-d)}; \mathbf{k}_h^{(-d)}\big)$ be the localization estimator of the sliced covariance matrix $\mathbf{\Sigma}_{s_d}^{(-d)}$, where $h^{(-d)}$ is a $(d-1)$-variate localization function and $\mathbf{k}_h^{(-d)} = (k_{h1},\dots,k_{h\,d-1})^\top$ is a $(d-1)$-dimensional scaling vector. The following corollary shows the convergence rate of estimating sliced covariances under the spectral norm.

COROLLARY 1. *Suppose the conditions of Theorem 1 hold for random tensor $\mathbb{X}_{m,s_d}^{(-d)}$ and localization function $h^{(-d)}$. For $d \ge 2$, let $\mathbf{p}_{-d} = (p_1,\dots,p_{d-1})^\top$ and $V(\mathbf{k}') = \prod_{\ell=1}^{d-1} k'_\ell$. The localization estimator with the scaling vector $\mathbf{k}_h^{(-d)} = \arg\min_{\mathbf{k}'\in\mathcal{H}_{d-1}(\mathbf{p}_{-d})}\{\tau_{-d}^2(\mathbf{k}') + V(\mathbf{k}')/n\}$ satisfies*

$$\sup_{s_d\in\{1,2,\dots,p_d\}} \sup_{\mathbf{\Sigma}\in\mathcal{U}(d,\tau,\epsilon)} \mathbb{E}\big\|\mathcal{L}_{h^{(-d)}}\big(\mathbf{S}_{n,s_d}^{(-d)}; \mathbf{k}_h^{(-d)}\big) - \mathbf{\Sigma}_{s_d}^{(-d)}\big\|^2 \le C\varepsilon_{n,\mathbf{p}_{-d}} + C\frac{\log(\prod_{\ell=1}^{d-1} p_\ell)}{n},$$

where $\varepsilon_{n,\mathbf{p}_{-d}} = \min_{\mathbf{k}'\in\mathcal{H}_{d-1}(\mathbf{p}_{-d})}\{\tau_{-d}^2(\mathbf{k}') + V(\mathbf{k}')/n\}$ and $\tau_{-d}(\mathbf{k}') = \tau\big(k'_1,\dots,k'_{d-1},p_d\big)$.

Corollary 1 shows that estimating a fixed-slice covariance matrix $\mathbf{\Sigma}_{s_d}^{(-d)}$ is a genuinely $(d-1)$-way tensor problem, mimicking the results for the localization estimator for the full covariance matrix of d-variate tensors. Consequently, whenever the dth direction contributes nontrivially to the full covariance estimation, the fixed-slice covariance estimator enjoys a faster convergence rate with $\varepsilon_{n,\mathbf{p}}$ reduced to $\varepsilon_{n,\mathbf{p}_{-d}}$ and $\log p_d/n$ removed.

The first term of the upper bound $\varepsilon_{n,\mathbf{p}}$ in (4.6) can have an explicit form in the following cases, where the specific details can be found in Section S7.1.

EXAMPLE 3 (Convergence rate under polynomially decayed covariances). *For the covariance decay function $\tau(\mathbf{k})$ in (3.2), the upper bound term $\varepsilon_{n,\mathbf{p}}$ can be obtained by solving*

$$\min_{\mathbf{k}\in\mathcal{H}_d(\mathbf{p})} \Big\{\Big(\sum_{\ell=1}^{d} k_\ell^{-\alpha_\ell}\mathbb{I}\{1 \le k_\ell < p_\ell\}\Big)^2 + n^{-1}\prod_{\ell=1}^{d} k_\ell\Big\} \tag{4.9}$$

within the closed set $\mathcal{H}_d(\mathbf{p})$. Then, if $p_\ell > n^{\alpha_\ell^{-1}(2+\sum_{\ell'=1}^d \alpha_{\ell'}^{-1})^{-1}}$ for $\ell = 1, 2, \ldots, d$, it follows that $\varepsilon_{n,\mathbf{p}} \asymp n^{-2(2+\sum_{\ell=1}^d \alpha_\ell^{-1})^{-1}}$ by choosing $k_{h\ell} \asymp n^{\alpha_\ell^{-1}(2+\sum_{\ell'=1}^d \alpha_{\ell'}^{-1})^{-1}}$, where each $\{\alpha_\ell\}$ jointly influences $\varepsilon_{n,\mathbf{p}}$ and the order of the optimal scaling vector $\mathbf{k}_h$. Otherwise, if $p_\ell \le n^{\alpha_\ell^{-1}(2+\sum_{\ell'=1}^d \alpha_{\ell'}^{-1})^{-1}}$ for an ℓ, it is sufficient to take $k_\ell = p_\ell$ and solve (4.9) for $\ell' \neq \ell$. A detailed discussion for the degenerate case of $d = 2$ is given in Section S7.1.3.

For the homogeneous case, when $\alpha_\ell = \alpha$ and $p_\ell > n^{1/(2\alpha+d)}$ for all ℓ, then $\varepsilon_{n,\mathbf{p}} \asymp n^{-2\alpha/(2\alpha+d)}$ and the upper bound $n^{-2\alpha/(2\alpha+d)} + \log p/n$ would involve d, the order of the lattice, through the first term. Then, when the dimension $p = \prod_{\ell=1}^d p_\ell$ is relatively small, say $\log p \le n^{d/(2\alpha+d)}$, $\varepsilon_{n,\mathbf{p}} \asymp n^{-2\alpha/(2\alpha+d)}$ becomes the leading term while p has no effect on the upper bound. In contrast, for large p that satisfies $\log p \ge n^{d/(2\alpha+d)}$, the leading term $\log p/n$ will be influenced by both the dimension p and the sample size n.

For estimating the fixed-slice covariance matrix $\mathbf{\Sigma}_{s_d}^{(-d)}$, the same argument yields that if $p_\ell > n^{\alpha_\ell^{-1}(2+\sum_{\ell'=1}^{d-1} \alpha_{\ell'}^{-1})^{-1}}$ for $\ell = 1, 2, \ldots, d-1$, then $\varepsilon_{n,\mathbf{p}_{-d}} \asymp n^{-2(2+\sum_{\ell=1}^{d-1} \alpha_\ell^{-1})^{-1}}$. In the homogeneous case, this gives the convergence rate $n^{-2\alpha/(2\alpha+d-1)} + \log(\prod_{\ell=1}^{d-1} p_\ell)/n$, which is faster than estimating the whole covariance matrix.

EXAMPLE 4 (Convergence rate under exponentially decayed covariances). *For the covariance decay function $\tau(\mathbf{k})$ in (3.3), the upper bound term $\varepsilon_{n,\mathbf{p}}$ can be obtained by solving*

$$\min_{\mathbf{k}\in\mathcal{H}_d(\mathbf{p})} \Big\{ \Big(\sum_{\ell=1}^d \beta_\ell^{-k_\ell} \mathbb{I}\{1 \le k_\ell < p_\ell\} \Big)^2 + n^{-1} \prod_{\ell=1}^d k_\ell \Big\}. \tag{4.10}$$

Specifically, if $\beta_\ell = \beta$ and $p_\ell > \log n$ for all ℓ, then minimizing (4.10) suggests choosing $k_{h\ell} \asymp \log n$ for all $\ell = 1, 2, \ldots, d$, which leads to $\varepsilon_{n,\mathbf{p}} = O\{(\log n)^d/n\}$. For estimating the fixed-slice covariance matrix $\mathbf{\Sigma}_{s_d}^{(-d)}$, this reduces to $\varepsilon_{n,\mathbf{p}_{-d}} = O\{(\log n)^{d-1}/n\}$. In particular, for $d = 1$, $\varepsilon_{n,\mathbf{p}} \asymp \min\{\log n/n, p/n\}$ by choosing $k_h \asymp \min\{\log n, p\}$. One can see that the exponentially decayed $\tau(\mathbf{k})$ offers a much faster covariance decay rate than the polynomially decayed case, and thus requires a more restrictive regularization with the localization scaling parameters of a smaller order.

REMARK 1. *The results in Theorem 1 can be extended to data observed on an irregular subset of a regular lattice. Denote by $\tilde{\mathcal{S}} = \{\tilde{\mathbf{s}}_1, \tilde{\mathbf{s}}_2, \ldots, \tilde{\mathbf{s}}_{p'}\}$ an irregular lattice containing p' arbitrarily arranged d-variate coordinates. Suppose $\tilde{\mathcal{S}} \subseteq \mathcal{S}_d(\mathbf{p})$ with $\mathbf{p} = (p_1, p_2, \ldots, p_d)^\top$. Under conditions in Theorem 1, the localization estimator (3.6) is consistent and retains the same spectral-norm convergence rate in (4.6) at a conservative cost of raising the dimension p' to $p = \prod_{\ell=1}^d p_\ell$, the number of entries in the ambient lattice $\mathcal{S}_d(\mathbf{p})$. Such irregular subsets commonly arise in scientific applications, for example, land masks and seafloor topography may break a regularly shaped lattice in oceanography (Moore et al., 2011).*

We next study the minimax properties of the covariance estimation of tensor data within specific covariance classes. The following theorem demonstrates that the minimax upper bound under the spectral norm in (4.6) cannot be further improved for the heterogeneous covariance decay function $\tau(k) = C \sum_{\ell=1}^d k_\ell^{-\alpha_\ell} \mathbb{I}\{1 \le k_\ell < p_\ell\}$ in (3.2).

THEOREM 2. *For Gaussian-distributed $\mathbf{X}_1$, under Assumption 1, for covariance decay function $\tau(\mathbf{k})$ in (3.2), if $\log p = o(n)$ and $p_\ell > n^{\alpha_\ell^{-1}(2+\sum_{\ell'=1}^d \alpha_{\ell'}^{-1})^{-1}}$ for $\ell = 1, 2, \ldots, d$, the minimax risk of estimating the covariance matrix $\mathbf{\Sigma} \in \mathcal{U}(d, \tau, \epsilon)$ satisfies*

$$\inf_{\hat{\mathbf{\Sigma}}} \sup_{\mathbf{\Sigma}\in\mathcal{U}(d,\tau,\epsilon)} \mathbb{E} \big\| \hat{\mathbf{\Sigma}} - \mathbf{\Sigma} \big\|^2 \ge C \varepsilon_{n,\mathbf{p}} + C \frac{\log p}{n}, \tag{4.11}$$

where $\varepsilon_{n,\mathbf{p}} = \min_{\mathbf{k}\in\mathcal{H}_d(\mathbf{p})}\{\tau^2(\mathbf{k}) + V(\mathbf{k})/n\}$.

Theorem 2 establishes a minimax lower bound for tensor covariance estimation under the spectral norm. Therein, to handle the difficulty owing to the tensor's multi-way dependence structure, we construct a new least favorable prior specifically tailored to tensor covariances. This prior incorporates block perturbations along an upper contour set of the tensor entries, with the perturbation strength respecting the covariance decay pattern. This result on the lower bound, together with the upper bound in Theorem 1, implies that

$$\inf_{\hat{\mathbf{\Sigma}}} \sup_{\mathbf{\Sigma}\in\mathcal{U}(d,\tau,\epsilon)} \mathbb{E}\big\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\big\|^2 \asymp \varepsilon_{n,\mathbf{p}} + \frac{\log p}{n}$$

for the heterogeneous decay $\tau(\mathbf{k})$ in (3.2), under $\log p = o(n)$ and $p_\ell > n^{\alpha_\ell^{-1}(2+\sum_{\ell'=1}^d \alpha_{\ell'}^{-1})^{-1}}$ for $\ell = 1, 2, \ldots, d$. Hence, the localization estimator $\mathcal{L}_h(\mathbf{S}_n; \mathbf{k}_h)$ in (3.6) is minimax rate optimal under the spectral norm.

4.2. *Covariance estimation under the Frobenius norm.* Past high-dimensional statistics research has paid attention to the optimal rate of convergence under the Frobenius norm, such as in Cai, Zhang and Zhou (2010) and Zhang, Shen and Kong (2023). Furthermore, there were also tuning parameter selection procedures developed based on the expectation of the Frobenius loss (Yi and Zou, 2013; Qiu and Chen, 2015; Sun et al., 2024).

Our analysis on the tensor covariance estimation under the Frobenius norm is made for the following covariance class

$$\mathcal{V}(\boldsymbol{\alpha}, d, \epsilon, C) = \Big\{\mathbf{\Sigma} = [\sigma_{ij}]_{p\times p}: \ \text{(i) } |\sigma_{ij}| \le C\prod_{\ell=1}^d (1+\delta_{ij\ell})^{-\alpha_\ell - 1} \text{ for } i, j = 1, 2, \ldots, p;$$
$$\text{(ii) } 0 \le \lambda_{\min}(\mathbf{\Sigma}) \le \lambda_{\max}(\mathbf{\Sigma}) \le \epsilon^{-1}\Big\},$$

where $\{\alpha_\ell\}_{\ell=1}^d$ and ϵ are positive constants and $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_d)^\top$. The covariance class $\mathcal{V}(\boldsymbol{\alpha}, d, \epsilon, C)$ replaces Condition (i) of the multi-bandable class $\mathcal{U}(d, \tau, \epsilon)$ in (3.4) with a more restrictive condition on the off-diagonal entries. One can see that $\mathcal{V}(\boldsymbol{\alpha}, d, \epsilon, C) \subset \mathcal{U}(d, \tau, \epsilon)$ with the covariance decay function $\tau(\mathbf{k}) = C\sum_{\ell=1}^d k_\ell^{-\alpha_\ell}\mathbb{I}\{1 \le k_\ell < p_\ell\}$ in (3.2). The covariance class $\mathcal{V}(\boldsymbol{\alpha}, d, \epsilon, C)$ is inspired by the bandable covariance class $\mathcal{U}_2(\alpha, \epsilon, C)$ and the separably bandable covariance class $\mathcal{U}_{3,2}(\alpha_1, \alpha_2, \epsilon, C)$. The following theorem leads to the consistency of the localization estimator under the Frobenius norm.

THEOREM 3. *Under Assumptions 2 and 4, if* $V(\mathbf{k}_h) = o(n)$*, the localization estimator* (3.6) *satisfies*

(4.12)
$$\sup_{\mathbf{\Sigma}\in\mathcal{V}(\boldsymbol{\alpha},d,\epsilon,C)} p^{-1}\mathbb{E}\big\|\mathcal{L}_h(\mathbf{S}_n; \mathbf{k}_h) - \mathbf{\Sigma}\big\|_F^2 \le C\sum_{\ell=1}^d k_{h\ell}^{-2\alpha_\ell - 1}\mathbb{I}\{1 \le k_{h\ell} < p_\ell/c_\ell\} + C\frac{V(\mathbf{k}_h)}{n},$$

where $\mathbf{c} = (c_1, c_2, \ldots, c_d)^\top$ *is defined in Assumption 2 (ii). Moreover,*

$$\inf_{\mathbf{k}_h} \sup_{\mathbf{\Sigma}\in\mathcal{V}(\boldsymbol{\alpha},d,\epsilon,C)} p^{-1}\mathbb{E}\big\|\mathcal{L}_h(\mathbf{S}_n; \mathbf{k}_h) - \mathbf{\Sigma}\big\|_F^2 \le C\varepsilon'_{n,\mathbf{p}}, \tag{4.13}$$

where $\varepsilon'_{n,\mathbf{p}} \asymp \min_{\mathbf{k}\in\mathcal{H}_d(\mathbf{p})}\big\{\sum_{\ell=1}^d k_\ell^{-2\alpha_\ell-1}\mathbb{I}\{1 \le k_\ell < p_\ell\} + n^{-1}\prod_{\ell=1}^d k_\ell\big\}$.

Theorem 3 demonstrates that the Frobenius risk of the localization estimator is controlled by a polynomially decayed term $\sum_{\ell=1}^{d} k_{h\ell}^{-2\alpha_\ell-1}\mathbb{I}\{1 \le k_{h\ell} < p_\ell/c_\ell\}$ and $V(\mathbf{k}_h)/n$. The former term on the right-hand side of (4.12) converges to 0 if the scaling parameters $k_{h\ell} \to \infty$ as n and $p_\ell \to \infty$, while the latter is guaranteed to vanish for $V(\mathbf{k}_h) = o(n)$. The upper bound $\varepsilon'_{n,\mathbf{p}}$ in (4.13) is tighter than the upper bound term $\varepsilon_{n,\mathbf{p}}$ in (4.6) with $\tau(\mathbf{k})$ given by (3.2). The reason is that $\varepsilon_{n,\mathbf{p}}$ directly controls the bias of the localization estimator by $\tau^2(\mathbf{k}_h \circ \mathbf{c})$, which can be improved as $\sum_{\ell=1}^{d} k_\ell^{-2\alpha_\ell-1}\mathbb{I}\{1 \le k_\ell < p_\ell/c_\ell\}$ for the Frobenius loss by averaging the bias of all the entries for the specific case $\mathbf{\Sigma} \in \mathcal{V}(\boldsymbol{\alpha}, d, \epsilon, C)$. Specifically, if $p_\ell > n^{(2\alpha_\ell+1)^{-1}\{1+\sum_{\ell'=1}^{d}(2\alpha_{\ell'}+1)^{-1}\}^{-1}}$ for all ℓ, then $\varepsilon'_{n,\mathbf{p}} \asymp n^{-\{1+\sum_{\ell=1}^{d}(2\alpha_\ell+1)^{-1}\}^{-1}}$, which is determined by the sample size n and each α_ℓ. See Section S7.2 of the SM for explicit examples on the Frobenius-norm convergence rate of $\mathbf{\Sigma} \in \mathcal{V}(\boldsymbol{\alpha}, d, \epsilon, C)$.

If an additional separability condition $\mathbf{\Sigma} = \mathbf{\Sigma}_1 \otimes \mathbf{\Sigma}_2 \otimes \cdots \otimes \mathbf{\Sigma}_d$ is assumed, a faster convergence rate for covariance estimation can be achieved as shown in Tsiligkaridis, Hero III and Zhou (2013); Zhang, Shen and Kong (2023) and McCormack and Hoff (2025). This is because the separability structure reduces the number of covariance parameters from $O(\prod_{\ell=1}^{d} p_\ell^2)$ to $O\{\sum_{\ell=1}^{d} p_\ell^2\}$. In contrast, the proposed multi-bandable covariance matrix $\mathcal{U}(d, \tau, \epsilon)$ is more general as it allows nonseparable dependence across all tensor directions. The rate in (4.13) may be viewed as a worst-case rate over a broad nonseparable class, which becomes slower as d increases, reflecting the intrinsic difficulty of estimating a fully nonseparable covariance matrix.

REMARK 2. *Although the GC function does not satisfy Assumption 2 (ii), for* $\mathbf{\Sigma} \in \mathcal{V}(\boldsymbol{\alpha}, d, \epsilon, C)$*, the localization estimator with the localization function* $GC(2\|\mathbf{z}\|)$ *can attain the upper bound of the estimation error under the spectral norm* (4.6) *if* $0 < \alpha_\ell < 2$ *for* $\ell = 1, 2, \ldots, d$*, and under the Frobenius norm* (4.13) *if* $0 < \alpha_\ell < 3/2$ *for* $\ell = 1, 2, \ldots, d$*. The key is to control the bias* $\|\mathcal{L}_{\mathrm{GC}}(\mathbf{\Sigma}; \mathbf{k}_{\mathrm{GC}}) - \mathbf{\Sigma}\| \le C\sum_{\ell=1}^{d} k_{\mathrm{GC},\ell}^{-\alpha_\ell}$ *and* $p^{-1}\|\mathcal{L}_{\mathrm{GC}}(\mathbf{\Sigma}; \mathbf{k}_{\mathrm{GC}}) - \mathbf{\Sigma}\|_F^2 \le C\sum_{\ell=1}^{d} k_{\mathrm{GC},\ell}^{-2\alpha_\ell-1}$*. See Section S7.3 of the SM for more discussion.*

For the minimax lower bound under the Frobenius norm, we consider the following covariance class with homogeneous polynomial decay

$$\mathcal{W}(\alpha, d, \epsilon, C) = \Big\{\mathbf{\Sigma} = [\sigma_{ij}]_{p\times p} : \text{(i) } |\sigma_{ij}| \le C\|\boldsymbol{\delta}_{ij}\|^{-\alpha-d} \text{ for } \boldsymbol{\delta}_{ij} \ne \mathbf{0}_d,$$
$$\text{(ii) } 0 \le \lambda_{\min}(\mathbf{\Sigma}) \le \lambda_{\max}(\mathbf{\Sigma}) \le \epsilon^{-1}\Big\},$$

where α and ϵ are positive constants. One can verify that $\mathcal{W}(\alpha, d, \epsilon, C)$ is a subset of $\mathcal{V}\big((\alpha/d)\mathbf{1}_d, d, \epsilon, \max\{2^{(\alpha+d)/d}C, \epsilon^{-1}\}\big)$. The following theorem establishes the minimax lower bound of estimating $\mathbf{\Sigma} \in \mathcal{W}(\alpha, d, \epsilon, C)$ under the Frobenius norm.

THEOREM 4. *For Gaussian-distributed* $\mathbf{X}_1$*, if* $p_\ell > n^{1/(2\alpha+2d)}$ *for all* $\ell = 1, 2, \ldots, d$*, the minimax risk of estimating* $\mathbf{\Sigma} \in \mathcal{W}(\alpha, d, \epsilon, C)$ *satisfies*

$$\inf_{\hat{\mathbf{\Sigma}}} \sup_{\mathbf{\Sigma}\in\mathcal{W}(\alpha,d,\epsilon,C)} p^{-1}\mathbb{E}\big\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\big\|_F^2 \ge Cn^{-\frac{2\alpha+d}{2\alpha+2d}}. \tag{4.14}$$

Specifically, the localization estimator $\mathcal{L}_h(\mathbf{S}_n; \mathbf{k}_h)$ in (3.6) with h satisfying Assumption 2 and $k_{h\ell} \asymp n^{1/(2\alpha+2d)}$ can attain the minimax rate of convergence specified in (4.14). It then implies that the localization estimator $\mathcal{L}_h(\mathbf{S}_n; \mathbf{k}_h)$ is minimax rate optimal for estimating $\mathbf{\Sigma} \in \mathcal{W}(\alpha, d, \epsilon, C)$ under the Frobenius norm. The detailed discussions for the minimax upper bounds and the general case when $\mathbf{\Sigma} \in \mathcal{V}(\boldsymbol{\alpha}, d, \epsilon, C)$ are provided in Section S7.2 of the SM.

4.3. *Adaptive localization estimator.* The scaling vector $\mathbf{k}_h$ in $\mathcal{L}_h(\mathbf{S}_n;\mathbf{k}_h)$ plays a pivotal role in achieving the convergence rates in Theorem 1. However, without prior knowledge of the covariance decay function $\tau(\mathbf{k})$, the explicit order of $\mathbf{k}_h$ cannot be specified and should be selected by a data-driven approach. To this end, we develop an adaptive localization estimator for multiplicative banding and tapering localization functions h, which builds on an empirical selection for the scaling vector $\mathbf{k}_h$ based on the Goldenshluger–Lepski method (Goldenshluger and Lepski, 2008), an extension of Lepski's method (Lepski, 1992) employed in Liu and Ren (2020) for banding widths selection.

The multiplicative banding and tapering functions are $h(\mathbf{z}) = \prod_{\ell=1}^{d} \mathbb{I}\{z_\ell < 1\}$ and $h(\mathbf{z}) = \prod_{\ell=1}^{d} \varphi(z_\ell; 0.5, 1)$, respectively, where φ is the tapering function in (2.5). We consider a candidate set

$$\mathcal{K}_{n,p} = \big\{\mathbf{k} \in \mathbb{N}^d : V(\mathbf{k}) < n/\log p \text{ and } 1 \le k_\ell \le p_\ell, \text{ for } \ell = 1, 2, \ldots, d\big\}.$$

to select their scaling vectors. Define an empirical comparison function

$$\hat{R}(\mathbf{k}) = \sup_{\mathbf{k}' \in \mathcal{K}_{n,p}} \left\{ \big\| \mathcal{L}_h(\mathbf{S}_n; \mathbf{k} \wedge \mathbf{k}') - \mathcal{L}_h(\mathbf{S}_n; \mathbf{k}') \big\|^2 - C_{\mathrm{GL}} \frac{\log p + V(\mathbf{k}')}{n} \right\}_+, \tag{4.15}$$

where $\mathbf{k} \wedge \mathbf{k}' = (\min\{k_1, k_1'\}, \min\{k_2, k_2'\}, \ldots, \min\{k_d, k_d'\})^\top$, $\{z\}_+ = \max\{z, 0\}$ and $C_{\mathrm{GL}} > 0$ is a sufficiently large constant. The quantity $\hat{R}(\mathbf{k})$ serves as a data-driven proxy for the squared bias associated with the localization scale $\mathbf{k}$. Specifically, for each comparison candidate $\mathbf{k}'$, the difference between $\mathcal{L}_h(\mathbf{S}_n; \mathbf{k} \wedge \mathbf{k}')$ and $\mathcal{L}_h(\mathbf{S}_n; \mathbf{k}')$ measures the discrepancy caused by imposing the localization scale $\mathbf{k}$ on $\mathcal{L}_h(\mathbf{S}_n; \mathbf{k}')$, treating $\mathcal{L}_h(\mathbf{S}_n; \mathbf{k}')$ as the true covariance. The penalty term $-C_{\mathrm{GL}}\{\log p + V(\mathbf{k}')\}/n$ compensates for the stochastic fluctuation involved in the comparison. Thus, a small value of $\hat{R}(\mathbf{k})$ indicates a small bias of choosing the candidate $\mathbf{k}$.

The adaptive scaling vector is then selected by

$$\hat{\mathbf{k}} \in \arg\min_{\mathbf{k} \in \mathcal{K}_{n,p}} \left\{ \hat{R}(\mathbf{k}) + C_{\mathrm{GL}} \frac{\log p + V(\mathbf{k})}{n} \right\}. \tag{4.16}$$

The criterion in (4.16) balances the estimated squared bias $\hat{R}(\mathbf{k})$ in (4.15) and the variance term corresponding to the candidate $\mathbf{k}$. Therefore, the adaptive scaling vector $\hat{\mathbf{k}}$ is selected as the candidate that achieves the empirical "bias–variance" trade-off over $\mathcal{K}_{n,p}$, without requiring prior knowledge of the covariance decay function $\tau(\mathbf{k})$.

THEOREM 5. *Under Assumptions 1 and 4, for* $\log p = o(n)$ *and* h *being the multiplicative banding or tapering function, the localization estimator* (3.6) *with adaptive scaling vector* (4.16) *satisfies*

$$\sup_{\mathbf{\Sigma} \in \mathcal{U}(d,\tau,\epsilon)} \mathbb{E} \big\| \mathcal{L}_h(\mathbf{S}_n; \hat{\mathbf{k}}) - \mathbf{\Sigma} \big\|^2 \le C \inf_{\mathbf{k} \in \mathcal{K}_{n,p}} \left\{ \tau^2(\mathbf{k}) + \frac{V(\mathbf{k})}{n} \right\} + C \frac{\log p}{n}.$$

Moreover, if there exists $\mathbf{k}_0 \in \mathcal{K}_{n,p}$ *such that* $\tau^2(\mathbf{k}_0) + V(\mathbf{k}_0)/n \le C_0 \varepsilon_{n,\mathbf{p}}$ *for some constant* $C_0 > 0$, *then*

$$\sup_{\mathbf{\Sigma} \in \mathcal{U}(d,\tau,\epsilon)} \mathbb{E} \big\| \mathcal{L}_h(\mathbf{S}_n; \hat{\mathbf{k}}) - \mathbf{\Sigma} \big\|^2 \le C \varepsilon_{n,\mathbf{p}} + C \frac{\log p}{n}.$$

Theorem 5 shows that the adaptive localization estimator for multiplicative banding and tapering functions attains the same spectral-norm convergence rate as in Theorem 1 without knowing the decay function $\tau(\mathbf{k})$. Thus, the data-driven selector $\hat{\mathbf{k}}$ in (4.16) automatically

balances the localization bias and stochastic variation. In particular, under the conditions in Theorem 2, the adaptive localization estimator is minimax rate optimal under the spectral norm. The adaptive estimator for general localization function h could be constructed in a similar way, which is left as a future investigation.

5. Simulation study. This section reports results from simulation experiments designed to evaluate the performance of the proposed covariance localization estimator. To gauge relative performance, the tapering estimator and two separable estimators were considered. The tapering estimator in Cai, Zhang and Zhou (2010) was employed for lattice order $d = 1$, 2 and 3, despite it was originally proposed for $d = 1$. The separable estimators were taken to be the separably tapering estimator in Zhang, Shen and Kong (2023) for $d = 2$ and the separable maximum likelihood estimator (MLE) in McCormack and Hoff (2025) for $d = 3$.

The vectorized tensor data $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ were generated from Gaussian and the multivariate T_{10} distributions with zero mean and covariance matrices specified below. We employed the localization function $h(\mathbf{z}) = \prod_{\ell=1}^{d} \varphi(z_\ell; 0.5, 1)$ with φ being the tapering function in (2.5). The scaling vector $\mathbf{k}_h$ was selected by the data splitting scheme introduced in Bickel and Levina (2008a). Specifically, for each of N random splits of the original sample of size n, the n observations were randomly partitioned into a training subsample of size $n_1 = \lfloor n/3 \rfloor$ and a test subsample of size $n_2 = n - n_1$. Let $\mathbf{S}_{n_1,b}^{(\mathrm{tr})}$ and $\mathbf{S}_{n_2,b}^{(\mathrm{test})}$ denote the sample covariance matrices of the training and test subsamples in the bth random split, respectively. Then, the scaling vector was selected by

$$\hat{\mathbf{k}}_h = \arg\min_{\mathbf{k}} \sum_{b=1}^{N} \left\| \mathcal{L}_h\big(\mathbf{S}_{n_1,b}^{(\mathrm{tr})}; \mathbf{k}\big) - \mathbf{S}_{n_2,b}^{(\mathrm{test})} \right\|_1, \tag{5.1}$$

where $\|\cdot\|_1$ denotes the matrix L_1 norm and N was set to be 50. The banding widths for the tapering and separably tapering estimators were selected similarly and we report the results in Table 3 along with those obtained from the adaptive method.

We considered three settings for the covariance matrix $\mathbf{\Sigma}$ for the Gaussian and T-distributed tensor data. The first setting had $\mathbf{\Sigma} = [\sigma_{ij}]_{p\times p}$ with

$$\text{Setting (i)}: \ \sigma_{ij} = \sqrt{a_i a_j}\exp(-\|\mathbf{s}_i - \mathbf{s}_j\|^2/2)$$

where $\{a_i\}_{i=1}^{p}$ were randomly drawn from $\mathrm{Unif}(0.5, 1.5)$ and were kept fixed once generated. Since $\mathbf{\Sigma} = \mathbf{ACA}$ with $\mathbf{A} = \mathrm{diag}(\sqrt{a_1}, \sqrt{a_2}, \ldots, \sqrt{a_p})$ and $\mathbf{C}$ being positive definite, this covariance is positive definite despite the randomness of generating $\{a_i\}$. Tensor orders $d = 1, 2$ and 3 were experimented with $p = p_1 = 64$, 729 and 4096 for $d = 1$, $(p_1, p_2) = (8, 8), (27, 27)$ and $(64, 64)$ for $d = 2$ and $(p_1, p_2, p_3) = (4, 4, 4), (9, 9, 9)$ and $(16, 16, 16)$ for $d = 3$, leading to $p = 64$, 729 and 4096 entries for each $d = 1$, 2 and 3.

Table 1 reports the empirical estimation errors of the proposed localization estimator under the spectral and the Frobenius norms for Simulation Setting (i) for Gaussian-distributed data, while those for the T-distributed data exhibited similar patterns of results and are hence reported in Section S8 of the SM. Table 1 shows that for each combination of dimension p and tensor order d, the estimation errors under both the spectral and the Frobenius norms decreased as the sample size n increased, with a similar accompanying trend for the standard deviations of the estimation errors. These indicated the consistency of the localization estimators. When the dimension p and sample size n were fixed, both the spectral and the Frobenius norms of the estimation errors increased as the tensor order d got larger, which was consistent with the results in Section 4 that a larger d leads to a slower rate of convergence. Furthermore, simulation results of the localization estimator using the multiplicative banding function $h(\mathbf{z}) = \prod_{\ell=1}^{d} \mathbb{I}\{z_\ell < 1\}$ and the GC function $h(\mathbf{z}) = \mathrm{GC}(2\|\mathbf{z}\|)$ are reported in

TABLE 1

Average empirical estimation errors and their standard deviations (in parentheses) of the proposed localization covariance estimator under the spectral and the Frobenius norms for Simulation Setting (i) for Gaussian-distributed data with respect to the dimension p, the sample size n and the order of the lattice d.

p	n	$\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|$			$\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_F$		
		$d=1$	$d=2$	$d=3$	$d=1$	$d=2$	$d=3$
64	50	1.07(0.24)	1.88(0.3)	2.95(0.52)	2.92(0.26)	4.4(0.37)	5.83(0.57)
	100	0.77(0.13)	1.48(0.22)	2.37(0.4)	2.03(0.22)	3.45(0.25)	4.56(0.36)
	250	0.5(0.09)	0.97(0.17)	1.7(0.33)	1.33(0.12)	2.16(0.23)	3.25(0.3)
	500	0.37(0.06)	0.73(0.13)	1.2(0.26)	1(0.08)	1.6(0.13)	2.24(0.26)
	1000	0.28(0.05)	0.58(0.1)	0.94(0.21)	0.79(0.07)	1.27(0.09)	1.69(0.15)
	2000	0.19(0.04)	0.48(0.09)	0.78(0.16)	0.48(0.04)	1.05(0.09)	1.39(0.11)
729	50	1.47(0.21)	2.58(0.34)	5.06(0.35)	10.01(0.26)	16.11(0.41)	23.7(0.66)
	100	1.08(0.15)	2(0.15)	4.46(0.28)	6.79(0.2)	13.1(0.27)	19.15(0.4)
	250	0.68(0.08)	1.35(0.14)	3.47(0.32)	4.49(0.12)	7.79(0.19)	14.54(0.49)
	500	0.49(0.06)	1.03(0.1)	2.28(0.22)	3.4(0.08)	5.94(0.14)	9.42(0.23)
	1000	0.36(0.04)	0.84(0.07)	1.96(0.16)	2.71(0.06)	4.77(0.11)	7.41(0.17)
	2000	0.26(0.03)	0.71(0.05)	1.75(0.12)	1.66(0.05)	4.06(0.08)	6.15(0.14)
4096	50	1.78(0.22)	2.97(0.33)	5.86(0.24)	23.67(0.25)	38.83(0.4)	59.64(0.66)
	100	1.29(0.16)	2.22(0.15)	5.23(0.15)	16.04(0.19)	31.68(0.24)	48.42(0.42)
	250	0.78(0.08)	1.53(0.11)	4.18(0.12)	10.64(0.12)	18.83(0.2)	37.39(0.24)
	500	0.56(0.06)	1.17(0.08)	2.74(0.15)	8.08(0.08)	14.39(0.14)	24.09(0.24)
	1000	0.41(0.03)	0.94(0.06)	2.34(0.1)	6.44(0.06)	11.57(0.1)	18.96(0.18)
	2000	0.3(0.03)	0.78(0.04)	2.1(0.07)	3.94(0.04)	9.86(0.08)	15.79(0.14)

Section S8 of the SM, respectively, which exhibit similar patterns of results as Table 1 and suggest that the empirical performance was robust to the choice of localization function.

To assess the comparative performance of the estimators, Figure 4 displays the estimation errors of the localization estimator for Setting (i) with $d = 2$ and 3 along with the sample covariance, the tapering estimator for $d = 2$ and 3, and the separably tapering estimators of Zhang, Shen and Kong (2023) for $d = 2$ and the separable MLE of McCormack and Hoff (2025) for $d = 3$. The figure shows that, although all the covariance estimators had their estimation errors reduced as the sample size n was increased, the sample covariance and the tapering estimator exhibited quite subdued reductions and tended to plateau for large n, and had much larger estimation errors than those of the localization estimator. The separable estimators improved upon the performance of the sample covariance and the tapering estimator, but still had larger estimation errors than the localization estimator. The proposed localization estimator achieved the smallest estimation errors in almost all the settings. It is noted that the tapering estimator designed for $d = 1$ was not sufficiently tailored to capture the multi-bandable structure experimented in the simulation as argued in Section 4.

We then considered the second simulation setting for $\mathbf{\Sigma} = [\sigma_{ij}]$ for $d = 2$ with entries

$$\text{Setting (ii)}: \ \sigma_{ij} = 2\{(\lceil i/p_1 \rceil)/(p_2+1)\}^{|i-j|}\mathbb{I}\{\lceil i/p_1 \rceil = \lceil j/p_1 \rceil\}$$

where $\lceil \cdot \rceil$ is the ceiling function. This yields a block diagonal covariance matrix for $d = 2$ that included p_2 block matrices of size p_1 with the entries of the uth block ($u = \lceil i/p_1 \rceil$) being $\sigma_{ij} = 2\{u/(p_2+1)\}^{|i-j|}$, where (p_1, p_2) were considered as $(10, 20)$, $(10, 50)$ and $(20, 50)$, respectively. We compared the localization estimator with the tapering and the separably tapering estimators in this highly nonseparable setting.

Table 2 reports the average empirical estimation errors of the proposed localization estimator under the spectral and the Frobenius norms for Simulation Setting (ii) along with the sample covariance, the tapering and the separably tapering estimators for $d = 2$. Table 2 shows that, although all the estimators obtained decreasing estimation errors under both the spectral and the Frobenius norms as either the dimensions p_1 and p_2 decreased or the sample

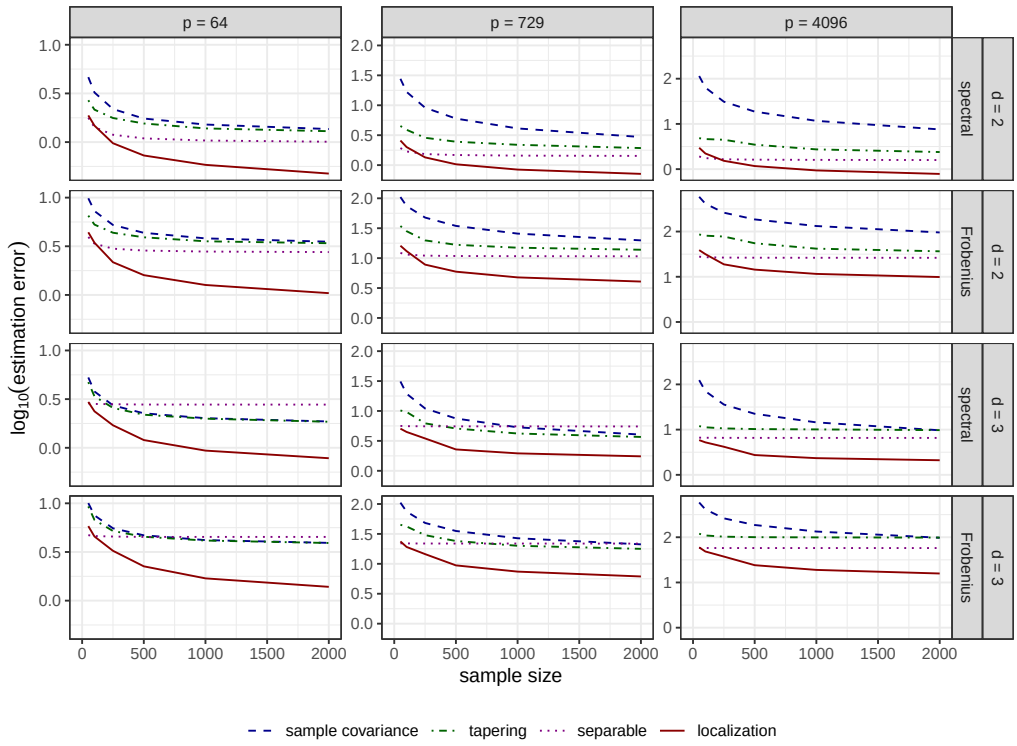


Fig 4: Average empirical estimation errors (in $\log_{10}$-scale) of the proposed localization estimator (red solid lines), the sample covariance (blue dashed lines), the tapering estimator (green dashed-dotted lines), and the separably tapering estimator for $d = 2$ and the separable MLE for $d = 3$ (both denoted as separable, purple dotted lines) under the spectral and the Frobenius norms with respect to the dimension p and the sample size n for Setting (i) under Gaussian data.

size n increased, the proposed localization estimator outperformed the other estimators in all the settings. In particular, the separably tapering estimator had the largest estimation errors under the spectral and the Frobenius norms for $n = 2000$ as it tended to construct the same covariance pattern across all the blocks in Simulation Setting (ii), which thus introduced non-negligible bias. These results demonstrated that the localization estimator is a good choice for the nonseparable case.

To evaluate the effects of the scaling parameter selection on the localization estimators, we considered the third simulation setting for $\mathbf{\Sigma} = [\sigma_{ij}]$ with

$$\text{Setting (iii)}: \ \sigma_{ij} = \sqrt{a_i a_j} \prod_{\ell=1}^{3} \{1 + |s_{i\ell} - s_{j\ell}|\}^{-\alpha_\ell - 1},$$

where $\{a_i\}$ were similarly generated as in Simulation Setting (i) for $d = 3$ and $p_1 = p_2 = p_3 = 10$, and $(\alpha_1, \alpha_2, \alpha_3) = (0.4, 0.6, 0.8)$. We evaluated the performance of the localization estimators with the data splitting and the adaptive estimation schemes based on the Goldenshluger-Lepski method.

Table 3 reports the average estimation errors of the localization estimator with the empirically selected localization scaling parameters by the sample splitting and the adaptive methods for Simulation Setting (iii) with Gaussian-distributed data. It shows that for both the data splitting and the adaptive schemes, the estimation errors declined as the sample size n

TABLE 2

Average empirical estimation errors and their standard deviations (in parentheses) of the localization estimator (proposed), the sample covariance (sample) and the tapering (tapering) and separably tapering (separable) estimators under the spectral and Frobenius norms for Simulation Setting (ii) for $d = 2$ and Gaussian-distributed data with respect to the dimension (p_1, p_2) and the sample size n.

n	$\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|$				$\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_F$			
	sample	tapering	separable	proposed	sample	tapering	separable	proposed
	$(p_1, p_2) = (10, 20)$							
50	20.07(2.03)	7.21(1.75)	6.31(1.95)	5.67(1.95)	57.43(1.41)	18.02(1.93)	18.25(1.57)	15.07(2.18)
100	12.98(1.21)	4.95(1.39)	5.31(1.39)	3.78(1.53)	40.46(0.77)	12.95(1.41)	17.03(0.89)	10.47(1.72)
250	7.45(0.73)	2.98(0.81)	4.54(0.67)	2.22(0.75)	25.48(0.38)	8.27(0.7)	16.31(0.19)	6.34(0.58)
500	5.08(0.49)	2.03(0.6)	4.33(0.4)	1.56(0.48)	18.02(0.25)	5.92(0.49)	16.14(0.07)	4.46(0.31)
1000	3.48(0.29)	1.45(0.31)	4.18(0.24)	1.07(0.32)	12.72(0.17)	4.29(0.23)	16.06(0.03)	3.13(0.21)
2000	2.42(0.21)	1(0.28)	4.09(0.15)	0.76(0.22)	8.99(0.11)	3.06(0.23)	16.02(0.02)	2.22(0.15)
	$(p_1, p_2) = (10, 50)$							
50	40.39(2.49)	8.53(1.94)	7.75(2.01)	7.2(2.13)	143.21(2.25)	29.76(3.68)	30.27(3.04)	24.78(4.29)
100	24.98(1.46)	6.24(1.58)	6.52(1.49)	4.98(1.69)	100.72(1.12)	21.82(3.15)	28.23(1.76)	17.38(3.44)
250	13.89(0.84)	3.78(0.94)	5.52(0.7)	2.9(0.94)	63.54(0.51)	13.72(1.49)	26.93(0.56)	10.45(1.74)
500	9.17(0.52)	2.5(0.68)	5.18(0.35)	1.89(0.5)	44.88(0.3)	9.64(0.83)	26.57(0.11)	7.08(0.51)
1000	6.15(0.33)	1.8(0.37)	5.08(0.2)	1.33(0.35)	31.7(0.18)	6.98(0.34)	26.45(0.03)	4.96(0.23)
2000	4.2(0.22)	1.22(0.32)	5.01(0.16)	0.96(0.24)	22.41(0.12)	4.96(0.36)	26.4(0.02)	3.53(0.16)
	$(p_1, p_2) = (20, 50)$							
50	73.94(4.07)	18.37(4.67)	16.06(5.02)	13.96(5.15)	286.07(3.23)	52.38(5.92)	52.07(5.13)	45.88(6.19)
100	45.25(2.63)	13.81(3.93)	13.56(3.74)	9.61(4.01)	201.33(1.67)	39.93(5)	48.94(2.71)	32.73(4.99)
250	24.95(1.55)	8.06(2.53)	11.1(1.91)	5(1.83)	126.94(0.74)	25.52(2.64)	46.8(0.55)	19.57(1.78)
500	16.38(1.04)	5.15(1.62)	10.44(1.15)	3.21(0.94)	89.62(0.42)	18.01(1.23)	46.39(0.11)	13.53(0.48)
1000	11(0.68)	3.38(0.89)	10.17(0.88)	2.32(0.67)	63.37(0.25)	12.88(0.55)	46.25(0.06)	9.64(0.35)
2000	7.52(0.47)	2.33(0.6)	9.94(0.64)	1.63(0.43)	44.8(0.15)	9.33(0.34)	46.18(0.02)	6.87(0.23)

TABLE 3

Average empirical estimation errors and the selected scaling parameters, with their respective standard deviations reported in parentheses, for the localization estimator with data splitting and the adaptive schemes for Simulation Setting (iii) for $d = 3$ and the Gaussian-distributed data with respect to the dimension of the lattice $p_1 = p_2 = p_3 = 10$ and the sample size n ranging from 50 to 2000. (Zero standard deviations indicate values below 0.005)

n	$\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|$	$\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_F$	k_1	k_2	k_3
	data splitting				
50	9.96(0.48)	30.24(1.42)	2.23(1.29)	1.68(0.9)	1.16(0.38)
100	8.99(0.86)	26.50(1.76)	2.51(0.5)	1.78(0.74)	1.41(0.49)
250	8.26(0.80)	22.99(1.92)	2.58(0.49)	2.17(0.75)	1.75(0.71)
500	6.87(0.64)	18.52(1.56)	3.03(0.72)	2.65(0.55)	2.29(0.67)
1000	5.93(0.49)	15.32(0.93)	3.66(0.79)	3.03(0.65)	2.59(0.49)
2000	5.13(0.49)	12.92(0.84)	4.55(1.28)	3.45(0.63)	2.89(0.33)
	adaptive				
50	9.24(0.2)	28.33(0.64)	2.21(0.43)	1.86(0.46)	1.2(0.42)
100	7.4(0.21)	24.34(0.31)	2.9(0.3)	2.02(0.13)	2(0)
250	5.65(0.18)	19.6(0.22)	4.01(0.23)	3.02(0.13)	2.98(0.15)
500	4.5(0.21)	17.25(0.43)	5.44(0.75)	3.7(0.46)	3.45(0.5)
1000	3.42(0.19)	14.9(0.24)	6.74(0.91)	4.66(0.49)	4.06(0.23)
2000	2.65(0.22)	12.5(0.49)	7.71(1.4)	5.94(0.82)	4.96(1.11)

increased. The adaptive scheme achieved smaller estimation errors and variation than the data splitting scheme among all sample sizes considered under both the spectral and the Frobenius norms. This confirmed the advantage of doing the adaptive estimation with the data-driven scaling parameter selection. For both schemes, the selected scaling parameters k_1, k_2 and k_3 increased with the sample size n, which confirmed the results in Theorems 1 and 3 of Section 4 that the scaling parameters should scale with n. It is noted that the relative sizes of the estimated k_ℓ reflected inversely the size of α_ℓ, which defined the speed of the dependence decay at the lattice direction ℓ in that a slower decay required larger k_ℓ. These results demonstrate that the proposed localization estimator and adaptive method can effectively accommodate the heterogeneous decay patterns in each tensor direction.

6. Case study. Oceanic eddies play a critical role in modulating ocean heat exchange, nutrient distribution and climate variability (Beech et al., 2022; He et al., 2024; Receveur et al., 2024). Reconstruction of the salinity fields of an ocean eddy helps to provide high-resolution insights into ocean dynamics and enhance oceanographic studies. We consider estimating the covariance matrix of the daily salinity changes associated with the eddy in the western Pacific (32°N–35°N and 158°E–161°E) from July 20 to September 12, 2024, with a spatial resolution of $1/12°$. The data were acquired from the daily salinity fields from GLORYS reanalysis products over the study area.

Since the eddy moved, we re-centered the eddy each day by aligning its center to the eddy center of the first day. The eddy center for each day was calculated as the spatial location where the daily averaged surface velocity was smallest. As we are interested in the salinity structure of the eddy, re-centering makes the daily salinity fields spatially comparable in the eddy-center coordinate over a fixed domain of $3° \times 3°$ spatially and $1\,\mathrm{km}$ in depth. The study domain was partitioned into a 37×37 longitude-latitude grid, coupled with 35 vertical layers that were gradually thinned out from $0.5\,\mathrm{m}$ resolution to $1000\,\mathrm{m}$; see the SM for details. We treated each day as a replication, which led to 54 observations of the daily salinity changes over a total of $p = 47915$ entries. The eddy's salinity field was temporally dependent, and taking the daily changes would reduce the dependence. We calculated the empirical autocorrelation coefficient (AC) for all tensor entries before and after taking the daily difference. The analysis in Section S9.1 of the SM showed that taking the daily difference reduced the average lag-1 AC from 0.83 to 0.17. We would treat the daily changes as independent data as our purpose here is to demonstrate the multi-bandable structure and the need for estimating the covariance for tensor data.

We considered the localization estimator $\mathcal{L}_h(\mathbf{S}_n; \mathbf{k}_h)$ in (3.6) with the localization function being the multiplicative banding function $h(\mathbf{z}) = \prod_{\ell=1}^{3} \mathbb{I}\{z_\ell < 1\}$ and the multiplicative tapering function $h(\mathbf{z}) = \prod_{\ell=1}^{3} \varphi(z_\ell; 0.5, 1)$. The scaling vector $\mathbf{k}_h = (k_{h1}, k_{h2}, k_{h3})^\top$ of the localization estimator was selected via the data splitting procedure in (5.1) with $N = 50$ replications and was chosen from the set $\{0.1, 0.2, \ldots, 1\}$ degrees for the longitude and the latitude, and $\{10, 20, \ldots, 200\}$ meters for the depth. Figures S9 and S10 in the SM display the objective function in (5.1) using the two localization functions for each combination of the scaling parameters. The optimal scaling parameters were chosen as $0.2°$ in longitude, $0.3°$ in latitude and $80\,\mathrm{m}$ in depth for the multiplicative banding function, and $0.3°$ in longitude, $0.4°$ in latitude and $200\,\mathrm{m}$ in depth for the multiplicative tapering function. We would like to interpret the selected parameters from an oceanography perspective. The chosen scaling parameters in the longitude and the latitude were around $30\,\mathrm{km}$, which were comparable to the first baroclinic Rossby radius of deformation, an important quantity in determining horizontal scales that characterize eddy sizes (Chelton et al., 1998). An illustration of the proposed covariance localization estimator of the sea surface of the daily salinity changes and an inset corresponding to the estimation of the sea surface is shown in Figure S8, which successfully captured a finer multi-bandable covariance structure for the ocean tensor data.

The precision of the covariance matrix estimation was evaluated via a three-dimensional variational assimilation (3DVar) framework, which critically depended on the quality of estimation on a key covariance matrix. To be specific, we estimated the covariance matrix $\mathbf{\Sigma}$ of the daily salinity changes using the daily changes in salinity over the first 53 days, while the data on September 12th, 2024, the last day of the study period, was used as the testing set. We randomly selected the salinity data on 5% of the entries while adding independent $\mathcal{N}(0, 0.01\mathbf{I})$ distributed noise to the observation on each observed entry. The aim was to reconstruct the salinity field on the remaining 95% entries. Denote by $\mathbf{Y}$ the observations on the observed entries and let $\mathbf{H}$ be a matrix that maps the state variables to the observations. The 3DVar assimilated the salinity field on the last day as

$$\hat{\mathbf{X}} = \mathbf{X}_0 + \hat{\mathbf{\Sigma}}\mathbf{H}^\top \left(\mathbf{H}\hat{\mathbf{\Sigma}}\mathbf{H}^\top + \mathbf{R}\right)^{-1} (\mathbf{Y} - \mathbf{H}\mathbf{X}_0), \tag{6.1}$$

TABLE 4

Average reconstruction errors and their 5% and 95% quantiles in parentheses of the reconstructed state variable $\hat{\mathbf{X}}$ in the salinity field on the last day by using the sample covariance, the banding, the tapering and the separable estimators and the localization estimators with multiplicative banding and tapering functions, including the L_2 distance, the L_1 distance and the Hamming distance divided by p.

Estimation	$\|\mathbf{X}-\hat{\mathbf{X}}\|$	$\|\mathbf{X}-\hat{\mathbf{X}}\|_1$	$p^{-1}H\{\text{sgn}(\mathbf{X}),\text{sgn}(\hat{\mathbf{X}})\}$
sample covariance	6.14(6.11,6.19)	918.5(905.97,932.76)	0.35(0.34,0.36)
banding	7.5(7.47,7.52)	1065.4(1063.1,1067.5)	0.48(0.48,0.49)
tapering	7.48(7.46,7.51)	1063.3(1061.1,1065.7)	0.48(0.48,0.49)
separable	5.85(5.75,5.96)	848.7(838.7,859)	0.22(0.21,0.23)
localization (banding)	4.59(4.46,4.74)	685.5(673.3,698.7)	0.23(0.22,0.24)
localization (tapering)	4.46(4.33,4.58)	664.5(652.4,676.5)	0.21(0.21,0.22)

where $\hat{\mathbf{X}}$ was the estimated salinity changes in the last day, $\mathbf{X}_0$ was a first guess for the increment which was set to be 0 psu as it reflected the salinity anomaly, $\hat{\mathbf{\Sigma}}$ was the covariance estimate of the daily salinity changes using the first 53 days' data and $\mathbf{R}=0.01\mathbf{I}$ represented the observational error covariance matrix.

The estimate $\hat{\mathbf{\Sigma}}$ was obtained based on the localization estimator $\mathcal{L}_h(\mathbf{S}_n;\mathbf{k}_h)$ employing the two localization functions $h(\mathbf{z})=\prod_{\ell=1}^3\mathbb{I}\{z_\ell<1\}$ and $h(\mathbf{z})=\prod_{\ell=1}^3\varphi(z_\ell;0.5,1)$, the sample covariance, the banding estimator (Bickel and Levina, 2008a), the tapering estimator (Cai, Zhang and Zhou, 2010) and the separable MLE (McCormack and Hoff, 2025). The scaling parameters of the localization estimator for the two functions were set to be the aforementioned optimized scaling parameters selected via the data splitting procedure in (5.1). The banding width parameters for the banding and tapering estimators were chosen as 3 and 4, respectively, via the data splitting procedure, whose objective function values are shown in Figure S11 of the SM. For each covariance matrix estimator, the 3DVar assimilation was replicated 500 times, in which the 5% observed entries and the noise added to the observations contained in $\mathbf{Y}$ were randomly generated in each repetition. We treated the last day as test data to verify the accuracy of the reconstruction on the 95% missing salinity values.

Table 4 summarizes the reconstruction errors for the 95% "missing" entries on the last day using the localization estimator, the sample covariance, the banding and tapering estimators, and the separable MLE for $\hat{\mathbf{\Sigma}}$ being used in (6.1). Three metrics were used to evaluate the reconstruction errors, namely the L_2 distance $\|\mathbf{X}-\hat{\mathbf{X}}\|$, the L_1 distance $\|\mathbf{X}-\hat{\mathbf{X}}\|_1$ and the Hamming distance between $\text{sgn}(\mathbf{X})$ and $\text{sgn}(\hat{\mathbf{X}})$, where $\text{sgn}(\mathbf{X})$ denotes the sign indicator function. The results demonstrated that the reconstruction using the two localization covariance estimators $\mathcal{L}_h(\mathbf{S}_n;\mathbf{k}_h)$ recovered the underlying salinity field more accurately with much smaller reconstruction errors in all three metrics than those using the sample covariance. In contrast, the banding and tapering estimators encountered larger reconstruction errors even than those using the sample covariance, which might be due to the fact that the two covariance estimators were designed for vector data and were not suited for the three-order tensor data that we were dealing with here. The separable covariance estimator obtained smaller reconstruction errors than the sample covariance and the banding and tapering estimators, but larger errors than the localization estimators, suggesting a potential nonseparable dependence structure in daily salinity changes.

7. Discussion. We study covariance estimation for high-dimensional tensor data with a flexible multi-bandable covariance structure, which allows nonseparable covariance matrices and heterogeneous decay patterns across tensor directions, thereby accommodating complex dependence structures encountered in tensor data. We establish the minimax bounds under both the spectral and the Frobenius norms for a general covariance decay class, and show that the proposed localization covariance estimator attains the optimal convergence rates.

As a byproduct, our theory also provides a new optimality result for the banding estimator for vector data. We also develop a data-driven adaptive localization estimator based on the Goldenshluger–Lepski method for selecting the scaling parameters, which achieves the same minimax optimal rate without requiring prior knowledge of the covariance decay structure.

There are several possible extensions of the proposed framework. First, although the convergence rates in Theorems 1 and 3 are insensitive to the choice of the localization function h, its choice can impact finite-sample performance. When non-negative definiteness is desired, the GC function may be preferable because it yields a non-negative definite covariance estimator. An ideal localization function should reflect the empirical pattern of covariance decay across tensor entries. In practice, we can provide empirical summaries of how the covariance magnitude decays with distance. If the tensor covariance satisfies an isotropic structure where covariances between two variables are only determined by their distances, we can estimate the decay function by calculating the average absolute empirical covariances $|\hat{\sigma}_{ij}|$ over the pairs of variables with the same distance. More generally, when the tensor covariance exhibits heterogeneous decay patterns across different directions or slices, analogous covariance-distance profiles can be constructed within each sliced covariance matrix. One direction is to develop fully data-adaptive procedures for estimating the optimal localization function h for covariance estimation.

Another direction is covariance estimation for temporally dependent tensor data as in the case study in Section 6. Even after detrending, temporal dependence may remain. Extending the present analysis from independent tensor observations to temporally dependent tensor data would be practically important. One possible formulation is to assume that the tensor observations satisfy a strongly mixing condition, under which the temporal dependence decays sufficiently fast as the time lag increases.

Acknowledgments. The authors would like to thank two anonymous referees, the Associate Editor and the Editor for constructive comments and suggestions, which have improved the presentation of the paper. Yumou Qiu is the corresponding author.

Funding. This study was supported by the NSFC Major Research Plan on West-Pacific Earth System Multispheric Interactions (Grant No. 92358303), the Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (Grant No. JYB2025XDXM801), and the NSFC Major Program (Grant Nos. 12292980 and 12292983). We thank for the technical support of the National Large Scientific and Technological Infrastructure "Earth System Numerical Simulation Facility".

SUPPLEMENTARY MATERIAL

Supplementary Material to "Localization Estimator for High-Dimensional Tensor Covariance Matrices"

In the supplementary material, we present technical details, proofs and additional results of the simulations and the case study.

REFERENCES

BAI, Z. D., SILVERSTEIN, J. W. and YIN, Y. Q. (1988). A note on the largest eigenvalue of a large dimensional sample covariance matrix. *J. Multivariate Anal.* **26** 166–168.

BAI, Z. D. and YIN, Y. Q. (1993). Limit of the smallest eigenvalue of a large dimensional sample covariance matrix. *Ann. Probab.* **21** 1275–1294.

BEECH, N., RACKOW, T., SEMMLER, T., DANILOV, S., WANG, Q. and JUNG, T. (2022). Long-term evolution of ocean eddy activity in a warming world. *Nat. Clim. Change* **12** 910–917.

BICKEL, P. J. and LEVINA, E. (2008a). Regularized estimation of large covariance matrices. *Ann. Statist.* **36** 199–227.

BICKEL, P. J. and LEVINA, E. (2008b). Covariance regularization by thresholding. *Ann. Statist.* **36** 2577–2604.

CAI, T. T. and YUAN, M. (2012). Adaptive covariance matrix estimation through block thresholding. *Ann. Statist.* **40** 2014 – 2042.

CAI, T. T., ZHANG, C.-H. and ZHOU, H. H. (2010). Optimal rates of convergence for covariance matrix estimation. *Ann. Statist.* **38** 2118 – 2144.

CHELTON, D. B., DESZOEKE, R. A., SCHLAX, M. G., EL NAGGAR, K. and SIWERTZ, N. (1998). Geographical variability of the first baroclinic Rossby radius of deformation. *J. Phys. Oceanogr.* **28** 433–460.

CRESSIE, N. and HUANG, H.-C. (1999). Classes of nonseparable, spatio-temporal stationary covariance functions. *J. Amer. Statist. Assoc.* **94** 1330–1339.

DALEY, R. and BARKER, E. (2001). NAVDAS: formulation and diagnostics. *Mon. Wea. Rev.* **129** 869–883.

FANG, B., GUNTUBOYINA, A. and SEN, B. (2021). Multivariate extensions of isotonic regression and total variation denoising via entire monotonicity and Hardy-Krause variation. *Ann. Statist.* **49** 769–792.

FURRER, R. and BENGTSSON, T. (2007). Estimation of high-dimensional prior and posterior covariance matrices in Kalman filter variants. *J. Multivariate Anal.* **98** 227–255.

GASPARI, G. and COHN, S. E. (1999). Construction of correlation functions in two and three dimensions. *Q. J. R. Meteorol. Soc.* **125** 723–757.

GOLDENSHLUGER, A. and LEPSKI, O. (2008). Universal pointwise selection rule in multivariate function estimation. *Bernoulli* **14** 1150–1190.

GUTTORP, P. and SCHMIDT, A. M. (2013). Covariance structure of spatial and spatiotemporal processes. *Wiley Interdiscip. Rev. Comput. Stat.* **5** 279–287.

HAKIM, G. J. (2005). Vertical structure of midlatitude analysis and forecast errors. *Mon. Wea. Rev.* **133** 567–578.

HAMILL, T. M., WHITAKER, J. S. and SNYDER, C. (2001). Distance-dependent filtering of background error covariance estimates in an ensemble Kalman filter. *Mon. Wea. Rev.* **129** 2776–2790.

HAVIV, D., REMŠÍK, J., GATIE, M., SNOPKOWSKI, C., TAKIZAWA, M., PEREIRA, N., BASHKIN, J., JOVANOVICH, S., NAWY, T., CHALIGNE, R. et al. (2025). The covariance environment defines cellular niches for spatial inference. *Nat. Biotechnol.* **43** 269–280.

HE, Q., ZHAN, W., FENG, M., GONG, Y., CAI, S. and ZHAN, H. (2024). Common occurrences of subsurface heatwaves and cold spells in ocean eddies. *Nature* **634** 1111–1117.

JOHNSTONE, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.* **29** 295–327.

LEPSKI, O. (1992). Asymptotically minimax adaptive estimation. I. Upper bounds. optimally adaptive estimates. *Theory Probab. Appl.* **36** 682–697.

LIU, Y. and REN, Z. (2020). Minimax estimation of large precision matrices with bandable Cholesky factor. *Ann. Statist.* **48** 2428–2454.

MCCORMACK, A. and HOFF, P. (2025). Information geometry and asymptotics for Kronecker covariances. *Bernoulli* **31** 3165–3186.

MOORE, A. M., ARANGO, H. G., BROQUET, G., POWELL, B. S., WEAVER, A. T. and ZAVALA-GARAY, J. (2011). The Regional Ocean Modeling System (ROMS) 4-dimensional variational data assimilation systems: Part I–System overview and formulation. *Prog. Oceanogr.* **91** 34–49.

QIU, Y. and CHEN, S. X. (2015). Bandwidth selection for high-dimensional covariance matrix estimation. *J. Amer. Statist. Assoc.* **110** 1160–1174.

RECEVEUR, A., MENKES, C., LENGAIGNE, M., ARIZA, A., BERTRAND, A., DUTHEIL, C., CRAVATTE, S., ALLAIN, V., BARBIN, L., LEBOURGES-DHAUSSY, A. et al. (2024). A rare oasis effect for forage fauna in oceanic eddies at the global scale. *Nat. Commun.* **15** 4834.

SUN, H.-X., WANG, S., ZHENG, X. and CHEN, S. X. (2024). High-dimensional ensemble Kalman filter with localization, inflation, and iterative updates. *Q. J. R. Meteorol. Soc.* **150** 4870–4884.

TSILIGKARIDIS, T., HERO III, A. O. and ZHOU, S. (2013). On convergence of Kronecker graphical lasso algorithms. *IEEE Trans. Signal Process.* **61** 1743–1755.

VASWANI, A., SHAZEER, N., PARMAR, N., USZKOREIT, J., JONES, L., GOMEZ, A. N., KAISER, Ł. and POLOSUKHIN, I. (2017). Attention is all you need. *Adv. Neural Inf. Process. Syst.* **30**.

VITALI, G. (1908). Sui gruppi di punti e sulle funzioni di variabili reali. *Atti Accad. Sci. Torino* **43** 75–92.

YI, F. and ZOU, H. (2013). SURE-tuned tapering estimation of large covariance matrices. *Comput. Statist. Data Anal.* **58** 339–351.

ZHANG, Y., SHEN, W. and KONG, D. (2023). Covariance estimation for matrix-valued data. *J. Amer. Statist. Assoc.* **118** 2620–2631.