Biometrika (2024), **111**, 3, *pp.* 843–864
https://doi.org/10.1093/biomet/asae002
Advance Access publication 20 January 2024

Efficient nonparametric estimation of Toeplitz covariance matrices

By K. KLOCKMANN and T. KRIVOBOKOVA

Department of Statistics and Operations Research, Universität Wien,
Oskar-Morgenstern-Platz 1, 1090 Wien, Austria
karolina.klockmann@univie.ac.at tatyana.krivobokova@univie.ac.at

Summary

A new efficient nonparametric estimator for Toeplitz covariance matrices is proposed. This estimator is based on a data transformation that translates the problem of Toeplitz covariance matrix estimation to the problem of mean estimation in an approximate Gaussian regression. The resulting Toeplitz covariance matrix estimator is positive definite by construction, fully data driven and computationally very fast. Moreover, this estimator is shown to be minimax optimal under the spectral norm for a large class of Toeplitz matrices. These results are readily extended to estimation of inverses of Toeplitz covariance matrices. Also, an alternative version of the Whittle likelihood for the spectral density based on the discrete cosine transform is proposed.

Some key words: Discrete cosine transform; Periodogram; Spectral density; Variance-stabilizing transform; Whittle likelihood.

1. Introduction

Estimation of covariance and precision matrices is a fundamental problem in statistical data analysis with countless applications in the natural and social sciences. A special type of covariance matrix that has descending diagonal constants, known as a Toeplitz matrix, arises in the study of stationary stochastic processes. Stationary stochastic processes are an important modelling tool in many applications, such as radar target detection, speech recognition, modelling internet economic activity, electrical brain activity or the motion of crystal structures (Grenander & Szegö, 1958, p.232; Franaszczuk et al., 1985; Quah, 2000; Roberts & Ephraim, 2000; Du et al., 2020).

The data for estimation are given as n independent and identically distributed realizations of a p-dimensional vector having a zero mean and a Toeplitz covariance matrix $\Sigma = (\sigma_{|i-j|})_{i,j=1}^{p}$. Thereby, p is assumed to grow while n may be equal to 1 or may tend to infinity as well. For $p \to \infty$ and $p/n \to c \in (0, \infty]$, the sample (auto)covariance matrix is known to be an inconsistent estimator of Σ in the spectral norm; see, e.g., Wu & Pourahmadi (2009, Ch. 2) and Pourahmadi (2013). Therefore, tapering, banding and thresholding of the sample covariance matrix have been proposed to regularize this estimator; see Wu & Xiao (2012) and Cai et al. (2013). The optimal rate of convergence for Toeplitz covariance matrix estimators was established in Cai et al. (2013), who in particular showed that tapering and banding estimators attain the minimax optimal convergence rate over certain spaces of Toeplitz covariance matrices. Optimality of the thresholded estimator was shown only for the case $n = 1$; see Wu & Xiao (2012). However, all of these estimators have several practical drawbacks that affect their performance in small samples. First, additional manipulations with the estimators must be performed to enforce positive definiteness; see, e.g., § 5 of Cai et al. (2013). Second, the data-driven choice of the tapering, banding or thresholding parameter is not trivial in practice. For $n > 1$, Bickel & Levina (2008) proposed a cross-validation criterion that approximates the risk of the estimator. Fang et al. (2016) compared this method with a bootstrap-based approximation of the risk in an intensive simulation study and recommended cross-validation over bootstrap. However, for small n, a cross-validated tuning parameter turns out to be very variable, while, already for moderate n, it becomes numerically very demanding. For $n = 1$, to the best of our knowledge, there is no fully data-driven approach for selecting the banding, tapering or thresholding parameter available. Wu & Pourahmadi (2009) suggested first splitting the time series into non-overlapping subseries and then applying the cross-validation criterion of Bickel & Levina (2008). However, the appropriate choice of the subseries length is crucial for this approach, but this cannot be chosen in a data driven manner.

In this work, an alternative way to estimate a Toeplitz covariance matrix and its inverse is proposed. Our approach exploits the one-to-one correspondence between Toeplitz covariance matrices and their spectral densities. First, the given data are transformed into approximate Gaussian random variables whose mean equals the logarithm of the spectral density. Then, the log-spectral density is estimated by a periodic smoothing spline with a data-driven smoothing parameter. Finally, the resulting spectral density estimator is transformed into an estimator for Σ or its inverse. It is shown that this procedure leads to an estimator that is fully data driven, automatically positive definite and achieves the minimax optimal convergence rate under the spectral norm over a large class of Toeplitz covariance matrices. In particular, this class includes Toeplitz covariance matrices that correspond to long-memory processes with bounded spectral densities. Moreover, the computation is very efficient, does not require iterative or resampling schemes and allows application of any inference and adaptive estimation procedures developed in the context of nonparametric Gaussian regression.

Estimation of the spectral density from a single stationary time series is a research topic with a long history. Earlier nonparametric methods are based on smoothing the (log-)periodogram, which itself is not a consistent estimator (Bartlett, 1950; Welch, 1967; Wahba, 1980; Thomson, 1982). Another line of nonparametric methods for estimating the spectral density is based on the Whittle likelihood, which is an approximation to the exact likelihood of the time series in the frequency domain. For example, Pawitan & O'Sullivan (1994) estimated the spectral density from a penalized Whittle likelihood, while Kooperberg et al. (1995) used polynomial splines to estimate the log-spectral density function maximizing the Whittle likelihood. Recently, Bayesian methods for spectral density estimation have been proposed (see Choudhuri et al., 2004; Edwards et al., 2019; Maturana-Russel & Meyer, 2021), but these may become very computationally intensive in large samples due to posterior sampling.

The minimax optimal convergence rate for nonparametric estimators of a Hölder continuous spectral density from a single Gaussian stationary time series was obtained by Bentkus (1985) under the L_p norm, $1 \leqslant p \leqslant \infty$. Only a few works on spectral density estimation show the optimality of the corresponding estimators. In particular, Pawitan & O'Sullivan (1994) and Kooperberg et al. (1995) derived convergence rates of their estimators for the log-spectral density under the L_2 norm, while neglecting the Whittle likelihood approximation error.

In general, most works on spectral density estimation do not exploit further the close connection to the corresponding Toeplitz covariance matrix estimation. In particular, an upper bound for the L_∞ risk of a spectral density estimator automatically provides an upper bound for the risk of the corresponding Toeplitz covariance matrix estimator under the spectral norm. This fact is used to establish the minimax optimality of our nonparametric estimator for Toeplitz covariance matrices. The main contribution of this work is to show that our proposed spectral density estimator is not only numerically very efficient, performing excellently in small samples, but also achieves the minimax optimal rate in the L_∞ norm, which in turn ensures the minimax optimality of the corresponding Toeplitz covariance matrix estimator.

2. Set-up and diagonalization of Toeplitz matrices

Let $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}_p(0_p, \Sigma)$, where Σ is a $p \times p$ positive semidefinite covariance matrix with a Toeplitz structure, that is, $\Sigma = (\sigma_{|i-j|})_{i,j=1}^{p} \succeq 0$. The sample size n may tend to infinity or be a constant. The case $n = 1$ corresponds to a single observation of a stationary time series and in this case the data are simply denoted by $Y \sim \mathcal{N}_p(0_p, \Sigma)$. The dimension p is assumed to grow. The spectral density function f, corresponding to a Toeplitz covariance matrix Σ with absolute summable sequence of covariances $(\sigma_k)_{k\in\mathbb{Z}}$, is given by

$$f(x) = \sigma_0 + 2\sum_{k=1}^{\infty} \sigma_k \cos(kx), \qquad x \in [-\pi, \pi].$$

The inverse Fourier transform implies that

$$\sigma_k = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)\cos(kx)\,\mathrm{d}x = \int_0^1 f(\pi x)\cos(k\pi x)\,\mathrm{d}x. \tag{1}$$

If $\sum_{h=-\infty}^{\infty} |\sigma_h| = \infty$, i.e., the corresponding stochastic process is a long-memory process, the spectral density function is directly defined as a 2π-periodic, nonnegative function $f\colon [-\pi,\pi] \to \mathbb{R}_{\geqslant 0}$ that satisfies condition (1). Hence, in the case that f exists, Σ is completely characterized by f. Furthermore, the nonnegativity of the spectral density function implies the positive semidefiniteness of the covariance matrix. Moreover, the decay of the covariances σ_k is directly connected to the smoothness of f. Finally, the convergence rate of a Toeplitz covariance estimator and that of the corresponding spectral density estimator are directly related via $\|\Sigma\| \leqslant \|f\|_\infty = \sup_{x\in[-\pi,\pi]} |f(x)|$, where $\|\cdot\|$ denotes the spectral norm; see Grenander & Szegö (1958, Ch. 5.2).

As in Cai et al. (2013), we consider the class of positive semidefinite Toeplitz covariance matrices with Hölder continuous spectral densities. For $\beta = \gamma + \alpha > 0$, where $\gamma \in \mathbb{N} \cup \{0\}$, $0 < \alpha \leqslant 1$ and $0 < M_0, M_1 < \infty$, let

$$\mathcal{P}_\beta(M_0, M_1) = \{f \mid f\colon [-\pi,\pi] \to \mathbb{R}_{\geqslant 0},\ \|f\|_\infty \leqslant M_0,\ \|f^{(\gamma)}(\cdot + h) - f^{(\gamma)}(\cdot)\|_\infty \leqslant M_1|h|^\alpha\}.$$

Furthermore, for $f \in \mathcal{P}_\beta(M_0, M_1)$, we denote by $\Sigma(f) \in \mathbb{R}^{p\times p}$ the corresponding $p \times p$ Toeplitz covariance matrix obtained with the inverse Fourier transform (1) and by $\Sigma^{-1}(f)$

the precision matrix. The optimal convergence rate for estimating Toeplitz covariance matrices $\Sigma(f)$, where $f \in \mathcal{P}_\beta(M_0, M_1)$, depends crucially on β. It is well known that the kth Fourier coefficient of a function whose γth derivative is Hölder continuous with exponent $\alpha \in (0, 1]$ decays at least with $\mathcal{O}(k^{-\beta})$; see Zygmund (2002). Hence, β determines the decay rate of the covariances σ_k, which are the Fourier coefficients of the spectral density f, as $k \to \infty$. For $\beta \in (0, 1/2]$, the class $\mathcal{P}_\beta(M_0, M_1)$ includes bounded spectral densities of certain long-memory processes.

A connection between Toeplitz covariance matrices and their spectral densities is further exploited in the following lemma.

Lemma 1. *Let $\Sigma = \Sigma(f)$ with $f \in \mathcal{P}_\beta(M_0, M_1)$ and $x_j = (j-1)/(p-1)$ for $j = 1, \dots, p$. Then*

$$(D^{\mathrm{T}}\Sigma D)_{i,j} = f(\pi x_j)\delta_{i,j} + \frac{1 + (-1)^{|i-j|}}{2}\mathcal{O}(p^{-1} + p^{-\beta}\log p),$$

where $\delta_{i,j}$ is the Kroneker delta, $\mathcal{O}(\cdot)$ terms are uniform over $i, j = 1, \dots, p$ and

$$D = \left(\frac{2}{p-1}\right)^{1/2}\left[\cos\left\{\pi(i-1)\frac{j-1}{p-1}\right\}\right]_{i,j=1}^{p},$$

divided by $2^{1/2}$ when i or j is 1 or p, is the discrete cosine transform I matrix.

The proof can be found in the Supplementary Material. This result shows that the discrete cosine transform I matrix approximately diagonalizes Toeplitz covariance matrices and that the diagonalization error depends to some extent on the smoothness of the corresponding spectral density.

In time series analysis the discrete Fourier transform matrix $F = p^{-1/2}\{\exp(2\pi \mathrm{i} ij/p)\}_{i,j=1}^{p}$, where i is the imaginary unit, is typically employed to approximately diagonalize Toeplitz covariance matrices. Using the fact that $(F^{\mathrm{T}}\Sigma F)_{i,i} = f(2\pi i/p) + o(1)$, Whittle (1957) introduced an approximation for the likelihood of a single Gaussian stationary time series ($n = 1$ case), the so-called Whittle likelihood:

$$\mathcal{L}(Y \mid f) \propto \exp\left\{-\sum_{j=1}^{\lfloor p/2 \rfloor}\log f\left(\frac{2\pi j}{p}\right) + \frac{I_j}{f(2\pi j/p)}\right\}. \tag{2}$$

The quantity $I_j = |F_j^{\mathrm{T}}Y|^2$, where F_j denotes the jth column of F, is known as the periodogram at the jth Fourier frequency. Because of periodogram symmetry, only $\lfloor p/2 \rfloor$ data points $I_1, \dots, I_{\lfloor p/2 \rfloor}$ are available for estimating the mean $f(2\pi j/p)$, $j = 1, \dots, \lfloor p/2 \rfloor$, where $\lfloor x \rfloor$ denotes the largest integer strictly smaller than x. The Whittle likelihood has become a popular tool for parameter estimation of stationary time series, e.g., for nonparametric and parametric spectral density estimation or for estimation of the Hurst exponent; see, e.g., Walker (1964) and Taqqu & Teverovsky (1997).

Lemma 1 yields the following alternative version of the Whittle likelihood:

$$\mathcal{L}(Y \mid f) \propto \exp\left\{-\sum_{j=1}^{p}\log f(\pi x_j) + \frac{W_j}{f(\pi x_j)}\right\}. \tag{3}$$

Here $W_j = (D_j^{\mathrm{T}} Y)^2$ with D_j denoting the jth column of D. This likelihood approximation is based on twice as many data points W_j as the standard Whittle likelihood. Thus, it allows for a more efficient use of the data Y to estimate the parameter of interest, such as the spectral density or the Hurst parameter. This is particularly advantageous in small samples.

Equations (2) and (3) invite the estimation of f by maximizing the (penalized) likelihood over certain linear spaces, e.g., spline spaces, as suggested by Pawitan & O'Sullivan (1994) and Kooperberg et al. (1995). However, such an approach requires well-designed numerical methods to solve the corresponding optimization problem, since the spectral density in the second term of (2) and (3) is in the denominator, which hinders derivation of a closed-form expression for the estimator and often leads to numerical instabilities. Also, the choice of smoothing parameter becomes challenging.

Therefore, we suggest an alternative approach that allows the spectral density to be estimated as a mean in an approximate Gaussian regression. Such estimators have a closed-form expression, do not require an iterative optimization algorithm and a smoothing parameter can be easily obtained with any conventional criterion. First if $Y \sim \mathcal{N}_p(0_p, \Sigma)$, with $\Sigma = \Sigma(f)$ and $f \in \mathcal{P}_\beta(M_0, M_1)$, then $D^{\mathrm{T}} Y \sim \mathcal{N}_p(0_p, D^{\mathrm{T}} \Sigma D)$. Hence, for $W_j = (D_j^{\mathrm{T}} Y)^2$, $j = 1, \ldots, p$, it follows with Lemma 1 that

$$W_j \sim \Gamma\{1/2, 2f(\pi x_j) + \mathcal{O}(p^{-1} + p^{-\beta} \log p)\}, \tag{4}$$

where $\Gamma(a, b)$ denotes the gamma distribution with shape parameter a and scale parameter b. The random variables $W_1, \ldots, W_p$ are only asymptotically independent. Obviously, $E(W_j) = f(\pi x_j) + o(1)$ for $j = 1, \ldots, p$. To estimate f from $W_1, \ldots, W_p$, one could use a generalized nonparametric regression framework with a gamma-distributed response; see, e.g., the classical monograph by Hastie & Tibshirani (1990). However, this approach requires an iterative procedure for estimation, e.g., a Newton–Raphson algorithm, with a suitable choice for the smoothing parameter at each iteration step. Deriving the L_∞ rate for the resulting estimator is also not a trivial task. Instead, we suggest employing a variance-stabilizing transform of Cai & Zhou (2010) that converts a gamma regression into an approximate Gaussian regression. In the next section we present the methodology in more detail for a general setting with $n \geqslant 1$.

3. Methodology

Let $L_\delta = \{f\colon \inf_x f(x) \geqslant \delta\}$ for some $\delta > 0$ and set $\mathcal{F}_\beta = \mathcal{P}_\beta(M_0, M_1) \cap L_\delta$. We consider estimation of Σ and $\Omega = \Sigma^{-1}$ from a sample $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}_p(0_p, \Sigma)$, where $\Sigma = \Sigma(f)$ with $f \in \mathcal{F}_\beta$. For $Y_i \sim \mathcal{N}_p(0_p, \Sigma)$, $i = 1, \ldots, n$, it was shown in the previous section that, with Lemma 1, the data can be transformed into gamma-distributed random variables $W_{i,j} = (D_j^{\mathrm{T}} Y_i)^2$, $i = 1, \ldots, n; j = 1, \ldots, p$, where, for each fixed i, the random variable $W_{i,j}$ has the same distribution as W_j given in (4). Now the approach of Cai & Zhou (2010) is adapted to the setting $n \geqslant 1$.

First, the transformed data points $W_{i,j}$ are binned, that is, fewer new variables Q_k, $k = 1, \ldots, T$, with $T < p$, are built via $Q_k = \sum_{j=(k-1)p/T+1}^{kp/T} \sum_{i=1}^{n} W_{i,j}$ for $k = 1, \ldots, T$. The number of observations in a bin is $m = np/T$. In Theorem 1 in §4, we show that setting $T = \lfloor p^\upsilon \rfloor$ for any $\upsilon \in (1 - \min\{1, \beta\}/3, 1)$ leads to the minimax optimal rate for the spectral density estimator. To simplify the notation, m is handled as an integer; otherwise, one can discard several observations in the last bin. Next, applying the variance-stabilizing transform $G(x) = 2^{-1/2}\log(x/m)$ to each Q_k yields new random variables $Y_k^* = 2^{-1/2}\log(Q_k/m)$ that are approximately Gaussian, as shown by Cai & Zhou (2010). Since the spectral density is a function that is symmetric around zero and periodic on $[-\pi, \pi]$, one can mirror the resulting observations to use $Y_T^*, \ldots, Y_2^*, Y_1^*, \ldots, Y_{T-1}^*$ for estimation. Renumerating the observations Y_k^* and scaling the design points into the interval $[0, 1)$ for convenience leads to the approximate Gaussian regression problem

$$Y_k^* \overset{\text{approx.}}{\sim} \mathcal{N}[H\{f(x_k)\}, m^{-1}], \qquad x_k = \frac{k-1}{2T-2};\ k = 1, \ldots, 2T-2,$$

where $H(y) = 2^{-1/2}\{\phi(m/2) + \log(2y/m)\}$ and ϕ is the digamma function (Cai & Zhou, 2010). Now, the scaled and shifted log-spectral density $H(f)$ can be estimated with a periodic smoothing spline

$$\widehat{H(f)}(x) = \underset{s \in S_{\text{per}}(2q-1)}{\arg\min} \left[\frac{1}{2T-2} \sum_{k=1}^{2T-2} \{Y_k^* - s(x_k)\}^2 + h^{2q} \int_0^1 \{s^{(q)}(x)\}^2 \, \mathrm{d}x \right], \tag{5}$$

where $h > 0$ denotes a smoothing parameter, $q \in \mathbb{N}$ is the penalty order and $S_{\text{per}}(2q-1)$ is a space of periodic splines of degree $2q-1$. More details on periodic smoothing splines can be found in the Supplementary Material.

Once an estimator $\widehat{H(f)}$ is obtained, application of the inverse transform function $H^{-1}(y) = m\exp\{2^{1/2}y - \phi(m/2)\}/2$ yields the spectral density estimator $\hat{f} = H^{-1}\{\widehat{H(f)}\}$. Finally, the inverse Fourier transform leads to the covariance matrix estimator

$$\hat{\Sigma} = (\hat{\sigma}_{|i-j|})_{i,j=1}^p \quad \text{with } \hat{\sigma}_k = \int_0^1 \hat{f}(x)\cos(k\pi x)\, \mathrm{d}x \text{ for } k = 0, \ldots, p-1. \tag{6}$$

The precision matrix Ω is estimated by the inverse Fourier transform of the reciprocal of the spectral density estimator, i.e.,

$$\hat{\Omega} = (\hat{\omega}_{|i-j|})_{i,j=1}^p \quad \text{with } \hat{\omega}_k = \int_0^1 \hat{f}(x)^{-1}\cos(k\pi x)\, \mathrm{d}x \text{ for } k = 0, \ldots, p-1. \tag{7}$$

The estimation procedure for $\hat{\Sigma}$ and $\hat{\Omega}$ can be summarized as follows.

Step 1 (*Data transformation*). Define $W_{i,j} = (D_j^{\mathrm{T}} Y_i)^2$, $i = 1, \ldots, n; j = 1, \ldots, p$, where D is the $p \times p$ discrete cosine transform I matrix as given in Lemma 1 and D_j is its jth column.

Step 2 (*Binning*). Set $T = \lfloor p^{\upsilon} \rfloor$ for any $\upsilon \in (1 - \min\{1, \beta\}/3, 1)$ and calculate

$$Q_k = \sum_{j=(k-1)p/T+1}^{kp/T} \sum_{i=1}^{n} W_{i,j}, \qquad k = 1, \ldots, T.$$

Step 3 (*Variance-stabilizing transform*). Set $Y_k^* = 2^{-1/2}\log(Q_k/m)$ for $k = 1, \ldots, T$ and $m = np/T$. Mirror the data to get approximately $2T - 2$ Gaussian random variables $Y_T^*, \ldots, Y_2^*, Y_1^*, \ldots, Y_{T-1}^*$.

Step 4 (*Gaussian regression*). Renumerate observations Y_k^*, scale the design points to [0, 1) and estimate $H(f)$ with a periodic smoothing spline of degree $2q-1$ in an approximate Gaussian regression model

$$Y_k^* = H\{f(x_k)\} + \epsilon_k, \qquad x_k = \frac{k-1}{2T-2};\ k = 1, \ldots, 2T-2,$$

where the ϵ_k are asymptotically independent and identically distributed Gaussian variables.

Step 5 (*Inverse transform*). Estimate the spectral density f with $\hat{f} = H^{-1}\{\widehat{H(f)}\}$, where

$$H^{-1}(y) = m \exp\{2^{1/2}y - \phi(m/2)\}/2$$

for a digamma function ϕ.

Step 6 (*Estimators*). Set $\hat{\Sigma} = (\hat{\sigma}_{|i-j|})_{i,j=1}^p$ with $\hat{\sigma}_k = \int_0^1 \hat{f}(x)\cos(k\pi x)\,\mathrm{d}x$ and $\hat{\Omega} = (\hat{\omega}_{|i-j|})_{i,j=1}^p$ with $\hat{\omega}_k = \int_0^1 \hat{f}(x)^{-1}\cos(k\pi x)\,\mathrm{d}x$ $(k = 0, \ldots, p-1)$.

The estimators $\hat{\Sigma}$ and $\hat{\Omega}$ are positive definite matrices by construction, since the spectral density estimator $\hat{f}$ is nonnegative by definition. For a detailed discussion on the choice of all parameters needed to obtain our estimators, see § 5.

4. Theoretical properties

In this section, we study the asymptotic properties of the estimators $\hat{f}$, $\hat{\Sigma}$ and $\hat{\Omega}$. Let $\hat{f} = m\exp\{2^{1/2}\widehat{H(f)} - \phi(m/2)\}/2$ be the spectral density estimator defined in § 3, where $\widehat{H(f)}$ is given in (5), $m = np/T$ and ϕ is the digamma function. Furthermore, let $\hat{\Sigma}$ be the Toeplitz covariance matrix estimator and $\hat{\Omega}$ the corresponding precision matrix defined in (6) and (7), respectively. The following theorem shows that both $\hat{\Sigma}$ and $\hat{\Omega}$ attain the minimax optimal rate of convergence over the class of Toeplitz matrices $\Sigma(f)$ such that $f \in \mathcal{F}_\beta$, $\beta > 0$.

Theorem 1. *Let* $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}_p(0_p, \Sigma)$, $n \geqslant 1$, *with* $\Sigma = \Sigma(f)$ *such that* $f \in \mathcal{F}_\beta$ *and* $\beta = \gamma + \alpha > 0$. *If* $h > 0$ *such that* $h \to 0$ *and* $hT \to \infty$, *then, with* $T = \lfloor p^\upsilon \rfloor$ *for any* $\upsilon \in (1 - \min\{1, \beta\}/3, 1)$ *and* $q = \max\{1, \gamma\}$, *the spectral density estimator* $\hat{f}$, *the corresponding covariance matrix estimator* $\hat{\Sigma}$ *and the precision matrix estimator* $\hat{\Omega}$ *satisfy, for* $p \to \infty$ *and* n *such that* $p^{\min\{1,\beta\}}/n \to c \in (0, \infty]$,

$$\sup_{f\in\mathcal{F}_\beta} E_f\|\hat{\Sigma} - \Sigma(f)\|^2 \leqslant \sup_{f\in\mathcal{F}_\beta} E_f\|\hat{f} - f\|_\infty^2 = \mathcal{O}\left\{\frac{\log(np)}{nph}\right\} + \mathcal{O}(h^{2\beta}),$$

$$\sup_{f\in\mathcal{F}_\beta} E_f\|\hat{\Omega} - \Sigma^{-1}(f)\|^2 = \mathcal{O}\left\{\frac{\log(np)}{nph}\right\} + \mathcal{O}(h^{2\beta}).$$

For $h \asymp \{\log(np)/(np)\}^{1/(2\beta+1)}$, *it follows that*

$$\sup_{f\in\mathcal{F}_\beta} E_f\|\hat{\Sigma} - \Sigma(f)\|^2 \leqslant \sup_{f\in\mathcal{F}_\beta} E_f\|\hat{f} - f\|_\infty^2 = \mathcal{O}\left[\left\{\frac{\log(np)}{np}\right\}^{2\beta/(2\beta+1)}\right],$$

$$\sup_{f\in\mathcal{F}_\beta} E_f\|\hat{\Omega} - \Sigma^{-1}(f)\|^2 = \mathcal{O}\left[\left\{\frac{\log(np)}{np}\right\}^{2\beta/(2\beta+1)}\right].$$

Theorem 1 is established under the asymptotic scenario with $p \to \infty$ and n such that $p^{\min\{1,\beta\}}/n \to c \in (0,\infty]$, i.e., the dimension p grows, while the sample size n either remains fixed or also grows, but not faster than $p^{\min\{1,\beta\}}$. This asymptotic scenario covers the setting when the sample covariance matrix is inconsistent. In particular, for $\beta \geqslant 1$, the sample size n does not grow faster than p and adding more samples improves the convergence rate. If $\beta \in (0,1)$ then increasing n with the rate faster than p^{β} will not lead to a faster convergence rate, due to the diagonalization error from Lemma 1, which can be improved only by making additional assumptions on the spectral density.

The minimax optimal convergence rates for estimating Σ and Σ^{-1} from n independent and identically distributed Gaussian vectors $Y_1, \ldots, Y_n$ with zero mean and a Toeplitz covariance matrix $\Sigma(f)$ with $f \in \mathcal{F}_\beta$ have been established by Cai et al. (2013). Since the lower bound rates given in Theorems 5 and 7 of Cai et al. (2013) match the upper bound rates obtained in Theorem 1, we conclude that our estimator is minimax optimal. For non-Gaussian data, the minimax optimal convergence rates for Σ and Σ^{-1} are not known. Note that $\mathcal{F}_\beta$ with $\beta > 0$ includes bounded spectral densities of long-memory processes.

The proof of Theorem 1 can be found in the Appendix and is the main result of our work. The most important part of this proof is the derivation of the convergence rate for the spectral density estimator $\hat{f}$ under the L_∞ norm. Cai & Zhou (2010) established an L_2 rate for a wavelet nonparametric mean estimator in a gamma regression where the data are assumed to be independent. In our work, the spectral density estimator $\hat{f}$ is based on the gamma-distributed data $W_{i,1}, \ldots, W_{i,p}$, which are only asymptotically independent. Moreover, the mean of these data is not exactly $f(\pi x_1), \ldots, f(\pi x_p)$, but is corrupted by the diagonalization error given in Lemma 1. This error adds to the error that arises via binning and variance stabilizing transformation and that describes the deviation from a Gaussian distribution, as derived by Cai & Zhou (2010). Finally, we need to obtain an L_∞ rather than an L_2 rate for our spectral density estimator. Overall, the proof requires different and partly novel tools than those used by Cai & Zhou (2010). A particular challenge is the treatment of the dependence of $W_{i,1}, \ldots, W_{i,p}$.

To get the L_∞ rate for $\hat{f}$, we first derive that for the periodic smoothing spline estimator $\widehat{H(f)}$ of the log-spectral density. To do so, we use a closed-form expression of its effective kernel obtained by Schwarz & Krivobokova (2016), thereby carefully treating various (dependent) errors that describe deviations from a Gaussian nonparametric regression with independent errors and mean $f(\pi x_i)$. Although the periodic smoothing spline estimator is obtained on T binned points, the rate is given in terms of the vector dimension p and the sample size n. Next, using the Cauchy–Schwarz inequality and a mean value argument, this rate is translated into the L_∞ rate for the spectral density estimator $\hat{f}$. To obtain the rate for the Toeplitz covariance matrix estimator, it is enough to note that $E\|\hat{\Sigma} - \Sigma\|^2 \leqslant E\|\hat{f} - f\|_\infty^2$.

5. Practical issues

Several choices must be made in practice to obtain our estimator. First, the data Y_i, $i = 1, \ldots, n$, are transformed into $W_{i,j} = (D_j^{\mathrm{T}} Y_i)^2$, $j = 1, \ldots, p$, and are subsequently binned. According to Theorem 1, the number of bins $T = \lfloor p^{\upsilon} \rfloor$ with any $\upsilon \in (1 - \min\{\beta, 1\}/3, 1)$ leads to a minimax optimal estimator. That is, for $\beta \geqslant 1$, one can take any $\upsilon \in (2/3, 1)$, while for $\beta < 1$, the interval depends on β; for example, for $\beta = 1/2$, the interval is $\upsilon \in (5/6, 1)$. In our simulation studies we observed that the results are quite robust for various values of υ.

If no knowledge about β is available, one can proceed as follows. Any standard test for long-range dependence can be performed and if the null hypothesis of long-range dependence is rejected, i.e., $\beta > 1/2$, then any $\upsilon \in (5/6, 1)$ can be taken. Otherwise, a smaller interval for υ should be considered. Additionally, one can always verify whether the chosen value of T is appropriate by generating quantile–quantile (Q-Q) plots of Y_k^*, which should ideally show little departure from the normality. In Fig. 1 in the Supplementary Material we show how the Q-Q plots change for Gaussian data depending on T.

Once the data are transformed, the mean of Y_k^*, $k = 1, \ldots, T$, i.e., the log-spectral density, is estimated with a periodic smoothing spline. For this, one needs to choose basis functions of the periodic spline space, the penalty order $q \in \mathbb{N}$ and the smoothing parameter $h > 0$.

The basis of a periodic spline space with knots put at the observations is a Fourier basis $\{2^{1/2}\cos(2\pi x), 2^{1/2}\sin(2\pi x)\}$, evaluated at $x_k = (k-1)/(2T-2)$ for $k = 1, \ldots, 2T-2$.

The smoothing parameter h can be chosen using any data-driven approach, such as (generalized) cross-validation or an empirical Bayes approach; see Wahba (1985).

According to Theorem 1, the choice of q should be related to the true smoothness β of the spectral density in order to obtain the minimax optimal estimator. Assume that q taken for estimation is larger than the true smoothness β. Then, the rate of convergence is determined by the true β and is minimax, independent of how large q is. In particular, if $\gamma = 0$ then $\beta = \alpha \in (0, 1)$ and taking $q = \max\{1, \gamma\} = 1$ would lead to a minimax optimal estimator. If q taken for the estimation is less than β then the rate will depend on q and on the choice of the smoothing parameter h. Assume that $\beta > q$ and that the smoothing parameter h is estimated with generalized cross-validation. Then, it has been shown by Wahba (1985) that a periodic spline estimator adapts to the unknown smoothness up to $2q$. That is, if $\beta < 2q$ then the rate will be minimax optimal, while for $\beta \geqslant 2q$, the rate will be determined by q. If $\beta > q$ and the smoothing parameter h is estimated by the empirical Bayesian approach, then Wahba (1985) has shown that the resulting estimator does not adapt to the extra smoothness. However, in small samples the empirical Bayes smoothing spline can perform similarly or even better than the cross-validated smoothing spline due to smaller constants in the risk; see Krivobokova (2013).

Since, in practice, β is not known exactly, it is also not known which rate the estimator will have with a chosen q. However, looking at the decay of the sample covariance functions one can get an idea of whether β is rather small, in the case of a very slow decay, or rather large, in the case of a very fast decay, and decide on the choice of q. A more attractive approach is to resort to adaptive estimation methods that lead to the best possible estimators without prior knowledge on β. For example, the empirical Bayesian framework allows estimation of the unknown γ under certain assumptions on the function space (Serra & Krivobokova, 2017). Other approaches for adaptive nonparametric estimation are aggregation and Lepski's method; see, e.g., Chagny (2016) for an overview and references. A detailed study of adaptive spectral density estimators is outside the scope of our work.

In our simulation study and the real data example, presented in the next two sections, we discuss the parameter choices explicitly.

6. Simulation study

In this section, we compare the performance of our proposed Toeplitz covariance estimator with the tapering estimator of Cai et al. (2013) and with the sample covariance matrix. A Monte Carlo simulation with 100 samples is performed using R (version 4.1.2, seed 42;

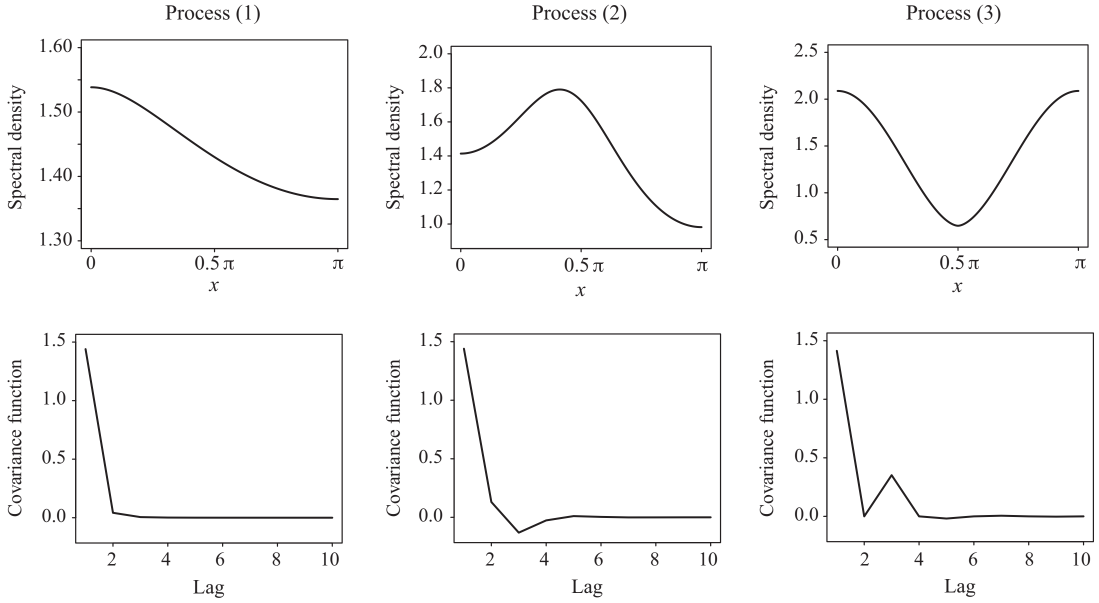

Fig. 1. Spectral density functions (first row) and covariance functions (second row) for processes (i)–(iii).

R Development Core Team, 2024). We consider Gaussian vectors $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}_p(0_p, \Sigma)$ with (A) $p = 5000$, $n = 1$, (B) $p = 1000$, $n = 50$ and (C) $p = 5000$, $n = 10$, and with the covariance functions $\sigma : \mathbb{Z} \to \mathbb{R}$, $k \mapsto \sigma_k$

(i) of a polynomial decay, i.e., $\sigma_k = 1.44(1 + |k|)^{-5.1}$,
(ii) of an autoregressive process $y_t = \epsilon_t + 0.1y_{t-1} - 0.1y_{t-2}$, where ϵ_t is independent and identically distributed Gaussian noise and $\text{var}(\epsilon_t) = 1.44$,
(iii) such that the corresponding spectral density is Lipschitz continuous, but not differentiable: $f(x) = 1.44\{|\sin(x + 0.5\pi)|^{1.7} + 0.45\}$.

In particular, $\sigma_0 = 1.44$ for all three processes. Figure 1 shows the spectral densities and the corresponding covariance functions for the three processes. To set the parameters, note that the spectral density of process (i) belongs to $\mathcal{F}_4$, the spectral density of process (ii) is an analytic function and the spectral density of process (iii) is from $\mathcal{F}_1$.

Since, for all three processes, $\beta \geqslant 1$, according to Theorem 1, any value of $\upsilon \in (2/3, 1)$ should lead to optimal estimation. We set the number of bins to $T = 500$ in all scenarios, which corresponds to $\upsilon \approx 0.73$ for scenarios (A) and (C) and $\upsilon \approx 0.90$ for scenario (B).

Setting $q = 2$ would lead to the minimax optimal rates for processes (1) and (3), if the smoothing parameter is chosen by generalized cross-validation; see the discussion in § 5. According to arguments from the previous section, the convergence rate for process (2) is the same as that for process (i), while process (iii) has a slower convergence rate. Hence, for a given scenario, we should observe similar magnitudes of the average norms for processes (i) and (ii), while somewhat larger values for process (iii). Across scenarios, we expect scenario (A) to have larger average norm values, as only $p = 5000$ observations are used, compared to $np = 50\,000$ data points in scenarios (B) and (C).

To select the regularization parameter for our estimator, we implemented the restricted maximum likelihood method, generalized cross-validation and the corresponding oracle versions, i.e., as if Σ were known.

The sample covariance matrix $\widetilde{\Sigma} = (\tilde{\sigma}_{|i-j|})_{i,j=1}^{p}$ is defined as $n^{-1}\sum_{i=1}^{n} Y_i Y_i^{\mathrm{T}}$ with averaged diagonals to obtain the Toeplitz structure. The tapering estimator with tapering parameter $k \leqslant p/2$ is defined as $\mathrm{Tap}_k(\widetilde{\Sigma}) = (\tilde{\sigma}_{|i-j|} w_{|i-j|})_{i,j=1}^{p}$, where $w_m = 1$ when $m = 0, \ldots, k/2$, $w_m = 2 - 2m/k$ when $k/2 < m \leqslant k$ and $w_m = 0$ otherwise; see Cai et al. (2013). Parameter k can be selected using cross-validation, see Bickel & Levina (2008) only if $n > 1$, that is, under scenarios (B) and (C). For this, the n observations are divided by 30 random splits into a training set of size $n_1 = 2n/3$ and a test set of size $n_2 = n/3$. Let Σ_1^{ν} and Σ_2^{ν} be the sample covariance matrices from the νth split. The tapering parameter k is then estimated as

$$\hat{k}_{\mathrm{cv}} = \arg \min_{k=2,3,\ldots,p/2} \frac{1}{30} \sum_{\nu=1}^{30} \|\mathrm{Tap}_k(\Sigma_1^{\nu}) - \Sigma_2^{\nu}\|,$$

where $\mathrm{Tap}_k(\Sigma_2^{\nu})$ is the tapering estimator with parameter k. If $n = 1$, that is, under scenario (A), Wu & Pourahmadi (2009) suggested splitting the time series Y into l non-overlapping subseries of length p/l and then proceeding as before to select the tuning parameter k. To the best of our knowledge, there is no data-driven method for the selection of l. Using the true covariance matrix Σ, we preselected oracle value $l = 30$ subseries for process (i) and $l = 15$ subseries for processes (ii) and (iii). Parameter k can then be chosen with cross-validation as above. We employ this approach under scenario (A) instead of an unavailable fully data-driven criterion and name it semi-oracle. Finally, for all three scenarios (A), (B) and (C), the oracle tapering parameter is computed using grid search for each Monte Carlo sample as $\hat{k}_{\mathrm{or}} = \arg\min_{k=2,3,\ldots,p/2} \|\mathrm{Tap}_k(\widetilde{\Sigma}) - \Sigma\|$, where $\widetilde{\Sigma}$ is the sample covariance matrix. To speed up the computation, one can replace the spectral norm by the ℓ_1 norm, as suggested by Bickel & Levina (2008).

In Table 1, the errors of the Toeplitz covariance estimators with respect to the spectral norm and the average computation time for one Monte Carlo sample for all three processes are reported for scenarios (A), (B) and (C), respectively. To illustrate the goodness of fit of the spectral density, the L_2 norm $\|\hat{f} - f\|_2$ is also computed.

The overall behaviour of our estimator is exactly as expected. Moreover, the tapering and our estimator perform similarly in terms of the spectral norm risk. This is not surprising as both estimators are proved to be rate optimal. The oracle estimators show similar behaviour, but are slightly less variable compared to the data-driven estimators. Clearly, both the tapering and our estimators are superior to the inconsistent sample covariance matrix. In terms of computational time, both methods are similarly fast for scenarios (A) and (B). For scenario (C), the tapering method is much slower due to the multiple high-dimensional matrix multiplications in the cross-validation method. It is anticipated that, for larger p, the tapering estimator will be much more computationally intensive compared to our method.

7. Application to non-Gaussian data

While we consider the rigorous theoretical study of our estimator for non-Gaussian data to be out of scope of this work, in this section we would like to discuss the application of our method to non-Gaussian data in practice.

To apply our method, one needs to ensure that the transformed data Y_k^* are approximately Gaussian with mean $H\{f(x_k)\}$. Our estimator will be minimax optimal if the deviation from Gaussianity of Y_k^* is sufficiently small. In general, one needs to ensure that

Table 1. *Errors of the Toeplitz covariance matrix and the spectral density estimators with respect to the spectral and L_2 norms. The average computation time of the covariance estimators given in seconds for one Monte Carlo sample is reported in the last column; all numbers are multiplied by* 100 *except for the last column*

Scenario (A): $p = 5000, n = 1$

	Process (i)		Process (ii)		Process (iii)		Time
	$\lVert\hat{\Sigma} - \Sigma\rVert^2$	$\lVert\hat{f} - f\rVert_2^2$	$\lVert\hat{\Sigma} - \Sigma\rVert^2$	$\lVert\hat{f} - f\rVert_2^2$	$\lVert\hat{\Sigma} - \Sigma\rVert^2$	$\lVert\hat{f} - f\rVert_2^2$	(s)
Our method (GCV)	0.688	0.255	1.591	0.439	3.401	0.606	4.235
Our method (ML)	0.591	0.224	1.559	0.417	3.747	0.628	4.224
Tapering (semi-oracle)	0.558	0.216	2.325	0.674	3.551	0.979	4.617
Sample covariance	16 240.895	3810.680	20 291.809	3694.392	24 438.036	3809.486	0.342
Our method (GCV oracle)	0.421	0.175	1.373	0.378	3.321	0.575	
Our method (ML oracle)	0.464	0.186	1.487	0.391	3.781	0.624	
Tapering (oracle)	0.418	0.171	1.045	0.288	1.547	0.371	

Scenario (B): $p = 1000, n = 50$

	Process (i)		Process (i)		Process (iii)		Time
	$\lVert\hat{\Sigma} - \Sigma\rVert^2$	$\lVert\hat{f} - f\rVert_2^2$	$\lVert\hat{\Sigma} - \Sigma\rVert^2$	$\lVert\hat{f} - f\rVert_2^2$	$\lVert\hat{\Sigma} - \Sigma\rVert^2$	$\lVert\hat{f} - f\rVert_2^2$	(s)
Our method (GCV)	0.100	0.028	0.205	0.050	0.531	0.082	27.194
Our method (ML)	0.076	0.024	0.230	0.051	0.611	0.089	27.098
Tapering (CV)	0.110	0.031	0.218	0.055	0.348	0.073	23.908
Sample covariance	79.603	56.262	95.090	61.000	127.528	61.001	0.141
Our method (GCV oracle)	0.062	0.020	0.163	0.043	0.462	0.074	
Our method (ML oracle)	0.067	0.021	0.221	0.050	0.603	0.088	
Tapering (oracle)	0.057	0.020	0.133	0.037	0.265	0.055	

Scenario (C): $p = 5000, n = 10$

	Process (i)		Process (ii)		Process (iii)		Time
	$\lVert\hat{\Sigma} - \Sigma\rVert^2$	$\lVert\hat{f} - f\rVert_2^2$	$\lVert\hat{\Sigma} - \Sigma\rVert^2$	$\lVert\hat{f} - f\rVert_2^2$	$\lVert\hat{\Sigma} - \Sigma\rVert^2$	$\lVert\hat{f} - f\rVert_2^2$	(s)
Our method (GCV)	0.088	0.026	0.217	0.050	0.593	0.079	4.260
Our method (ML)	0.078	0.023	0.231	0.050	0.677	0.086	4.251
Tapering (CV)	0.143	0.034	0.217	0.051	0.422	0.070	635.345
Sample covariance	673.122	370.946	792.714	360.687	1 014 071.587	375.728	1.189
Our method (GCV oracle)	0.062	0.020	0.172	0.043	0.500	0.071	
Our method (ML oracle)	0.069	0.021	0.224	0.048	0.663	0.085	
Tapering (oracle)	0.055	0.018	0.147	0.039	0.257	0.051	

GCV, generalized cross-validation; ML, restricted maximum likelihood method.

(i) the mean of $(D_j^{\mathrm{T}} Y_i)^2$ is the spectral density up to a negligible error $f(\pi x_j)\{1 + o(1)\}$,
(ii) the central limit theorem is applicable to Q_k after appropriate centring and scaling,
(iii) the log-transform is variance stabilizing for $(D_j^{\mathrm{T}} Y_i)^2$.

The first point is always satisfied, as long as both moments of $D_j^{\mathrm{T}} Y_i$ exist. Indeed, suppose that Y_i has a Toeplitz covariance matrix Σ and that the marginal distribution is non-Gaussian with mean zero. Then, $E\{(D_j^{\mathrm{T}} Y_i)^2\} = (D^{\mathrm{T}} \Sigma D)_{jj} = f(\pi x_j)\{1 + o(1)\}$. The last two points can be checked explicitly, if the distribution of Y_i is known. For example, for gamma-distributed data, the second point is clearly satisfied and the third point is proved in the Supplementary Material.

Of course, in practice, the distribution of Y_i is rarely known. Since our method relies on the asymptotic normality of Y_k^*, one can simply check whether the data deviate from normality strongly. Application of normality tests for Y_k^* might be misleading in small samples, since the Y_k^* are only asymptotically Gaussian, even when the Y_i are Gaussian. Therefore, we suggest generating a Q-Q plot of Y_k^*. If this Q-Q plot shows little deviations from Gaussianity then our method can be safely applied in practice. If deviations from normality are substantial, this might indicate that the log-transform of the data is not suitable. To find an appropriate variance-stabilizing transformation, one can employ a Box–Cox transform. To estimate the Box–Cox transformation parameter, one must take into account correlation of the binned data, e.g., by using the method of Guerrero (1993) developed for time series.

In the Supplementary Material we provide a small simulation study, as well as several Q-Q plots of Y_k^*, where the distribution of $\Sigma^{-1/2}Y_i$ was taken to be gamma and uniform. It can be observed that in all examples the Q-Q plots show little departure from the Gaussian distribution and all simulation results look very similar, independent of the distribution of Y_i.

8. Application to protein dynamics

We revisit the data analysis of protein dynamics performed by Krivobokova et al. (2012) and Singer et al. (2016). We consider data generated by the molecular dynamics simulations for the yeast aquaporin, the gated water channel of the yeast *Pichi pastoris*. Molecular dynamics simulations are an established tool for studying biological systems at the atomic level on timescales of nano- to microseconds. The data are given as Euclidean coordinates of all 783 atoms of the aquaporin observed in a 100-ns time frame, split into 20 000 equidistant observations. Additionally, the diameter of the channel y_t at time t is given, measured by the distance between two centres of mass of certain residues of the protein. The aim of the analysis is to identify the collective motions of the atoms responsible for the channel opening. In order to model the response variable y_t, which is a distance, based on the motions of the protein atoms, we chose to represent the protein structure by distances between atoms and certain fixed base points instead of Euclidean coordinates. That is, we calculated

$$X_t = \{d(A_{t,1}, B_1), \ldots, d(A_{t,783}, B_1), d(A_{t,1}, B_2), \ldots, d(A_{t,783}, B_y)\} \in \mathbb{R}^{4\cdot 783},$$

where $A_{t,i} \in \mathbb{R}^3$, $i = 1, \ldots, 783$, denotes the ith atom of the protein at time t, $B_j \in \mathbb{R}^3$, $j = 1, \ldots, 4$, is the jth base point and $d(\cdot, \cdot)$ is the Euclidean distance. Figure 2 shows the diameter y_t and the distance between the first atom and the first centre of mass.

It can be concluded that a linear model $Y = Xb + \epsilon$ holds, where $Y = (y_1, \ldots, y_{20\,000})^{\mathrm{T}}$, $X = (X_1^{\mathrm{T}}, \ldots, X_{20\,000}^{\mathrm{T}})^{\mathrm{T}}$, $b \in \mathbb{R}^{4\cdot 783}$, $\epsilon \in \mathbb{R}^{20\,000}$. This linear model has two specific features that are intrinsic to the problem: first, the observations are not independent over time and, second, X_t is high dimensional at each t and only few columns of X are relevant for Y. Krivobokova et al. (2012) have shown that the partial least-squares, PLS, algorithm performs exceptionally well on this type of data, leading to a small dimensional and robust representation of proteins, which is able to identify the atomic dynamics relevant for Y. Singer et al. (2016) studied the convergence rates of the PLS algorithm for dependent observations and showed that decorrelating the data before running the PLS algorithm improves its performance. Since Y is a linear combination of columns of X, it can be assumed that Y and all columns of X have the same correlation structure. Hence, it is sufficient to estimate $\Sigma = \mathrm{cov}(Y)$ to decorrelate the data for the PLS algorithm, i.e., $\Sigma^{-1/2}Y = \Sigma^{-1/2}Xb + \Sigma^{-1/2}\epsilon$ results in a standard linear regression with independent errors.

Our goal now is to estimate Σ and compare the performance of the PLS algorithm on original and decorrelated data. For this purpose, we divided the dataset into a training and

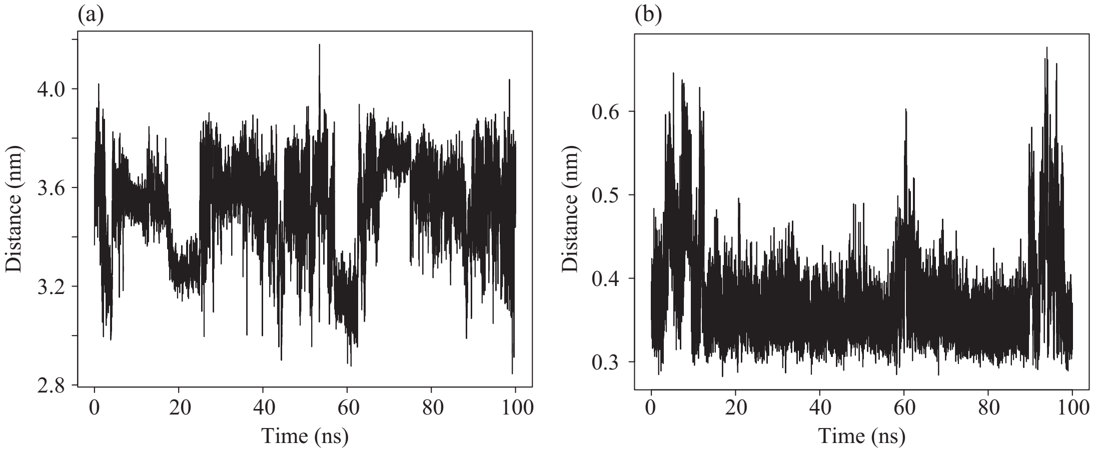

Fig. 2. Distance between (a) the first atom and the first centre of mass of aquaporin and (b) the opening diameter y_t over time t.

a test set, each with $p = 10\,000$ observations. First, we tested whether the data are stationary. The augmented Dickey–Fuller test confirmed stationarity for Y with a p-value < 0.01. The Hurst exponent of Y is 0.85, indicating moderate long-range dependence supported by a rather slow decay of the sample covariances; see the grey line in Fig. 3(a). Therefore, we set $q = 1$ for our estimator to match the low smoothness of the corresponding spectral density. Application of the R package `forecast`, which implements the approach of Guerrero (1993) to estimate the Box–Cox transform parameter, confirms that the log-transform is appropriate for these data. The bin number is set to $T = 2500$, i.e., $\upsilon \approx 0.85$. The smoothing parameter of the log-spectral density is selected with generalized cross-validation. The black line in Fig. 3(a) confirms that the covariance matrix estimated with our method almost completely decorrelates the channel diameter Y on the training dataset. Next, we estimated the regression coefficients b with the usual partial least-squares algorithm, ignoring the dependence in the data. Finally, we estimated b with the partial least-squares algorithm that takes into account dependence using our covariance estimator $\hat{\Sigma}$. Based on these regression coefficient estimators, the prediction on the test set was calculated. Figure 2(b) shows the Pearson correlation between the true channel diameter on the test set and the prediction on the same test set based on raw (in grey) and decorrelated data (in black). Obviously, the performance of the partial least-squares algorithm on the decorrelated data is significantly better for smaller numbers of components. In particular, with just one component, the correlation between the true opening diameter on the test set and its prediction that takes into account the dependence in the data is already 0.45, while it is close to zero for the partial least-squares method that ignores the dependence in the data. Krivobokova et al. (2012) showed that the estimator of b based on one PLS component is exactly the ensemble-weighted maximally correlated mode, which is defined as the collective mode of atoms that has the highest probability to achieve a specific alteration of response Y. Therefore, an accurate estimator of this quantity is crucial for the interpretation of the results and can only be achieved if the dependence in the data is taken into account.

Estimating Σ with a tapered covariance estimator has two practical problems. First, since we only have a single realization of a time series Y, i.e., $n = 1$, there is no data-driven method for selecting the tapering parameter. Second, the tapering estimator

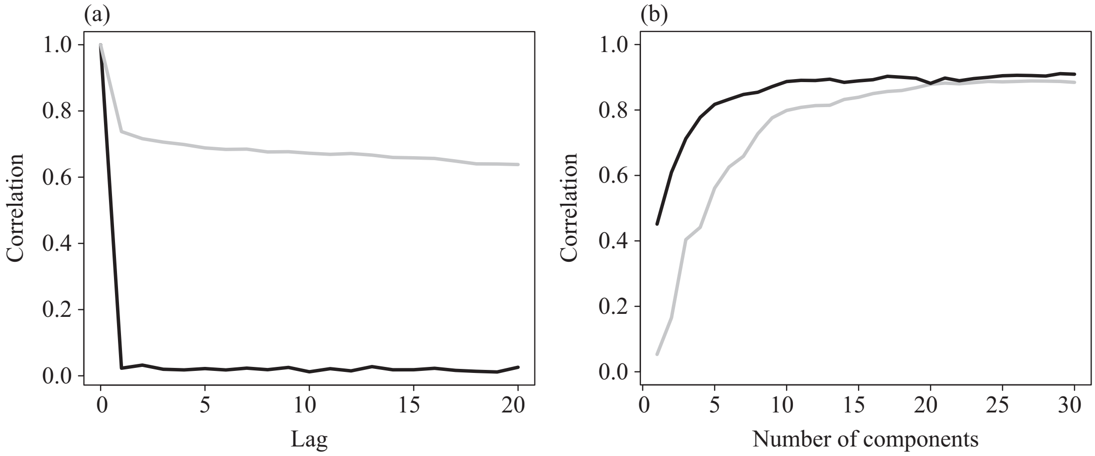

Fig. 3. (a) The correlation function of Y (grey) and of $\hat{\Sigma}^{-1/2} Y$ (black), where $\hat{\Sigma}$ is estimated with our method. (b) Correlation between the true values on the test dataset and the prediction based on partial least squares (grey) and corrected partial least squares (black).

turned out to be not positive definite for the data at hand. To solve the second problem, we truncated the corresponding spectral density estimator $\hat{f}_{\text{tap}}$ to a small positive value, i.e., $\hat{f}^{+}_{\text{tap}} = \max\{\hat{f}_{\text{tap}}, 1/\log p\}$ (McMurry & Politis, 2010; Cai et al., 2013). To select the tapering parameter with cross-validation, we experimented with different subseries lengths and found that the tapering estimator is very sensitive to this choice. For example, estimating the tapered covariance matrix based on subseries of length 8/15/30 yields a correlation of 0.42/0.53/0.34 between the true diameter and the first component, respectively.

In summary, our proposed estimator is fully data driven, fast even for large sample sizes, automatically positive definite and can handle certain long-memory processes. The protein data example and further simulations in the Supplementary Material suggest that our approach yields robust estimators even when data are not normally distributed. In contrast, the tapering estimator is not data driven in all data scenarios and must be manipulated to become positive definite. To the best of our knowledge, the tapering estimator has not been studied for non-Gaussian data. Our method is implemented in the R package `vstdct`.

Acknowledgement

The authors thank the reviewer, associate editor and Benedikt Pötscher for insightful comments that improved our manuscript. Funding by the German Research Association is also gratefully acknowledged.

Supplementary material

The Supplementary Material includes a summary of periodic smoothing splines, proofs of Lemma 1 and Theorem 1, proofs of the auxiliary results used in the proof of Theorem 1, simulation results for non-Gaussian data, and Q-Q plots of the binned and transformed data Y_k^*.

Appendix

A.1. *Proof of Theorem 1*

Throughout the Appendix, we denote by $c, c_1, C, C_1, \ldots$ etc. generic constants that are independent of n and p. To simplify the notation, the constants are sometimes skipped and we write $\lesssim$ for less than or equal to up to constants.

The structure of the proof is as follows. First, we derive the L_∞ rate of the periodic smoothing spline estimator $\widehat{H(f)}$. Then, using the Cauchy–Schwarz inequality and a mean value argument, the convergence rate of the spectral density estimator $\hat{f}$ is established. With the relationship $E\|\hat{\Sigma} - \Sigma\| \leqslant E\|\hat{f} - f\|_\infty^2$, the first claim of the theorem follows. Finally, we prove the second statement on the precision matrices. For the sake of clarity, the proofs of the auxiliary Lemmas 2, 3 and 4, and the proof of the convergence rate of $\hat{f}$ and of the precision matrices are listed separately in the Supplementary Material.

First we derive an upper bound on $E\|\widehat{H(f)} - H(f)\|_\infty^2$.

Proposition 1. *Let $\Sigma = \Sigma(f)$ with $f \in \mathcal{F}_\beta$ such that $\beta > 0$. If $h > 0$ such that $h \to 0$ and $hT \to \infty$, then, with $T = \lfloor p^\upsilon \rfloor$ for any $\upsilon \in (1 - \min\{1, \beta\}/3, 1)$, estimator $\widehat{H(f)}$ described in § 3 with $q = \max\{1, \gamma\}$ satisfies, for $p \to \infty$ and n such that $p^{\min\{1,\beta\}}/n \to c \in (0, \infty]$,*

$$E\|\widehat{H(f)} - H(f)\|_\infty^2 = \mathcal{O}\{\log(np)/(nph)\} + \mathcal{O}(h^{2\beta}).$$

Proof. Application of the triangle inequality yields a bias-variance decomposition

$$E\|\widehat{H(f)} - H(f)\|_\infty^2 \leqslant 2E\|\widehat{H(f)} - E\{\widehat{H(f)}\}\|_\infty^2 + 2\|E\{\widehat{H(f)}\} - H(f)\|_\infty^2.$$

Set $\tilde{T} = 2T - 2$ and $x_k = (k-1)/\tilde{T}$ for $k = 1, \ldots, \tilde{T}$. The periodic smoothing spline estimator $\widehat{H(f)}$ can be represented as a kernel estimator with respect to a kernel function $W(x, y)$ or the scaled version $K_h(x, t) = hW(x, t)$. An explicit representation of the kernel is derived by Schwarz & Krivobokova (2016). A summary of their results is given in the Supplementary Material. Lemma 4 listed in the Supplementary Material gives the following decomposition of $\widehat{H(f)}(x)$ for $x \in [0, 1]$ into a deterministic, a Gaussian and a non-Gaussian part:

$$\widehat{H(f)}(x) = \frac{1}{\tilde{T}} \sum_{k=1}^{\tilde{T}} W(x, x_k) Y_k^* = \frac{1}{\tilde{T}h} \sum_{k=1}^{\tilde{T}} K_h(x, x_k)[H\{f(x_k)\} + \epsilon_k + \zeta_k + \xi_k].$$

Here $|\epsilon_k| \lesssim (np)^{-1} + (np)^{-\beta} \log p$, $\zeta_k \sim \mathcal{N}(0, m^{-1})$ with $\text{cov}(\zeta_k, \zeta_l) = \mathcal{O}\{p^{-2} + p^{-2\beta}(\log p)^2\}$ for $k \neq l$. The random variable ξ_k satisfies $E|\xi_k|^\ell \lesssim (\log m)^{2\ell}\{m^{-\ell} + (T^{-1} + T^{-1}p^{1-\beta} \log p)^\ell\}$ for each integer $\ell > 1$ and has mean zero. Mirroring and renumerating $\zeta_k, \eta_k, \epsilon_k$ is similar to that for Y_k^*, $k = 1, \ldots, \tilde{T}$.

Now we derive an upper bound on the variance. Using the above representation, one can write

$$E\|\widehat{H(f)} - E\{\widehat{H(f)}\}\|_\infty^2 = E\left\| \frac{1}{\tilde{T}} \sum_{k=1}^{\tilde{T}} W(x, x_k)(\zeta_k + \xi_k) \right\|_\infty^2. \qquad \text{(A1)}$$

First, we bound the supremum by a maximum over a finite number of points. If $q > 1$ then $W(\cdot, x_k)$ is Lipschitz continuous with constant $L > 0$. In this case, it holds almost surely that

$$\sup_{x \in [0,1]} \left| \frac{1}{\tilde{T}} \sum_{k=1}^{\tilde{T}} W(x, x_k)(\zeta_k + \xi_k) \right|^2$$

$$\leqslant \left[\max_{1 \leqslant j \leqslant \tilde{T}} \left| \frac{1}{\tilde{T}} \sum_{k=1}^{\tilde{T}} W(x_j, x_k)(\zeta_k + \xi_k) \right| \right.$$

$$+ \sup_{x,x'\in[0,1],\,|x-x'|<1/\tilde{T}} \left|\frac{1}{\tilde{T}}\sum_{k=1}^{\tilde{T}}\{W(x,x_k)-W(x',x_k)\}(\zeta_k+\xi_k)\right|\Bigg]^2$$

$$\leqslant 2\max_{1\leqslant j\leqslant \tilde{T}}\left|\frac{1}{\tilde{T}}\sum_{k=1}^{\tilde{T}}W(x_j,x_k)(\zeta_k+\xi_k)\right|^2+\frac{2L^2}{\tilde{T}^4}\left(\sum_{k=1}^{\tilde{T}}|\zeta_k+\xi_k|\right)^2.$$

Using $E\{(\sum_{k=1}^{\tilde{T}}|\zeta_k+\xi_k|)^2\}\lesssim E\{(\sum_{k=1}^{\tilde{T}}|\zeta_k|)^2\}+E\{(\sum_{k=1}^{\tilde{T}}|\xi_k|)^2\}\lesssim \tilde{T}^2m^{-1}$, one gets

$$(A1)\lesssim E\left\{\max_{1\leqslant j\leqslant\tilde{T}}\left|\frac{1}{\tilde{T}}\sum_{k=1}^{\tilde{T}}W(x_j,x_k)(\zeta_k+\xi_k)\right|^2\right\}+o\{(np)^{-1}\}. \tag{A2}$$

If $q=1$ then $\sum_{k=1}^{\tilde{T}}W(\cdot,x_k)$ is a piecewise linear function with knots at $x_j=j/\tilde{T}$. The factor $(\zeta_k+\xi_k)$ can be considered as stochastic weights that do not affect the piecewise linear property. Thus, the supremum is attained at one of the knots $x_j=j/\tilde{T}$ for $j=1,\ldots,\tilde{T}$, and (A2) is also valid for $q=1$. Again with $(a+b)^2\leqslant 2a^2+2b^2$, we obtain

$$(A1)\lesssim E\left\{\max_{1\leqslant j\leqslant\tilde{T}}\left|\frac{1}{\tilde{T}}\sum_{k=1}^{\tilde{T}}W(x_j,x_k)\zeta_k\right|^2\right\}+E\left\{\max_{1\leqslant j\leqslant\tilde{T}}\left|\frac{1}{\tilde{T}}\sum_{k=1}^{\tilde{T}}W(x_j,x_k)\xi_k\right|^2\right\}+o\{(np)^{-1}\}.$$

We start with bounding $T_1=E\{\max_{1\leqslant j\leqslant\tilde{T}}|(\tilde{T}h)^{-1}\sum_{k=1}^{\tilde{T}}K_h(x_j,x_k)\zeta_k|^2\}$ with Lemma 1.6 of Tsybakov (2009). This requires a bound on $\|(\tilde{T}h)^{-1}\sum_{k=1}^{\tilde{T}}K_h(x_j,x_k)\zeta_k\|_{\psi_2}^2$, where $\|\cdot\|_{\psi_2}$ denotes the sub-Gaussian norm. In the case of a Gaussian random variable the norm equals the variance. See Vershynin (2018) for further details on the sub-Gaussian distribution.

Using the properties of the kernel function K_h stated in Lemma 2 of the Supplementary Material, we obtain

$$\begin{aligned}\left\|\frac{1}{\tilde{T}h}\sum_{k=1}^{\tilde{T}}K_h(x_j,x_k)\zeta_k\right\|_{\psi_2}^2&=\operatorname{var}\left\{\frac{1}{\tilde{T}h}\sum_{k=1}^{\tilde{T}}K_h(x_j,x_k)\zeta_k\right\}\\&=\frac{1}{\tilde{T}^2h^2}\sum_{k=1}^{\tilde{T}}K_h(x_j,x_k)^2\operatorname{var}(\zeta_k)\\&\quad+\frac{1}{\tilde{T}^2h^2}\sum_{k=1}^{\tilde{T}}K_h(x_j,x_k)\sum_{l=1}^{\tilde{T}}K_h(x_j,x_l)\operatorname{cov}(\zeta_k,\zeta_l)\\&\leqslant C(Thm)^{-1}+C'\{p^{-2}+p^{-2\beta}(\log p)^2\}.\end{aligned}$$

Lemma 1.6 of Tsybakov (2009) then yields

$$T_1\lesssim\log(2\tilde{T})[C(Thm)^{-1}+C'\{p^{-2}+(\log p)^2p^{-2\beta}\}]=\mathcal{O}\{(Thm)^{-1}\log(2\tilde{T})+h^{2\beta}\}.$$

To see this, note that $p^{\min\{1,\beta\}}/n\to c\in(0,\infty]$ implies that $p^{-1}=\mathcal{O}(n^{-1})$. Thus, if $\beta>1$ then

$$\log(2\tilde{T})\{p^{-2}+p^{-2\beta}(\log p)^2\}\lesssim\log(2\tilde{T})p^{-2}\lesssim\log(2\tilde{T})(Tm)^{-1}.$$

Now consider $0<\beta\leqslant 1$. Recall that $T=\lfloor p^{\upsilon}\rfloor$ for some fixed $\upsilon\in(1-\min\{1,\beta\}/3,1)$. One can find a constant $a=C_\upsilon$ depending on υ, but not on n,p such that the inequality $\log(x)\leqslant x^a/a$ implies that $\log(2\tilde{T})\log(p)^2p^{-2\beta}T^{2\beta}=\mathcal{O}(1)$. Thus, $\log(2\tilde{T})\log(p)^2p^{-2\beta}=\mathcal{O}(h^{2\beta})$ since $hT\to\infty$ by assumption.

Next, we derive a bound for the second term $T_2 = E\{\max_{1\leqslant j\leqslant \tilde{T}} |(\tilde{T}h)^{-1} \sum_{k=1}^{\tilde{T}} K_h(x_j, x_k)\xi_k|^2\}$. The exponential decay property of the kernel K_h, see Lemma 2 in the Supplementary Material, yields

$$T_2 \lesssim E\left\{ \max_{1\leqslant j\leqslant \tilde{T}} \left| \frac{1}{\tilde{T}h} \sum_{k=1}^{\tilde{T}} \gamma_h(x_j, x_k)|\xi_k| \right|^2 \right\},$$

where $\gamma_h(x,t) = \gamma^{|x-t|/h} + \gamma^{1/h}\{\gamma^{(x-t)/h} + \gamma^{(t-x)/h}\}(1-\gamma^{1/h})^{-1}$ and $\gamma \in (0,1)$ is a constant. For some threshold $R > 0$ specified later, define $\xi_k^- = |\xi_k|\mathbb{1}\{|\xi_k| \leqslant R\}$ and $\xi_k^+ = |\xi_k|\mathbb{1}\{|\xi_k| > R\}$. Then,

$$T_2 \lesssim E\left\{ \max_{1\leqslant j\leqslant \tilde{T}} \left| \frac{1}{\tilde{T}h} \sum_{k=1}^{\tilde{T}} \gamma_h(x_j, x_k)\xi_k^- \right|^2 \right\} + E\left\{ \max_{1\leqslant j\leqslant \tilde{T}} \left| \frac{1}{\tilde{T}h} \sum_{k=1}^{\tilde{T}} \gamma_h(x_j, x_k)\xi_k^+ \right|^2 \right\}. \tag{A3}$$

The first term in (A3) can be bounded, again with Lemma 1.6 of Tsybakov (2009). We use the fact that, for not necessarily independent random variables $X_1, \ldots, X_N$ with $|X_i| < R$, $i = 1, \ldots, N$, it holds that $\| \sum_{i=1}^N a_i X_i\|_{\psi_2}^2 \leqslant 4R^2 \sum_{i=1}^N a_i^2$, where $a_1, \ldots, a_N \in \mathbb{R}$ and $R > 0$ are constants. This is a consequence of Lemma 1 of Azuma (1967), which yields $E\{\exp(\lambda \sum_{i=1}^N a_i X_i)\} \leqslant \exp(\lambda^2 R^2 \sum_{i=1}^N a_i^2)$ for all $\lambda \in \mathbb{R}$. Thus, for all $t > 0$, $\mathrm{pr}(|\sum_{i=1}^N a_i X_i| > t) \leqslant 2\exp(-\lambda t + \lambda^2 R^2 \sum_{i=1}^N a_i^2)$ holds. Setting $\lambda = (t/2)(R^2 \sum_{i=1}^N a_i^2)^{-1}$, it follows that $\sum_{i=1}^N a_i X_i$ has a sub-Gaussian distribution and the sub-Gaussian norm is bounded by $2R(\sum_{i=1}^N a_i^2)^{1/2}$. Together, we get $\|(\tilde{T}h)^{-1}\sum_{k=1}^{\tilde{T}} \gamma_h(x_j, x_k)\xi_k^-\|_{\psi_2}^2 \leqslant 4R^2(\tilde{T}h)^{-2}\sum_{k=1}^{\tilde{T}} \gamma_h(x_j,x_k)^2 \lesssim R^2(\tilde{T}h)^{-1}$, where we have used the bound on γ_h from Lemma 2. Applying Lemma 1.6 of Tsybakov (2009) then yields

$$E\left\{ \max_{1\leqslant j\leqslant \tilde{T}} \left| \frac{1}{\tilde{T}h} \sum_{k=1}^{\tilde{T}} \gamma_h(x_j, x_k)\xi_k^- \right|^2 \right\} \lesssim \frac{R^2}{\tilde{T}h}\log(2\tilde{T}). \tag{A4}$$

To bound the second term in (A3), we use the moment bounds for ξ_k derived in Lemma 4. Then, for all integers $\ell > 1$,

$$\begin{aligned}
&E\left\{ \max_{1\leqslant j\leqslant \tilde{T}} \left| \frac{1}{\tilde{T}h} \sum_{k=1}^{\tilde{T}} \gamma_h(x_j, x_k)\xi_k^+ \right|^2 \right\} \\
&\quad = E\left\{ \max_{1\leqslant j\leqslant \tilde{T}} \left| \frac{1}{\tilde{T}h} \sum_{k=1}^{\tilde{T}} \gamma_h(x_j, x_k)(\xi_k^+)^{\ell}(\xi_k^+)^{1-\ell} \right|^2 \right\} \\
&\quad \leqslant R^{2-2\ell} E\left\{ \max_{1\leqslant j\leqslant \tilde{T}} \frac{1}{\tilde{T}^2 h^2} \sum_{k=1}^{\tilde{T}} \gamma_h(x_j, x_k)^2 \sum_{k=1}^{\tilde{T}} (\xi_k^+)^{2\ell} \right\} \\
&\quad \leqslant R^{2-2\ell} E\left(\frac{C_1}{\tilde{T}h} \sum_{k=1}^{\tilde{T}} \xi_k^{2\ell} \right) \\
&\quad \lesssim h^{-1}R^{2-2\ell}(\log m)^{4\ell}\{m^{-2\ell} + (T^{-1} + T^{-1}p^{1-\beta}\log p)^{2\ell}\}.
\end{aligned} \tag{A5}$$

Combining the error bounds (A4) and (A5) and choosing $R = m^{-1/2}$ gives

$$T_2 \lesssim \frac{\log T}{mTh} + \frac{(\log m)^{4\ell}}{hm^{1-\ell}} \times \begin{cases} \dfrac{(\log p)^{2\ell} p^{2\ell(1-\beta)}}{T^{2\ell}} + m^{-2\ell}, & 0 < \beta \leqslant 1, \\ T^{-2\ell} + m^{-2\ell}, & 1 < \beta. \end{cases}$$

By assumption, $T = \lfloor p^{\upsilon} \rfloor$ and $m = \lfloor np^{(1-\upsilon)} \rfloor$ for some fixed $\upsilon \in (1 - \min\{1, \beta\}/3, 1)$. If ℓ is an integer such that $\ell \geqslant 1/(1-\upsilon)$ then

$$m^{\ell-1}(\log m)^{4\ell} m^{-2\ell} \lesssim m^{-l} \lesssim n^{-\ell} p^{-(1-\upsilon)\ell} = \mathcal{O}\{(np)^{-1}\},$$

where we have used $\log x \leqslant x^a/a$ with $a = 1/(4\ell)$. Now, consider $0 < \beta \leqslant 1$. Then, $n \leqslant cp^{\beta}$ by assumption. Let $0 < \chi < 1$ be a small constant. Using the inequality $\log x \leqslant x^a/a$ twice with $a = \chi/(2\ell)$ gives

$$\begin{aligned} m^{\ell-1}(\log m)^{4\ell}(\log p)^{2\ell} p^{2\ell(1-\beta)} T^{-2\ell} &\lesssim n^{\ell-1+\chi} p^{(1-\upsilon)(\ell-1+\chi)-2\ell\upsilon+2\ell(1-\beta)+\chi} \\ &\lesssim p^{\ell(3-3\upsilon-\beta)+(\beta+\chi+1)(1-v)+\chi}. \end{aligned} \tag{A6}$$

For $\upsilon \in (1 - \beta/3, 1)$, the condition $\ell(3 - 3\upsilon - \beta) + (\beta + \chi + 1)(1 - v) + \chi < -2$ is equivalent to $\ell > \{-2 - (\beta + \chi + 1)(1 - v) - \chi\}/(3 - 3\upsilon - \beta)$. Thus, for any fixed υ, one can find an integer ℓ that is independent of n, p such that the right-hand side of (A6) is of $\mathcal{O}(p^{-2})$. For the same choice of ℓ, it holds that $m^{\ell-1}(\log m)^{4\ell} T^{-2\ell} = \mathcal{O}(p^{-2})$.

In total, choosing an integer $\ell \geqslant \max[1/(1-\upsilon), \{-2 - (\beta + \chi + 1)(1 - v) - \chi\}/(3 - 3\upsilon - \beta)]$ gives

$$\mathbb{E}\|H(\hat{f}) - \mathbb{E}\{H(\hat{f})\}\|_{\infty}^2 = \mathcal{O}\left\{\frac{\log(np)}{nph} + h^{2\beta}\right\}. \tag{A7}$$

Next we derive an upper bound on the bias. Using the representation in Lemma 4 once more gives, for each $x \in [0, 1]$,

$$E\{H(\hat{f})(x)\} - H\{f(x)\} = \frac{1}{\tilde{T}h} \sum_{k=1}^{\tilde{T}} K_h(x, x_k)[H\{f(x_k)\} + \epsilon_k] - H\{f(x)\}.$$

The bounds on ϵ_k imply that

$$\left| \frac{1}{\tilde{T}h} \sum_{k=1}^{\tilde{T}} K_h(x, x_k)\epsilon_k \right| \lesssim \frac{1}{\tilde{T}h} \sum_{k=1}^{\tilde{T}} \gamma_h(x_j, x_k)|\epsilon_k| \lesssim (np)^{-1} + (np)^{-\beta} \log p.$$

Consider the case in which $\beta \geqslant 1$. In particular, $q = \gamma$ and $f^{(q)}$ is α-Hölder continuous. Since f is a periodic function with $f(x) \in [\delta, M_0]$ and $H(y) \propto \phi(m/2) + \log(2y/m)$, it follows that $\{H(f)\}^{(q)}$ is also α-Hölder continuous. Extending $g = H(f)$ to the entire real line, we get

$$\frac{1}{\tilde{T}h} \sum_{k=1}^{\tilde{T}} K_h(x, x_k) g(x_k) = \int_{-\infty}^{\infty} h^{-1} \mathcal{K}_h(x, t) g(t)\, \mathrm{d}t + \mathcal{O}(\tilde{T}^{-\beta}),$$

where $\mathcal{K}_h$ is the extension of K_h to the entire real line (Schwarz & Krivobokova, 2016). Expanding $g(t)$ in a Taylor series around x and using the fact that $h^{-1}\mathcal{K}_h$ is a kernel of order $2q$ (see Lemma 2(iii)), it follows that, for any $x \in [0, 1]$,

$$\begin{aligned} \frac{1}{\tilde{T}h} \sum_{k=1}^{\tilde{T}} K_h(x, x_k) g(x_k) &= g(x) + \int_{-\infty}^{\infty} h^{-1} \mathcal{K}_h(x, t)(x - t)^q \frac{g^{(q)}(\xi_{x,t})}{q!}\, \mathrm{d}t + \mathcal{O}(\tilde{T}^{-\beta}) \\ &= g(x) + \int_{-\infty}^{\infty} h^{-1} \mathcal{K}_h(x, t)(x - t)^q \frac{g^{(q)}(\xi_{x,t}) - g^{(q)}(x)}{q!}\, \mathrm{d}t + \mathcal{O}(\tilde{T}^{-\beta}) \\ &= g(x) + \sum_{l=-\infty}^{\infty} \int_{x+(l-1)h}^{x+lh} \mathcal{K}_h(x, t)(x - t)^q \frac{g^{(q)}(\xi_{x,t}) - g^{(q)}(x)}{hq!}\, \mathrm{d}t + \mathcal{O}(\tilde{T}^{-\beta}), \end{aligned}$$

where $\xi_{x,t}$ is a point between x and t. Using the facts that kernel $\mathcal{K}_h$ decays exponentially and that $g^{(q)}$ is α-Hölder continuous on $[\delta, M_0]$ with some constant L, one obtains

$$\begin{aligned}\left|\frac{1}{\tilde{T}h}\sum_{k=1}^{\tilde{T}} K_h(x,x_k)g(x_k) - g(x)\right| &\leqslant \frac{CL}{q!}\sum_{l=-\infty}^{\infty}\int_{x+(l-1)h}^{x+lh} \gamma^{|x-t|/h}|x-t|^q \frac{|\xi_{x,t}-x|^\alpha}{hq!}\,\mathrm{d}t + \mathcal{O}(\tilde{T}^{-\beta})\\ &\leqslant \frac{CL}{q!}\sum_{l=-\infty}^{\infty}\int_{x+(l-1)h}^{x+lh} \gamma^{|x-t|/h}\frac{|x-t|^\beta}{h}\,\mathrm{d}t + \mathcal{O}(\tilde{T}^{-\beta})\\ &\leqslant h^\beta \frac{CL}{q!}\sum_{l=-\infty}^{\infty}\gamma^{|l-1|}|l|^\beta + \mathcal{O}(\tilde{T}^{-\beta})\\ &= \mathcal{O}(h^\beta) + \mathcal{O}(\tilde{T}^{-\beta}).\end{aligned}$$

If $0 < \beta \leqslant 1$ then $q = 1$ and g is β-Hölder continuous. Since $f(x) \in [\delta, M_0]$ and the logarithm is Lipschitz continuous on a compact interval, it follows that $g = H(f)$ is β-Hölder continuous. Expanding g to the entire line and using Lemma 2(iii) with $m = 0$ gives

$$\frac{1}{\tilde{T}h}\sum_{k=1}^{\tilde{T}} K_h(x,x_k)g(x_k) - g(x) = \int_{-\infty}^{\infty}\mathcal{K}_h(x,t)\frac{g(t)-g(x)}{h}\,\mathrm{d}t + \mathcal{O}(\tilde{T}^{-\beta}).$$

In a similar way as before, one obtains

$$\left|\frac{1}{\tilde{T}h}\sum_{k=1}^{\tilde{T}} K_h(x,x_k)g(x_k) - g(x)\right| = \mathcal{O}(h^\beta) + \mathcal{O}(\tilde{T}^{-\beta}).$$

Note that $\tilde{T}^{-\beta} = o(h^\beta)$ as $hT \to \infty$ and $h \to 0$ by assumption. Since the derived bounds are uniform for $x \in [0, 1]$, it holds that

$$\|E\{H(\hat{f})(x)\} - H\{f(x)\}\|_\infty = \mathcal{O}(h^\beta) + \mathcal{O}\{(np)^{-1} + (np)^{-\beta}\log p\}. \tag{A8}$$

Putting bounds (A7) and (A8) together gives

$$E[\|H(\hat{f}) - H(f)\|_\infty^2] = \mathcal{O}\{\log(np)/(nph)\} + \mathcal{O}(h^{2\beta}).$$

This completes the proof of Proposition 1. □

Now we consider an upper bound on $E\|\hat{f} - f\|_\infty^2$.

Proposition 2. *Let* $\Sigma = \Sigma(f)$ *with* $f \in \mathcal{F}_\beta$ *such that* $\beta > 0$. *If* $h > 0$ *such that* $h \to 0$ *and* $hT \to \infty$, *then, with* $T = \lfloor p^\upsilon \rfloor$ *for any* $\upsilon \in (1 - \min\{1,\beta\}/3, 1)$, *estimator* $\hat{f}$ *described in* § 3 *with* $q = \max\{1,\gamma\}$ *satisfies, for* $p \to \infty$ *and* n *such that* $p^{\min\{1,\beta\}}/n \to c \in (0,\infty]$,

$$E\|\hat{f} - f\|_\infty^2 = \mathcal{O}\{\log(np)/(nph)\} + \mathcal{O}(h^{2\beta}).$$

The proof of Proposition 2 is given in the Supplementary Material.

For an upper bound on $E\|\hat{\Sigma} - \Sigma\|^2$, using the facts that the spectral norm of a Toeplitz matrix is upper bounded by the supremum norm of its spectral density, and that Proposition 2 holds for every $f \in \mathcal{F}_\beta$, we get

$$\sup_{f\in\mathcal{F}_\beta} E_f\|\hat{\Sigma} - \Sigma(f)\|^2 \leqslant \sup_{f\in\mathcal{F}_\beta} E_f\|\hat{f} - f\|_\infty^2 = \mathcal{O}\{\log(np)/(nph)\} + \mathcal{O}(h^{2\beta}). \tag{A9}$$

Minimizing the right-hand side of (A9) for the bandwidth parameter h yields $h \asymp \{\log(np)/(np)\}^{1/(2\beta+1)}$. In particular, $h \to 0$. Furthermore, $hT \geqslant cp^{1-\min\{\beta,1\}/3-(1+\min\{1,\beta\})/(2\beta+1)}\{\log(np)\}^{1/(2\beta+1)} \to \infty$ since, by assumption, $n \lesssim p^{\min\{1,\beta\}}$ and $T = \lfloor p^{\upsilon} \rfloor$ with $\upsilon \in (1-\min\{1,\beta\}/3, 1)$. Thus, substituting $h \asymp \{\log(np)/(np)\}^{1/(2\beta+1)}$ into (A9) gives the second result.

References

Azuma, K. (1967). Weighted sums of certain dependent random variables. *Tohoku Math. J.* **19**, 357–67.

Bartlett, M. S. (1950). Periodogram analysis and continuous spectra. *Biometrika* **37**, 1–16.

Bentkus, R. (1985). Rate of uniform convergence of statistical estimators of spectral density in spaces of differentiable functions. *Lith. Math. J.* **25**, 209–19.

Bickel, P. J. & Levina, E. (2008). Regularized estimation of large covariance matrices. *Ann. Statist.* **36**, 199–227.

Cai, T. T. & Zhou, H. H. (2010). Nonparametric regression in natural exponential families. In *Borrowing Strength: Theory Powering Applications—A Festschrift for Lawrence D. Brown*, vol. 6, Ed. J. O. Berger, T. T. Cai and I. M. Johnstone, pp. 199–215. Hayward, CA: Institute of Mathematical Statistics.

Cai, T. T., Ren, Z. & Zhou, H. H. (2013). Optimal rates of convergence for estimating Toeplitz covariance matrices. *Prob. Theory Rel. Fields* **156**, 101–43.

Chagny, G. (2016). An introduction to nonparametric adaptive estimation. *Grad. J. Math.* **1**, 105–20.

Choudhuri, N., Ghosal, S. & Roy, A. (2004). Bayesian estimation of the spectral density of a time series. *J. Am. Statist. Assoc.* **99**, 1050–9.

Du, X., Aubry, A., De Maio, A. & Cui, G. (2020). Toeplitz structured covariance matrix estimation for radar applications. *IEEE Sig. Proces. Lett.* **27**, 595–9.

Edwards, M. C., Meyer, R. & Christensen, N. (2019). Bayesian nonparametric spectral density estimation using B-spline priors. *Statist. Comp.* **29**, 67–78.

Fang, Y., Wang, B. & Feng, Y. (2016). Tuning-parameter selection in regularized estimations of large covariance matrices. *J. Statist. Comp. Simul.* **86**, 494–509.

Franaszczuk, P. J., Blinowska, K. J. & Kowalczyk, M. (1985). The application of parametric multichannel spectral estimates in the study of electrical brain activity. *Biol. Cybern.* **51**, 239–47.

Grenander, U. & Szegö, G. (1958). *Toeplitz Forms and Their Applications*. Berkeley, CA: University of California Press.

Guerrero, V. M. (1993). Time-series analysis supported by power transformations. *J. Forecasting* **12**, 37–48.

Hastie, T. & Tibshirani, R. (1990). *Generalized Additive Models*. Boca Raton, FL: Chapman and Hall/CRC Press.

Kooperberg, C., Stone, C. J. & Truong, Y. K. (1995). Rate of convergence for logspline spectral density estimation. *J. Time Ser. Anal.* **16**, 389–401.

Krivobokova, T. (2013). Smoothing parameter selection in two frameworks for penalized splines. *J. R. Statist. Soc.* B **75**, 725–41.

Krivobokova, T., Briones, R., Hub, J. S., Munk, A. & de Groot, B. L. (2012). Partial least-squares functional mode analysis: application to the membrane proteins AQP1, Aqy1, and CLC-ec1. *Biophys. J.* **103**, 786–96.

Maturana-Russel, P. & Meyer, R. (2021). Bayesian spectral density estimation using p-splines with quantile-based knot placement. *Comp. Statist.* **36**, 2055–77.

McMurry, T. L. & Politis, D. N. (2010). Banded and tapered estimates for autocovariance matrices and the linear process bootstrap. *J. Time Ser. Anal.* **31**, 471–82.

Pawitan, Y. & O'Sullivan, F. (1994). Nonparametric spectral density estimation using penalized Whittle likelihood. *J. Am. Statist. Assoc.* **89**, 600–10.

Pourahmadi, M. (2013). *High-Dimensional Covariance Estimation*. Hoboken, NJ: John Wiley and Sons.

Quah, D. (2000). Internet cluster emergence. *Eur. Econ. Rev.* **44**, 1032–44.

R Development Core Team (2024). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0, http://www.R-project.org.

Roberts, W. J. & Ephraim, Y. (2000). Hidden Markov modeling of speech using Toeplitz covariance matrices. *Speech Commun.* **31**, 1–14.

Schwarz, K. & Krivobokova, T. (2016). A unified framework for spline estimators. *Biometrika* **103**, 121–31.

Serra, P. & Krivobokova, T. (2017). Adaptive empirical Bayesian smoothing splines. *Bayesian Anal.* **12**, 219–38.

Singer, M., Krivobokova, T., Munk, A. & De Groot, B. (2016). Partial least squares for dependent data. *Biometrika* **103**, 351–62.

Taqqu, M. S. & Teverovsky, V. (1997). Robustness of Whittle-type estimators for time series with long-range dependence. *Commun. Statist.* **13**, 723–57.

Thomson, D. J. (1982). Spectrum estimation and harmonic analysis. *Proc. IEEE* **70**, 1055–96.

Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation*. New York: Springer.

Vershynin, R. (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science*, vol. 47. Cambridge: Cambridge University Press.

Wahba, G. (1980). Automatic smoothing of the log periodogram. *J. Am. Statist. Assoc.* **75**, 122–32.

Wahba, G. (1985). A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. *Ann. Statist.* **13**, 1378–402.

Walker, A. (1964). Asymptotic properties of least-squares estimates of parameters of the spectrum of a stationary non-deterministic time-series. *J. Aust. Math. Soc.* **4**, 363–84.

Welch, P. (1967). The use of fast Fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. *IEEE Trans. Audio Electroacoust.* **15**, 70–3.

Whittle, P. (1957). Curve and periodogram smoothing. *J. R. Statist. Soc.* B **19**, 38–47.

Wu, W. B. & Pourahmadi, M. (2009). Banding sample autocovariance matrices of stationary processes. *Statist. Sinica* **19**, 1755–68.

Wu, W. B. & Xiao, H. (2012). Covariance matrix estimation in time series. In *Handbook of Statistics*, vol. 30, Ed. T. S. Rao, S. S. Rao and C. R. Rao, pp. 187–209. Amsterdam: Elsevier.

Zygmund, A. (2002). *Trigonometric Series*, vol. 1. Cambridge: Cambridge University Press.

[*Received on* 24 *March* 2023. *Editorial decision on* 3 *January* 2024]