universität wien

DISSERTATION / DOCTORAL THESIS

Titel der Dissertation / Title of the Doctoral Thesis

On Periodograms and Efficient Nonparametric Toeplitz Covariance Matrix Estimators

verfasst von / submitted by

Karolina Hella Klockmann, B.Sc. M.Sc.

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Doctor of Philosophy (PhD)

Wien, 2023 / Vienna, 2023

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	A 794 370 136
Dissertationsgebiet lt. Studienblatt / field of study as it appears on the student record sheet:	Statistik und Operations Research
Betreut von / Supervisor:	Univ.-Prof. Dr. Tatyana Krivobokova

Acknowledgements

I would like to express my profound gratitude to everyone who has supported me throughout my doctoral journey and contributed to the completion of this dissertation.

First and foremost, I want to thank my supervisor, Tatyana Krivobokova, for her constant support, thoughtful guidance, and scientific expertise. I have learned a lot about how to tackle complex theoretical problems, how to present my results confidently at a conference, or how to write a scientific article for a journal. Her standard email response to any problems was always "*come over to my office*", for which I am deeply grateful. Even when I worked for weeks on some technical detail, Tatyana was immediately into the topic and ready to discuss next steps or find ways out of the dead end. After four years, she has finally made it so that my approach to solving all of the worlds problems is now also a Taylor expansion.

My doctoral journey has taken me to various nice places, from the lovely Göttingen to the beautiful city of Vienna and even led me to a research stay in Paris. All these places were made special by the nice colleagues and new friends I found there.

I remember with pleasure many cheerful coffee breaks, funny discussions, and joint activities as the treasure hunt in Göttingen. I would especially like to thank Peter Kramlinger for all the plants we killed and for the great time we spent in our joint offices in Göttingen and Vienna. No matter how hard Peter tried, he always arrived five minutes later than me, which already provided the first laugh of the day. Thanks to Georg Köstenberger, who always had an open ear to both mathematical and non-mathematical matters. It was reassuring to know that Julia Naas, Rupsa Basu, and I, despite being at three different universities after leaving Göttingen, struggle with exactly the same things as a PhD student. Thank you for all the fantastic Asian and Indian dishes that you have cooked and that I was unaware of their existence. Finally, I very much enjoyed the lunches with the "*K-group*" at the Danube Canal. Thank you, Daniel, both Fabios, both Georgs, Gianluca, Ilaria, Nico, Peter, and Tom, for the great time we shared together.

I am very grateful to Sabine Sobotka-Tompits and Svetlana Mihajlovic. Their enthusiasm to immediately find a solution to any administrative or IT problem is greatly appreciated and has been very helpful to me.

Very important was the support from my family, Dorothea, Ulrich and Oliver Klockmann - the saying "*they have always supported me*"- leads always to laughter and is, of course, very, very true. Many thanks to my father, for all the relocations you've made with me since my first semester in mathematics. Finally, I would like to thank Thomas, who endured all the moods linked to the progress of this work, and who always provided an exceptionally cheerful atmosphere. After one (of the very few) silent dinners, where I was still pondering problems in Lemma 2.1, he said, "*Ok, let's fight the cotangent together.*", and after some terms were brought under control with a Taylor expansion, we still had a fun evening and also comics of the cotangent.

Abstract

The estimation of covariance and precision matrices is a fundamental problem in statistical data analysis with numerous applications in the natural and social sciences. For stationary stochastic processes, the covariance matrix takes the simple form of a Toeplitz matrix. This cumulative dissertation presents two novel data-driven nonparametric methods for estimating Toeplitz covariance matrices in different scenarios and discusses their theoretical properties.

The main idea is to establish a connection between the estimation of the Toeplitz covariance matrix and the estimation of the (log-)spectral density function. Specifically, it is demonstrated that the Discrete Cosine Transform matrix and the Demmler-Reinsch basis matrix diagonalize Toeplitz matrices asymptotically. Utilizing this result, two modified versions of the periodogram and the Whittle likelihood are derived, enabling the estimation of the spectral density function on a grid with a twice as fine frequency resolution.

In the first scenario, the data consist of n independent and identically distributed realizations of a stationary Gaussian process with zero mean and a bounded, α-Hölder continuous spectral density function. In particular, this setting includes certain processes with long-range dependence. The Toeplitz covariance matrix and the precision matrix are obtained through an estimator of the log-spectral density function in an approximate Gaussian regression. The main theoretical result is the proof that all proposed estimators are minimax optimal: the spectral density estimator attains the minimax optimal convergence rate under the L_∞ norm, the Toeplitz covariance and the precision matrix estimator attain the minimax optimal rate under the spectral norm. As a separate result, we derive an approximate expression for the covariance of the log-average periodogram.

In the second scenario, a single stationary Gaussian time series with short-term dependence and an unknown smooth mean function is considered. The covariance matrix and the mean function are simultaneously estimated using an adaptive, empirical

Bayesian approach.

Both methods are entirely data-driven and computationally efficient. Their performance is evaluated through simulations and real data examples. The two method are implemented in the R packages *vstdct* and *eBsc*.

Kurzfassung

Die Schätzung von Kovarianz- und Präzisionsmatrizen ist ein grundlegendes Problem der statistischen Datenanalyse mit unzähligen Anwendungen in den Natur- und Sozialwissenschaften. Für die Klasse der stationären stochastischen Prozesse weist die Kovarianzmatrix eine besonders einfache Struktur auf: sie ist eine Toeplitz-Matrix. In dieser kumulativen Dissertation werden zwei neue datengetriebene nicht-parametrische Methoden zur Schätzung von Toeplitz-Kovarianzmatrizen in zwei unterschiedlichen Szenarien vorgeschlagen und deren theoretische Eigenschaften diskutiert.

Der Hauptansatz besteht darin, das Problem der Schätzung der Toeplitz-Kovarianzmatrix mit dem Problem der Schätzung der (Log-)Spektraldichtefunktionen in Verbindung zu setzen. Insbesondere wird gezeigt, dass die diskrete Kosinustransformationsmatrix und die Demmler-Reinsch-Basismatrix Toeplitz-Matrizen asymptotisch diagonalisieren. Unter Verwendung dieses Ergebnisses werden zwei modifizierte Versionen des Periodogramms und der Whittle-Likelihood hergeleitet, die eine Schätzung der Spektraldichtefunktion mit doppelt so feiner Frequenzauflösung ermöglichen.

Im ersten Szenario liegen die Daten als n unabhängige und identisch verteilte Realisierungen eines stationären Gauß-Prozesses mit einem Mittelwert von Null und einer α-Hölder stetigen, beschränkten Spektraldichtefunktion vor. Insbesondere beinhaltet dieses Szenario bestimmte Prozesse mit Langzeitpersistenz. Die Toeplitz-Kovarianzmatrix und die Präzisionsmatrix werden aus einem Schätzer der Log-Spektraldichtefunktion in einer approximativen Gaußschen Regression gewonnen. Das wichtigste theoretische Ergebnis ist der Beweis, dass alle vorgeschlagenen Schätzer minimax-optimal sind: der Spektraldichte-Schätzer erreicht die minimax-optimale Konvergenzrate unter der L_∞-Norm, der Toeplitz-Kovarianz- und der Präzisionsmatrix-Schätzer erreichen die minimax-optimale Rate unter der Spektralnorm. Als separates Ergebnis leiten wir einen Näherungsausdruck für die Kovarianz des Log-Durchschnitt-Periodogramms her.

Im zweiten Szenario wird eine einzelne stationäre Gauß-Zeitreihe mit kurzfristiger Abhängigkeit und einer unbekannten stetigen Mittelfunktion betrachtet. Die Kovarianzmatrix und die Mittelfunktion werden gleichzeitig mithilfe eines adaptiven, empirischen Bayes-Ansatzes geschätzt.

Beide Methoden sind vollständig datengetrieben und rechentechnisch sehr effizient. Ihre Leistungsfähigkeit wird anhand von Simulationen und Beispielen echter Daten evaluiert. Die beiden Methoden sind in den R-Paketen *vstdct* und *eBsc* implementiert.

Contents

1 Introduction

The covariance matrix is a fundamental statistical tool used to quantify the relationship between random variables. It plays a crucial role in various scientific fields, including finance, physics, medicine and engineering. In many applications, data are observed over time, and the measurements at each time point are influenced by the preceding measurements. Examples include stock prices, rainfall measurements, electrical brain activity data, or power consumption data. The covariance matrix is the primary tool for describing the dependence between measurements at different time points. In particular, the covariance matrix serves as input for various statistical algorithms. Accurate estimation of the covariance structure is essential for interpreting the data correctly, decorrelating the data, fitting appropriate models, and conducting reliable inference.

This dissertation focuses on the estimation of covariance matrices for data with temporal dependence in two different settings. We propose two novel estimation methods and thoroughly analyze their theoretical properties. The research results presented in this cumulative dissertation consist of three publications, each contributing to the theory of covariance matrix estimation and the related problem of spectral density estimation. Before delving into the specific contributions of each publication, we first introduce the relevant terminology and concepts necessary for a comprehensive understanding of the subsequent chapters.

1.1 Data settings

Stock prices, rainfall data or electrical brain activity data differ substantially in their nature. They vary in terms of measurement units, observation frequency, and the presence of periodic components, among other factors. When analyzing electrical brain activity data, usually only one observation is available, which is measured for a specific person. In contrast, meterology institutes typically posses large databases of rainfall data, such as the records for the month of October in Vienna, for example. To put this in a common framework, we represent the observed data as a set of n real-valued vectors denoted as $\mathbf{Y}_1, ..., \mathbf{Y}_n$, each with a length of p. Here, $\mathbf{Y}_i = (Y_{i,1}, ..., Y_{i,p})^t$ for $i = 1, ..., n$ and the superscript denotes the transpose. If there is only one observation, i.e., $n = 1$, the index is omitted and we refer to the data as $\mathbf{Y} = (Y_1, ..., Y_p)^t$. From a probabilistic point of view, we assume that the observations $\mathbf{Y}_1, ..., \mathbf{Y}_n$ are realizations of a stochastic process $(Y_t)_{t \in \mathbb{Z}}$. For the asymptotic analysis of statistical estimators based on $\mathbf{Y}_1, ..., \mathbf{Y}_n$, as e.g., a covariance matrix estimator, the following cases are distinguished

1. **Time series case**: $n = 1$, only p grows and goes to ∞.

2. **Large n small p case**: $n \gg p$, n and p go to ∞ such that $p/n \to 0$.

3. n, p **have the same order**: n and p are of similar size, and $p/n \to c$ for some constant $0 < c < \infty$.

4. **Large p small n case**: $p \gg n$, p and n go to ∞ such that $p/n \to \infty$.

Typically, the different structural components occurring in the data as the mean function and periodic components are modeled separately. The stochastic remainder usually shows stationary behavior and is used to estimate the covariance structure of the data. Therefore, stationary processes are the main class of processes we consider in this thesis.

1.2 Stationary processes

A real-valued stochastic process $(Y_t)_{t \in \mathbb{Z}}$ with finite second moments, i.e., $\mathbb{E}[Y_t^2] < \infty$ for all $t \in \mathbb{Z}$, is called *(weakly) stationary* if there exists a function $\sigma : \mathbb{Z} \to \mathbb{R}, h \mapsto \sigma_h$ and $\mu \in \mathbb{R}$ such that, for all $t, h \in \mathbb{Z}$,

$$\mathbb{E}(Y_t) = \mu \quad \text{and} \quad \operatorname{Cov}(Y_{t+h}, Y_t) = \sigma_h. \tag{1.1}$$

Stationarity means that the covariance structure of the process $(Y_t)_{t \in \mathbb{Z}}$ depends only on the time interval between two observations Y_t and Y_{t+h}, but not on the exact value t. The function σ is called *autocovariance function*. The *autocovariance matrix* of the process is given by $\mathbf{\Sigma}_\infty = (\sigma_{|i-j|})_{i,j \in \mathbb{Z}}$. In particular, $\mathbf{\Sigma}_\infty$ is a positive semidefinite Toeplitz matrix, i.e., a matrix for which each diagonal is constant. For the vector $\mathbf{Y} = (Y_1, ..., Y_p)^t$, the autocovariance matrix is the $(p \times p)$-dimensional submatrix, denoted by $\mathbf{\Sigma}_p$. If the dimension of the matrix is clear from the context, the index is omitted. The prefix "auto" is typically used in the time series setting, i.e., there is only $n = 1$ observation $\mathbf{Y}$. If $n > 1$, the quantities σ, $\mathbf{\Sigma}_p$ and $\mathbf{\Sigma}_\infty$ are called *covariance function* and *covariance matrix*.

Processes that satisfy $|\sigma_h| \leqslant C(1 + |h|)^{-\beta}$ with $\beta > 1$ and for some constant $C > 0$ are called *short-range* processes. If σ_h decays as $1/h$ or slower for $h \to \infty$, the process is called *long-range* process.

The simplest example of a stationary process is a white noise process. That is a process with finite second moments fulfilling (1.1) with $\mu = 0$, $\sigma_0 > 0$ and $\sigma_h = 0$ for all $h \neq 0$. Another example are stationary Gaussian processes. That are processes $(Y_t)_{t\in\mathbb{Z}}$ which satisfy (1.1) and that for every finite set of indices $t_1, ..., t_k \in \mathbb{Z}$ the vector $(Y_{t_1}, \ldots, Y_{t_k})$ is a multivariate Gaussian random variable. Stationary Gaussian processes have the special property that they are completely characterized by their mean μ and the Toeplitz covariance matrix $\mathbf{\Sigma}_\infty$.

A different approach to characterize the properties of a stationary process is via its frequency domain. The frequency-domain representation of a problem often simplifies the mathematical analysis or states a problem differently. In particular, both proposed covariance matrix estimators in this thesis exploit the fact that the problem of estimating $\mathbf{\Sigma}$ corresponds to the estimation of a 2π periodic function in the frequency domain.

1.3 Frequency domain analysis

For a real-valued stationary processes $(Y_t)_{t\in\mathbb{Z}}$ with an absolutely summable sequence of covariance $\sum_{h=-\infty}^{\infty} |\sigma_h| < \infty$, i.e., a short range process, the sequence of covariances $(\sigma_h)_{h\in\mathbb{Z}}$ can be interpreted as the sequence of Fourier coefficients of a 2π-periodic, non-negative function $f : [-\pi, \pi] \to \mathbb{R}_{\geq 0}$. This function is called the *spectral density* of the process $(Y_t)_{t\in\mathbb{Z}}$ and is defined as

$$f(\lambda) = \sum_{h=-\infty}^{\infty} \sigma_h \exp(-\mathrm{i}h\lambda) = \sigma_0 + 2\sum_{h=1}^{\infty} \sigma_h \cos(h\lambda), \quad \lambda \in [-\pi, \pi], \tag{1.2}$$

where i denotes the imaginary unit. The summability of $|\sigma|$ implies that the series converges absolutely. In particular, the covariance function can be uniquely recovered from f by the inverse Fourier transform, i.e.,

$$\sigma_h = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)\exp(\mathrm{i}hx)\,dx = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)\cos(hx)\,dx. \tag{1.3}$$

For a real-valued stationary process $(Y_t)_{t\in\mathbb{Z}}$ with $\sum_{h=-\infty}^{\infty} |\sigma_h| = \infty$, i.e., a long-range process, the spectral density function is directly defined as a 2π-periodic, non-negative function $f : [-\pi, \pi] \to \mathbb{R}_{\geq 0}$ that satisfies condition (1.3). However, existence of such a function is not guaranteed.

The proposed two covariance matrix estimators in this thesis exploit the connection between the spectral density function and the covariance function. Therefore, in the following, we only consider stationary processes for which the spectral density function exists.

A particular connection between a function f and its Fourier coefficients σ_h is the following: the smoother the function the faster is the decay of its Fourier coefficients and vice-versa. This can be stated more precisely as follows. For $\beta > 0$, denote by $\lfloor\beta\rfloor$ the largest integer strictly less than β and define $\alpha = \beta - \lfloor\beta\rfloor$. Furthermore, denote by $f^{(n)}$ the n-th derivative for $n \in \mathbb{N}$. If f is an integrable, 2π periodic function with Fourier coefficients σ_h such that $|\sigma_h| \leqslant C(1 + |h|)^{-\beta-1}$ for all $h \in \mathbb{Z}$ and some constant $C > 0$, then f has derivatives of all orders up to $\gamma = \lfloor\beta\rfloor \in \mathbb{N}$ (inclusively), and $f^{(\gamma)}$ is $\tilde{\alpha}$-Hölder continuous for $0 < \tilde{\alpha} < \alpha$ (Grafakos, 2008, Prop. 3.3.12). On the other hand, if f is in the periodic Hölder class $\mathcal{H}_{\beta,\text{per}}$, i.e., f is 2π-periodic, has derivatives of all orders up to γ and $f^{(\gamma)}$ is α-Hölder continuous, then the Fourier coefficients decay like $|\sigma_h| \leqslant C_1(1 + |h|)^{-\beta}$, where $C_1 > 0$ is some constant (Katznelson, 2004).

Next, we discuss the difference between parametric and nonparametric methods to estimate the covariance matrix.

1.4 Parametric versus nonparametric covariance models

Parametric methods assume that the covariance matrix has a certain structural form described by a small number of parameters. For example, the covariance function of an AR(1) process with parameter ϕ decays exponentially fast, i.e, $\sigma_h = \sigma_0\phi^{|h|}$. For a MA($q$) process, the covariance σ_h is zero for $h > q$. Parametric models are recommended when there is prior knowledge or theoretical understanding of the underlying covariance structure.

Nonparametric assume that the covariance matrix belongs to a larger class characterized by some parameter such as the smoothness of their spectral density function. Imposing fewer assumptions, nonparametric models can capture more complex patterns and relationships in the data. On the other hand, they require more data than parametric methods as they have a slower rate of convergence. Nonparametric models are recommended when there is limited knowledge about the covariance structure or when the data does not conform well to specific assumptions of parametric models.

In this thesis, we use nonparametric methods to enable estimation of the covariance matrix for a broad class of processes. Specifically, we consider stationary processes with a covariance structure which belongs to one of the classes outlined below.

1.5 Classes of Toeplitz covariance matrices

The fact that Toeplitz covariance matrices can be characterized either by the smoothness of the spectral density f, or by the decay of the covariance function σ motivates the definition of the following two classes. Since the observed data is p-dimensional, the two classes are defined as collections of $(p \times p)$-dimensional submatrices $\mathbf{\Sigma}_p$ of $\mathbf{\Sigma}_\infty$ as follows

1. **$\mathbf{\Sigma}_p$ with covariance function σ of polynomial decay** Let $\beta > 0$ and $0 < M < \infty$ be some constants. The class of Toeplitz covariance matrices whose covariance function σ decays polynomially of order β is defined as

$$\mathcal{G}_\beta(M) = \left\{ \mathbf{\Sigma}_p \in \mathbb{R}^{p \times p} : |\sigma_h| \leqslant M(h+1)^{-(\beta+1)}, \mathbf{\Sigma}_\infty \succ 0 \right\},$$

where $\mathbf{\Sigma}_\infty \succ 0$ means positive semidefinite.

2. **$\mathbf{\Sigma}_p$ with Hölder continuous spectral density function**
Let $\beta = \gamma + \alpha > 0$ such that $\gamma \in \mathbb{N}$, $0 < \alpha \leq 1$, and $0 < M_0, M_1 < \infty$ are some constants. The class of Toeplitz covariance matrices which have a spectral density f that belongs to the periodic Hölder space is defined by

$$\begin{aligned} &\mathcal{P}_\beta(M_0, M_1) \\ &= \left\{ \mathbf{\Sigma}_p \in \mathbb{R}^{p \times p} : \|f\|_\infty \leqslant M_0, \|f^{(\gamma)}(\cdot + h) - f^{(\gamma)}(\cdot)\|_\infty \leqslant M_1 |h|^\alpha, \mathbf{\Sigma}_\infty \succ 0 \right\}, \end{aligned}$$

where $\|f\|_\infty = \sup_{x \in [-\pi,\pi]} |f(x)|$ is the L_∞ norm.

In particular, if $0 < \beta < 1$, then $\mathcal{P}_\beta(M_0, M_1)$ contains covariance matrices of long-range processes with a β-Hölder continuous and bounded spectral density function. Let

$\mathcal{H}^{+}_{\beta,\text{per}}(M_0, M_1)$ denote the space of 2π-periodic, non-negative, even Hölder continuous functions with Hölder constant M_1 and which are bounded by M_0, i.e.,

$$\mathcal{H}^{+}_{\beta,\text{per}}(M_0, M_1) = \Big\{ f : [0, 2\pi] \to \mathbb{R}_{\geq 0} : \text{ periodic, even, } \|f\|_\infty \leq M_0, \\ \|f^{(\gamma)}(\cdot + h) - f^{(\gamma)}(h)\|_\infty \leq M_1 \Big\} .$$

Since $f \in \mathcal{H}^{+}_{\beta,\text{per}}(M_0, M_1)$ is L_2-integrable, there is a one-to-one correspondence between f and its Fourier coefficients (Katznelson, 2004). Hence, for every $\mathbf{\Sigma} \in \mathcal{P}_\beta(M_0, M_1)$ the spectral density is obtained by (1.2) and belongs to $\mathcal{H}^{+}_{\beta,\text{per}}(M_0, M_1)$. Vice versa, for every $f \in \mathcal{H}^{+}_{\beta,\text{per}}(M_0, M_1)$ the corresponding Toeplitz matrix can be obtained with the inverse Fourier transform (1.3) and belongs to $\mathcal{P}_\beta(M_0, M_1)$. To simplify the notation, we use only the space $\mathcal{P}_\beta(M_0, M_1)$ and indicate which object is considered by writing $f \in \mathcal{P}_\beta(M_0, M_1)$ or $\mathbf{\Sigma} \in \mathcal{P}_\beta(M_0, M_1)$. The class $\mathcal{G}_\beta(M)$ contains Toeplitz covariance matrices of short range processes for all $\beta > 0$.

For any $M > 0$ and non-integer $\beta > 0$, it can be shown that there exists some constants M_0 and M_1 depending on M such that $\mathcal{G}_\beta(M) \subset \mathcal{P}_\beta(M_0, M_1)$. However, in general this is not true for integer β. See, e.g., Zygmund (2002). Conversely, for any $\mathbf{\Sigma}_p \in \mathcal{P}_\beta(M_0, M_1)$, we have $|\sigma_h| \leq M(h+1)^{-\beta}$ where M is some constant depending only on M_0 and M_1. Therefore, $\mathcal{P}_\beta(M_0, M_1) \subset \mathcal{G}_{\beta-1}(M)$ for some M depending on M_0 and M_1.

The two classes play a crucial role in this thesis. In the upcoming three chapters, we assume that the true covariance matrix $\mathbf{\Sigma}$ belongs either to the class $\mathcal{G}_\beta(M)$ or $\mathcal{P}_\beta(M_0, M_1)$. Furthermore, the minimax optimal convergence rates have been established with respect to these two spaces, as discussed in Section 1.7. First, we introduce the standard nonparametric estimators for σ, f and $\mathbf{\Sigma}$.

1.6 Standard nonparametric estimators

Let $\mathbf{Y} = (Y_1, ..., Y_p) \in \mathbb{R}^p$ be an observation of a stationary stochastic process with mean zero. The standard nonparametric estimators for the covariance function σ, the spectral density function f and the covariance matrix $\mathbf{\Sigma}$ are given by

- **sample covariance function**: $\hat{\sigma}_h = \frac{1}{p} \sum_{j=1}^{p-|h|} Y_j Y_{j+|h|}$ for $h = 0, ..., p$,

- **periodogram**: $\hat{f}\,(2\pi k/p) = \frac{1}{p}\left|\sum_{j=1}^{p} Y_j \exp\{2\pi \mathrm{i} jk/p\}\right|^2$ for $k \in \mathbb{N} \cap [-(p-1)/2, p/2]$,
- **sample covariance matrix**: $\hat{\mathbf{\Sigma}} = (\hat{\sigma}_{|i-j|})_{i,j=1}^{p}$.

In particular, the periodogram corresponds to the squared absolute value of the discrete Fourier transform of $\mathbf{Y}$. The frequencies $\omega_k = 2\pi k/p$, where k is any integer in $[-(p-1)/2, p/2]$, are also called the *Fourier frequencies*. The sample covariance function and the periodogram are connected as follows: at any non-zero Fourier frequency ω_k, the periodogram corresponds to the discrete Fourier transform of the sample covariance function (Brockwell and Davis, 2002, Prop. 4.2.1).
When n independent and identically distributed (i.i.d.) observations $\mathbf{Y}_1, ..., \mathbf{Y}_n$ are available, the extension of the estimators is straightforward, i.e., denoting by $Y_{i,j}$ the j-th entry of $\mathbf{Y}_i$, then $\hat{\sigma}_h = \frac{1}{np}\sum_{i=1}^{n}\sum_{j=1}^{p-|h|} Y_{i,j}Y_{i,j+|h|}$, $\hat{f}\,(2\pi k/p) = \frac{1}{np}\sum_{i=1}^{n}\left|\sum_{j=1}^{p} Y_{i,j}\exp\{2\pi \mathrm{i} jk/p\}\right|^2$ and $\hat{\mathbf{\Sigma}} = (\hat{\sigma}_{|i-j|})_{i,j=1}^{p}$.

However, $\hat{\sigma}, \hat{f}, \hat{\mathbf{\Sigma}}$ are inconsistent estimators for $n, p \to \infty$ except in the setting where the sample size n is large compared to the dimension p, and $p/n \to 0$ for $n, p \to \infty$ (Wu and Pourahmadi, 2009; Pourahmadi, 2013; Brockwell and Davis, 2009).

To obtain a consistent estimator for $\mathbf{\Sigma}$, the sample covariance matrix $\hat{\mathbf{\Sigma}}$ can be manipulated with several techniques: tapering estimators multiply the sample covariance function with linear weights, decreasing from 1 to 0, banding estimators cut the sample covariance function off after some fixed lag, and thresholding estimators set all entries of $\hat{\mathbf{\Sigma}}$ to zero whose absolute value is smaller than some threshold. See Xiao and Wu (2012); McMurry and Politis (2010) for the time series setting where $n = 1$, and Cai et al. (2013) for the scenario where $n > 1$.

To improve the periodogram, several kernel functions such as Bartletts's window, Welch's window, the triangular kernel, etc., have been used to smooth the (log-)periodogram and reduce the variance (Bartlett, 1950; Welch, 1967; Thomson, 1982; Wahba, 1980). Other approaches include nonparametric Bayesian and penalized likelihood methods (Pawitan and O'Sullivan, 1994; Choudhuri et al., 2004; Edwards et al., 2019).

1.7 Minimax optimal rates

Let $\mathbf{Y}_1, ..., \mathbf{Y}_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}_p(0, \mathbf{\Sigma})$ with $\mathbf{\Sigma} \in \mathcal{R}_\beta \in \{\mathcal{G}_\beta(M), \mathcal{P}_\beta(M_0, M_1)\}$. The minimax optimal convergence rate for estimating the Toeplitz covariance matrix $\mathbf{\Sigma}$ from $\mathbf{Y}_1, ..., \mathbf{Y}_n$ has been established by Cai et al. (2013). The result is derived for the spectral norm, which is the square root of the largest eigenvalue for a symmetric matrix. Here, and in the following, the spectral norm is denoted by $\| \cdot \|$.

Theorem (Theorem 5 of Cai et al. (2013)) *The minimax risk for estimating the Toeplitz covariance matrix $\mathbf{\Sigma}$ over $\mathcal{R}_\beta$ under the spectral norm satisfies*

$$\inf_{\hat{\Sigma}} \sup_{\Sigma \in \mathcal{R}_\beta} \mathbb{E}\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|^2 \geq c \left\{ \frac{\log(np)}{np} \right\}^{\frac{2\beta}{2\beta+1}}$$

for some constant $c > 0$.

In particular, this rate differs from the minimax optimal rate for covariance matrices with other structures such as bandable or sparse covariance matrices (Cai and Zhou, 2012; Cai et al., 2010). For non-Gaussian distributions, the minimax optimal rate is not known. The minimax optimal rate $\{\log(np)/(np)\}^{\frac{2\beta}{2\beta+1}}$ is attained by the tapering estimator of Cai et al. (2013) for both classes $\mathcal{G}_\beta(M)$ and $\mathcal{P}_\beta(M_0, M_1)$, but by the banding estimator of Cai et al. (2013) only for the class $\mathcal{G}_\beta(M)$.

To denote that an estimator attains a specific rate, the Big O notation will be used in this thesis. Recall, that for two functions $f(n, p)$ and $g(n, p)$ the notation $f(n, p) = \mathcal{O}\{g(n, p)\}$ for $n, p \to \infty$ means that there exist constants $c_1, c_2 > 0$ such that for all $m, n \geq c_2$ the inequality $|f(n, p)| \leq c_1 g(n, p)$ holds. Moreover, we write $f(n) = o\{g(n)\}$ as $n \to \infty$, if for every positive constant c_1 there exists a constant c_2 such that $|f(n)| \leq c_1 g(n)$ for all $n \geq c_2$.

Next, we give the minimax optimal rate for spectral density estimation. Denote by $\mathcal{H}^+_{\beta,\text{per}}$ the restriction of $\mathcal{H}_{\beta,\text{per}}$ to the functions which are non-negative and even. The difference to the space $\mathcal{H}^+_{\beta,\text{per}}(M_0, M_1)$ is, that for $f \in \mathcal{H}^+_{\beta,\text{per}}$ only $\|f\|_\infty < \infty$ and $\|f^{(\gamma)}(\cdot) - f^{(\gamma)}\|_\infty / |h|^\alpha < \infty$ is required. The minimax optimal convergence rate for estimating $f \in \mathcal{H}^+_{\beta,\text{per}}$ from a single Gaussian stationary time series $\mathbf{Y} \sim \mathcal{N}_p(0, \mathbf{\Sigma})$ was obtained by Bentkus (1985) under the $L_p, 1 \leq p \leq \infty$, norm.

Theorem (Theorem 2.1 of (Bentkus, 1985)) *Let* $\mathbf{Y} \sim \mathcal{N}_p(0, \boldsymbol{\Sigma})$ *such that the spectral density* f *belongs to* $\mathcal{H}^{+}_{\beta,per}$*, then for* $1 \leq p < \infty$

$$\inf_{\hat{f}} \sup_{f \in \mathcal{H}^{+}_{\beta,per}} \mathbb{E}\|\hat{f} - f\|_p^2 \geq c p^{-\frac{2\beta}{2\beta+1}}.$$

and for $p = \infty$

$$\inf_{\hat{f}} \sup_{f \in \mathcal{H}^{+}_{\beta,per}} \mathbb{E}\|\hat{f} - f\|_\infty^2 \geq c \left\{ \frac{\log(p)}{p} \right\}^{\frac{2\beta}{2\beta+1}}.$$

Despite the existence of several proposed methods for consistent covariance matrix estimation, the task of covariance estimation remains an active area of research. In the following, we will outline the main challenges.

1.8 Practical challenges for covariance matrix estimation

1. **Tuning Parameter Selection**: Nonparametric methods involve tuning parameters that control the regularization procedures. Selecting the optimal values for these parameters is crucial for obtaining accurate covariance estimates. However, determining the appropriate values can be challenging. In the time series setting with $n = 1$, to the best of our knowledge, there is no fully data-driven approach for selecting the banding/tapering/thresholding parameter, since cross-validation only works for large sample size n.

2. **Computational Complexity**: For high-dimensional data, i.e., n, p or both are large, estimation of the covariance matrix and tuning parameter selection can become computationally intensive, as it involves multiplication of large matrices and the Fourier transform.

3. **Positive Definiteness**: The estimated covariance matrix should be positive (semi-)definite. However, estimators such as tapering or banding estimators may yield estimated covariance matrices that violate this property. Ensuring positive (semi-)definiteness is crucial, because violation of this property can lead to problems in the subsequent analysis, such as unstable parameter estimates or incorrect inference. Additional steps, such as regularizing the estimates or enforcing positive (semi-)definiteness constraints, may be necessary to address this challenge.

1.9 Summary of the results

The dissertation is based on three publications, which can be found in Chapters 2, 3 and 4. Chapter 2 is based on the article Klockmann and Krivobokova (2023a), Chapter 3 on the article Klockmann and Krivobokova (2023b), and Chapter 4 on the article Krivobokova et al. (2022). The methods are implemented in the R packages *eBsc* and *vstdct* and are available on CRAN (Rosales et al., 2022; Klockmann and Krivobokova, 2023c).

In summary, we propose two new methods for estimating Toeplitz covariance matrices that are entirely data-driven, computationally efficient and guarantee that the resulting covariance matrix is positive semidefinite. The two methods are based on two new versions of the periodogram that allow more efficient use of the data. In particular, they provide estimates of the spectral density on a twice as fine grid as the Fourier frequencies, i.e., at p equi-spaced points in the interval $[0, \pi]$ instead of just $\lfloor (p-1)/2 \rfloor$. A rigorous theoretical analysis is performed, investigating consistency and convergence rates of the proposed methods. Additionally, an approximate expression for the covariance of the log-average periodogram is derived, providing a further improvement for one of the proposed covariance matrix estimators. In the following, a summary for each of the three chapters is given.

Chapter 2: Efficient nonparametric estimation of Toeplitz covariance matrices

A new nonparametric estimator for Toeplitz covariance matrices and its inverse is proposed. We consider the setting $\mathbf{Y}_1, \ldots, \mathbf{Y}_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}_p(\mathbf{0}_p, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} \in \mathcal{F}_\beta = \mathcal{P}_\beta(M_0, M_1) \cap \{f : \inf_x f(x) \geq \delta\}$ for some $\delta > 0$, i.e., the spectral density function f is Hölder smooth and bounded away from zero. The Toeplitz covariance estimator is based on a data transformation that translates the problem of Toeplitz covariance matrix estimation into an approximate Gaussian nonparametric regression for mean estimation.

First, we derive an alternative version of the periodogram based on the Discrete Cosine Transform matrix, which has a frequency resolution twice as fine as the original periodogram. The resulting gamma regression problem is transformed with data binning and variance stabilizing transforms into an approximate Gaussian regression setting for the log-spectral density. Data binning consists of dividing each $\mathbf{Y}_i$ in $T < p$ bins and averaging the observations in each bin for all n observations. Application of a variance stabilizing transform for the gamma distribution then yields a homoscedastic approximate Gaussian regression. A log-spectral density estimator is obtained with periodic smoothing splines with a data-driven smoothing parameter $h > 0$ and penalization order $q = 1, 2, ...$ The final estimators for the covariance matrix and its inverse are obtained via the inverse Fourier transform.

The main result of the chapter is the proof that all proposed estimators are minimax optimal: the spectral density estimator attains the minimax optimal convergence rate under the L_∞ norm, the Toeplitz covariance and the precision matrix estimator attain the minimax optimal rate under the spectral norm. Denote by $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$ the precision matrix. Then we obtain the following minimax optimal result.

Theorem *Let* $\mathbf{Y}_1, ..., \mathbf{Y}_n \overset{i.i.d.}{\sim} \mathcal{N}_p(0, \mathbf{\Sigma})$ *with* $\mathbf{\Sigma} = \mathbf{\Sigma}(f) \in \mathcal{F}_\beta$ *and* $\beta{=}\gamma{+}\alpha > 1/2$. *If* $h > 0$ *such that* $h \to 0$ *and* $hT \to \infty$, *then with* $T = \lfloor p^v \rfloor$ *for any* $v \in ((4 - 2\min\{\beta, 1\})/3, 1)$ *and* $q = \max\{1, \gamma\}$, *the spectral density estimator* $\hat{f}$, *the corresponding covariance matrix estimator* $\hat{\mathbf{\Sigma}}$ *and the precision matrix estimator* $\hat{\mathbf{\Omega}}$ *satisfy*

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

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

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

The estimators $\hat{\mathbf{\Sigma}}$ and $\hat{\mathbf{\Omega}}$ are positive definite by construction, fully data-driven and computationally very fast. The performance of the estimators is demonstrated in simulations and real data analysis. The method is implemented in the R package *vstdct*.

I derived most of the proofs independently and wrote the manuscript. My supervisor helped me develop the idea for the article and assisted me with some of the proofs. The simulations and the R package *vstdct* were implemented and designed by myself.

Chapter 3: On second-order statistics of the log-average periodogram

In this chapter, an approximate expression for the covariance of the log-average periodogram for a zero mean stationary Gaussian process is derived. The log-average periodogram consists of averaging local frequencies of the periodogram before applying the logarithm. It is a crucial quantity in various applications such as estimating the spectral density (see Chapter 2), the long memory parameter, or the cepstral coefficients. The covariance formula is obtained under the assumption that the spectral density function is Hölder continuous, bounded and bounded away from zero. In a simulation study, it is investigated how taking into account the derived covariance formula explicitly can improve the spectral density estimator of Chapter 2 in small samples. The findings offer promising insights for improving other estimators based on the log-average periodogram, particularly for scenarios involving smaller signal lengths. Moreover, a simple expression for the non-integer moments of a non-central chi-squared distribution is provided.

I had the idea for the article, derived the proof, and wrote the manuscript independently. My supervisor helped me weaken the assumptions and improve the simulations.

Chapter 4: Joint nonparametric estimation of mean and autocovariances for Gaussian processes

A new nonparametric method to simultaneously estimate the mean and the covariance function from a non-stationary Gaussian time series is proposed. The data is given by a single observation $\mathbf{Y} \sim \mathcal{N}_p(\mu, \mathbf{\Sigma})$ where $\mathbf{\Sigma} \in \mathcal{G}_\beta(M)$ or in $\mathcal{P}_\beta(M_0, M_1)$ with a spectral density bounded away from zero. The mean function μ is a smooth unknown function. The estimators are based on an empirical Bayes approach that is fully data-driven, adaptive, numerically efficient, and allows for the construction of confidence sets for the mean function. Another version of the periodogram is derived by showing that the Demmler-Reinsch basis matrix asymptotically diagonalizes Toeplitz matrices of the class $\mathcal{G}_\beta$. The performance of the estimators is demonstrated in simulations and real data analysis. The method is implemented in the R package *eBsc*.

My contribution includes the proof of Lemma 4.1 and part of the simulations.

2

Efficient nonparametric estimation of Toeplitz covariance matrices

This chapter is based on:
Klockmann, K. and Krivobokova, T. (2023a). Efficient nonparametric estimation of Toeplitz covariance matrices. *arXiv preprint arXiv:2303.10018*

Abstract

A new 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. The method is implemented in the R package *vstdct* that accompanies the paper.

Keywords: Discrete cosine transform, spectral density, Whittle likelihood, periodogram, variance stabilizing transform.

2.1 Introduction

Estimation of covariance and precision matrices is a fundamental problem in statistical data analysis with countless applications in the natural and social sciences. Covariance matrices with a Toeplitz structure arise in the study of stationary stochastic processes, which are an important modeling tool in many applications, such as radar target detection, speech recognition, modeling internet economic activity, electrical brain activity and the motion of crystal structures (Du et al., 2020; Roberts and Ephraim, 2000; Quah, 2000; Franaszczuk et al., 1985; Grenander and Szegö, 1958).

The data for estimation are given as n independent and identically distributed realizations of a p-dimensional vector having a zero mean and a covariance matrix $\mathbf{\Sigma}$ with a Toeplitz structure. Thereby, p is assumed to grow while n may be equal to 1 (a single realization of a stationary time series) 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 $\mathbf{\Sigma}$ in the spectral norm (Wu and Pourahmadi, 2009; Pourahmadi, 2013, chap. 2). Therefore, tapering, banding and thresholding of the sample covariance matrix have been proposed to regularize this estimator (Wu and Xiao, 2012; 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 $n = 1$ (Wu and Xiao, 2012). However, all of these estimators have several practical drawbacks. First, they are not guaranteed to be positive semidefinite. To enforce positive semidefiniteness, additional manipulations with the estimators must be performed, see e.g., Section 5 in 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 and Levina (2008b) 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. For $n = 1$, to the best of our knowledge, there is no fully data-driven approach for selecting the banding/tapering/thresholding parameter. Wu and Pourahmadi (2009) suggested first to split the time series into non-overlapping subseries and then apply the cross-validation criterion of Bickel and Levina (2008b). However, it turns out that the right choice of the subseries length is crucial for this approach, but there is no data-based method available for this.

In this work, an alternative way to estimate a Toeplitz covariance matrix and its inverse is chosen. 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 $\mathbf{\Sigma}$ 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-range processes with bounded spectral densities. Moreover, the computation is very efficient, does not require iterative or resampling schemes and allows to apply any inference and adaptive estimation procedures developed in the context of nonparametric Gaussian regression.

Estimation of the spectral density from a stationary time series is a research topic with a long history. Earlier nonparametric methods are based on smoothing of the (log-)periodogram, which itself is not a consistent estimator (Bartlett, 1950; Welch, 1967; Thomson, 1982; Wahba, 1980). 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 and 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 (Choudhuri et al., 2004; Edwards et al., 2019; Maturana-Russel and 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 Hölder continuous spectral densities from Gaussian stationary time series was obtained by Bentkus (1985) under the $L_p, 1 \leq p \leq \infty$, norm. Only few works on spectral density estimation show the optimality of the corresponding estimators. In particular, Kooperberg et al. (1995) and Pawitan and O'Sullivan (1994) 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 the Toeplitz covariance matrices. The main contribution of this work is to show that our proposed spectral density estimator is not only numerically very efficient, 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.

The paper is structured as follows. In Section 2.2, the model is introduced and approximate diagonalization of Toeplitz covariance matrices with the discrete cosine transform is discussed. Moreover, an alternative version of the Whittle's likelihood is proposed. In Section 2.3, new estimators for the Toeplitz covariance matrix and the precision matrix are derived, while in Section 2.4 their theoretical properties are presented. Section 2.5 contains simulation results, Section 2.6 presents a real data example, and Section 2.7 closes the paper with a discussion. The proofs and auxiliary results are given in Section 2.8.

2.2 Set up and diagonalization of Toeplitz matrices

Let $\mathbf{Y}_1, \ldots, \mathbf{Y}_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}_p(\mathbf{0}_p, \mathbf{\Sigma})$, where $\mathbf{\Sigma}$ is a $(p \times p)$-dimensional positive semidefinite covariance matrix with a Toeplitz structure, that is, $\mathbf{\Sigma} = (\sigma_{|i-j|})_{i,j=1}^{p} \succ 0$. The sample size n may tend to infinity or to 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 $\mathbf{Y} \sim \mathcal{N}_p(\mathbf{0}_p, \mathbf{\Sigma})$. The dimension p is assumed to grow.
The spectral density function f, corresponding to a Toeplitz covariance matrix $\mathbf{\Sigma}$, is given by

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

so that for $f \in L_2(-\pi, \pi)$ the inverse Fourier transform implies

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

Hence, $\mathbf{\Sigma}$ is completely characterized by f, and the non-negativity of the spectral density function implies the positive semidefiniteness of the covariance matrix. Moreover, the decay of the autocovariance 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 $\|\mathbf{\Sigma}\| \leq \|f\|_\infty := \sup_{x\in[-\pi,\pi]} |f(x)|$, where $\|\cdot\|$ denotes the spectral norm (Grenander and Szegö, 1958, Chap. 5.2).

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

$$\mathcal{P}_\beta(M_0, M_1) = \left\{ \mathbf{\Sigma}(f) \in \mathbb{R}^{p\times p} : \|f\|_\infty \leq M_0, \|f^{(\gamma)}(\cdot + h) - f^{(\gamma)}(\cdot)\|_\infty \leq M_1|h|^\alpha, \mathbf{\Sigma}_\infty \succ 0 \right\},$$

where $\mathbf{\Sigma}_\infty \succ 0$ means positive semidefinite. The optimal convergence rate for estimating Toeplitz covariance matrices over $\mathcal{P}_\beta(M_0, M_1)$ depends crucially on β. It is well known that the k-th Fourier coefficient of a function whose γ-th derivative is α-Hölder continuous decays at least with order $\mathcal{O}(k^{-\beta})$ (see Zygmund, 2002). Hence, β determines the decay rate of the autocovariances σ_k, which are the Fourier coefficients of the spectral density f, as $k \to \infty$. In particular, this implies that for $\beta \in (0, 1]$, the class $\mathcal{P}_\beta(M_0, M_1)$ includes Toeplitz covariance matrices corresponding to long-range processes with bounded spectral densities, since the sequence of corresponding autocovariances is not summable. As there is a one-to-one correspondence between $\mathbf{\Sigma}$ and its spectral density function f for $\mathbf{\Sigma} \in \mathcal{P}_\beta(M_0, M_1)$, we indicate in the following by $\mathbf{\Sigma} \in \mathcal{P}_\beta(M_0, M_1)$ and $f \in \mathcal{P}_\beta(M_0, M_1)$ which object is considered.

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

Lemma 2.1 *Let $\mathbf{\Sigma} \in \mathcal{P}_\beta(M_0, M_1)$ and let $x_j = (j-1)/(p-1)$, $j = 1, ..., p$, then*

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

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

$$\mathbf{D} = \sqrt{2}(p-1)^{-1/2}\left[\cos\left\{\pi(i-1)\frac{j-1}{p-1}\right\}\right]_{i,j=1}^{p} \text{ divided by } \sqrt{2} \text{ when } i, j \in \{1, p\}$$

is the Discrete Cosine Transform I (DCT-I) matrix.

The proof can be found in Section 2.8.1. This result shows that the DCT-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 the spectral density literature the Discrete Fourier Transform (DFT) matrix $\mathbf{F} = p^{-1/2}\{\exp(2\pi \mathrm{i} ij/p)\}_{i,j=1}^{p}$, where i is the imaginary unit, is typically employed to approximately diagonalize Toeplitz covariance matrices. Using the fact that $(\mathbf{F}\Sigma\mathbf{F}^t)_{i,i} = f(2\pi i/p) + \mathcal{O}(1)$, Whittle (1957) introduced an approximation for the likelihood of a single Gaussian stationary time series (case $n = 1$), the so-called Whittle likelihood

$$\mathcal{L}(\mathbf{Y}|f) \propto \exp\left[-\sum_{j=1}^{\lfloor p/2 \rfloor} \log\{f(2\pi j/p)\} + \frac{I_j}{f(2\pi j/p)}\right]. \tag{2.1}$$

The quantity $I_j = |\mathbf{F}_j^t\mathbf{Y}|^2$, where $\mathbf{F}_j$ denotes the j-th column of $\mathbf{F}$, is known as the periodogram at the j-th Fourier frequency. Note that due to periodogram symmetry, only $\lfloor p/2 \rfloor$ data points $I_1, ..., I_{\lfloor p/2 \rfloor}$ are available for estimating the mean $f(2\pi j/p)$, $j = 1, \ldots, \lfloor p/2 \rfloor$, 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); Taqqu and Teverovsky (1997).

Lemma 2.1 yields the following alternative version of the Whittle likelihood

$$\mathcal{L}(\mathbf{Y}|f) \propto \exp\left[-\sum_{j=1}^{p} \log\{f(\pi x_j)\} + \frac{W_j}{f(\pi x_j)}\right], \tag{2.2}$$

where $W_j = (\mathbf{D}_j^t\mathbf{Y})^2$. Note that 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 $\mathbf{Y}$ to estimate the parameter of interest, such as the spectral density or the Hurst parameter.

Equations (2.1) or (2.2) invite for the estimation of f by maximizing the (penalized) likelihood over certain linear spaces (e.g., spline spaces), as suggested by Kooperberg et al. (1995) or by Pawitan and O'Sullivan (1994). However, such an approach requires well-designed numerical methods to solve the corresponding optimization problem, since the spectral density in the second term of (2.1) or (2.2) is in the denominator, which does not allow to obtain a closed-form expression for the estimator and often leads to numerical instabilities. Also, the choice of the 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 note that if $\mathbf{Y} \sim \mathcal{N}_p(\mathbf{0}_p, \mathbf{\Sigma})$, with $\mathbf{\Sigma} \in \mathcal{P}_\beta(M_0, M_1)$, then $\mathbf{D}^t\mathbf{Y} \sim \mathcal{N}_p(\mathbf{0}_p, \mathbf{D}^t\mathbf{\Sigma}\mathbf{D})$. Hence, for $W_j = (\mathbf{D}_j^t\mathbf{Y})^2$, $j = 1, \dots, p$, it follows with Lemma 2.1 that

$$W_j \sim \Gamma\left(1/2, 2f(\pi x_j) + \mathcal{O}\left\{p^{-1} + \log(p)p^{-\beta}\right\}\right), \tag{2.3}$$

where $\Gamma(a, b)$ denotes a gamma distribution with a shape parameter a and a scale parameter b. Note that the random variables $W_1, \dots, W_p$ are only asymptotically independent. Obviously, $\mathbb{E}(W_j) = f(\pi x_j) + o(1)$, $j = 1, \dots, p$. To estimate f from $W_1, \dots, W_p$, one could use a generalized nonparametric regression framework with a gamma distributed response, see e.g., the classical monograph by Hastie and 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 to employ a variance stabilizing transform of Cai and Zhou (2010) that converts the gamma regression into an approximate Gaussian regression. In the next section we present the methodology in more detail for a general setting with $n \geq 1$.

2.3 Methodology

Let $L_\delta = \{f : \inf_x f(x) \geq \delta\}$ and set $\mathcal{F}_\beta = \mathcal{P}_\beta(M_0, M_1) \cap L_\delta$. We consider estimation of $\mathbf{\Sigma}$ and $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$ from a sample $\mathbf{Y}_1, ..., \mathbf{Y}_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}_p(\mathbf{0}_p, \mathbf{\Sigma})$ where $\mathbf{\Sigma} \in \mathcal{F}_\beta$. For $\mathbf{Y}_i \sim \mathcal{N}_p(\mathbf{0}_p, \mathbf{\Sigma})$, $i = 1, \dots, n$, it was shown in the previous section that with Lemma 2.1 the data can be transformed into gamma distributed random variables $W_{i,j} = (\mathbf{D}_j^t\mathbf{Y}_i)^2$, $i = 1, \dots, n$, $j = 1, \dots, p$, where for each fixed i the random variable $W_{i,j}$ has the same distribution as W_j given in (2.3). Now the approach of Cai and Zhou (2010) is adapted to the setting $n \geq 1$.

First, the transformed data points $W_{i,j}$ are binned, that is, fewer new variables Q_k, $k = 1, \dots, T$, with $T < p$, are built via $Q_k = \sum_{j=(k-1)p/T+1}^{kp/T} \sum_{i=1}^{n} W_{i,j}$, $k = 1, \dots, T$. Note that the number of observations in a bin is $m = np/T$. In Theorem 2.1 in Section 2.4, we show that setting $T = \lfloor p^v \rfloor$ for any $v \in ((4 - 2\min\{\beta, 1\})/3, 1)$ 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 (VST) $G(x) = \log(x/m)/\sqrt{2}$ to each Q_k, yields new random variables $Y_k^* = \log(Q_k/m)/\sqrt{2}$ that are approximately Gaussian, as shown in Cai and 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 $Y_{T-1}^*, \ldots, Y_2^*, Y_1^*, \ldots, Y_T^*$ for estimation. Renumerating the observations Y_k^* and scaling the design points into the interval $[0, 1]$ for convenience leads to an approximate Gaussian regression problem

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

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

$$\widehat{H(f)}(x) = \arg\min_{s \in S_{\text{per}}(2q-1)} \left[\frac{1}{2T-2}\sum_{k=1}^{2T-2}\{Y_k^* - s(x_k)\}^2 + h^{2q}\int_0^1 \{s^{(q)}(x)\}^2 \, \mathrm{d}x\right], \tag{2.4}$$

where $h > 0$ denotes a smoothing parameter, $q \in \mathbb{N}$ is the penalty order and $S_{\text{per}}(2q-1)$ a space of periodic splines of degree $2q - 1$. The smoothing parameter h can be chosen either with generalized cross-validation (GCV) as derived in Craven and Wahba (1978) or with the restricted maximum likelihood, see Krivobokova (2013).

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

$$\hat{\mathbf{\Sigma}} = (\hat{\sigma}_{|i-j|})_{i,j=1}^{p} \text{ with } \hat{\sigma}_j = \int_0^1 \hat{f}(x)\cos(k\pi x)\, \mathrm{d}x \text{ for } k = 1, ..., p. \tag{2.5}$$

The precision matrix $\mathbf{\Omega}$ is estimated by the inverse Fourier transform of the reciprocal of the spectral density estimator, i.e.,

$$\hat{\mathbf{\Omega}} = (\hat{\omega}_{|i-j|})_{i,j=1}^{p} \text{ with } \hat{\omega}_k = \int_0^1 \hat{f}(x)^{-1}\cos(k\pi x)\, \mathrm{d}x \text{ for } k = 1, ..., p. \tag{2.6}$$

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

1. **Data Transformation**: Define $W_{i,j} = (\mathbf{D}_j^t \mathbf{Y}_i)^2$, $i = 1, \ldots, n$, $j = 1, \ldots, p$, where $\mathbf{D}$ is the $(p \times p)$-dimensional DCT-I matrix as given in Lemma 2.1 and $\mathbf{D}_j$ is its j-th column.

2. **Binning**: Set $T = \lfloor p^v \rfloor$ for any $v \in ((4 - 2\min\{\beta, 1\})/3, 1)$ and calculate
$$Q_k = \sum_{j=(k-1)p/T+1}^{kp/T} \sum_{i=1}^{n} W_{i,j}, \; k = 1, \ldots, T.$$

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

4. **Gaussian Regression**: Renumerate observations Y_k^* and estimate $H(f)$ with a periodic smoothing spline in an approximate Gaussian regression model
$$Y_k^* = H\{f(x_k)\} + \epsilon_k, \;\; x_k = \frac{k-1}{2T-2}, \; k = 1, \ldots, 2T-2,$$
where ϵ_k are asymptotically i.i.d. Gaussian variables.

5. **Inverse VST**: Estimate the spectral density f with $\hat{f} = H^{-1}\left\{\widehat{H(f)}\right\}$, where $H^{-1}(y) = m\exp\left\{\sqrt{2}y - \phi(m/2)\right\}/2$, for a digamma function ϕ.

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

Note that $\hat{\mathbf{\Sigma}}$ and $\hat{\mathbf{\Omega}}$ are positive definite matrices by construction, since their spectral density functions $\hat{f}$ and $\hat{f}^{-1}$ are non-negative, respectively. Unlike the banding and tapering estimators, the autocovariance estimators $\hat{\sigma}_k$ are controlled by a smoothing parameter h, which can be estimated fully data-driven with several available automatic methods, which are numerically efficient and well-studied. In addition, one can also use methods for adaptive mean estimation, see e.g., Serra et al. (2017), which in turn leads to adaptive Toeplitz covariance matrix estimation. All inferential procedures developed in the Gaussian regression context can also be adopted accordingly.

2.4 Theoretical properties

In this section, we study the asymptotic properties of the estimators $\hat{f}$, $\hat{\mathbf{\Sigma}}$ and $\hat{\mathbf{\Omega}}$. The results are established under the asymptotic scenario where $p \to \infty$ and $p/n \to c \in (0, \infty]$, that is, the dimension p grows, while the sample size n either remains fixed or also grows but not faster than p. This corresponds to the asymptotic scenario when the sample covariance matrix is inconsistent. Let $\hat{f}$ be the spectral density estimator defined in Section 2.3, i.e., $\hat{f} = m \exp\{\sqrt{2}\widehat{H(f)} - \phi(m/2)\}/2$, where $\widehat{H(f)}$ is given in Equation (2.4), $m = np/T$ and ϕ is the digamma function. Furthermore, let $\hat{\mathbf{\Sigma}}$ be the Toeplitz covariance matrix estimator and $\hat{\mathbf{\Omega}}$ the corresponding precision matrix defined in equations (2.5) and (2.6), respectively. The following theorem shows that both $\hat{\mathbf{\Sigma}}$ and $\hat{\mathbf{\Omega}}$ attain the minimax optimal rate of convergence over the class $\mathcal{F}_\beta$.

Theorem 2.1 *Let $\mathbf{Y}_1, \ldots., \mathbf{Y}_n \overset{i.i.d.}{\sim} \mathcal{N}_p(\mathbf{0}_p, \mathbf{\Sigma})$ with $\mathbf{\Sigma} = \mathbf{\Sigma}(f) \in \mathcal{F}_\beta$ and $\beta = \gamma + \alpha > 1/2$. If $h > 0$ such that $h \to 0$ and $hT \to \infty$, then with $T = \lfloor p^v \rfloor$ for any $v \in ((4 - 2\min\{\beta, 1\})/3, 1)$ and $q = \max\{1, \gamma\}$, the spectral density estimator $\hat{f}$, the corresponding covariance matrix estimator $\hat{\mathbf{\Sigma}}$ and the precision matrix estimator $\hat{\mathbf{\Omega}}$ satisfy*

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

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

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

The proof of Theorem 2.1 can be found in Section 2.8 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. In the original work, Cai and 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 2.1. This error adds to the error that arises via binning and VST and that describes the deviation from a Gaussian distribution, as derived in Cai and 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 tools than those used in Cai and Zhou (2010).

To get the L_∞ rate for $\hat{f}$, we first derive the L_∞ rate 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 in Schwarz and Krivobokova (2016), thereby carefully treating various (dependent) errors that describe deviations from a Gaussian nonparametric regression with independent errors and mean $f(\pi x_i)$. Note also that although the periodic smoothing spline estimator is obtained on T binned points, the rate is given in terms of the vector dimension p. Then, 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 sufficient to note that $\mathbb{E}\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|^2 \leq \mathbb{E}\|\hat{f} - f\|_\infty^2$.

2.5 Simulation study

In this section, we compare the performance of the proposed Toeplitz covariance estimator, denoted as **VST-DCT**, with the **tapering** estimator of Cai et al. (2013) and with the **sample covariance** matrix. We consider Gaussian vectors $\mathbf{Y}_1, ..., \mathbf{Y}_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}_p(\mathbf{0}_p, \mathbf{\Sigma}_p)$ with **(A)** $p = 5000$, $n = 1$, **(B)** $p = 1000$, $n = 50$ and **(C)** $p = 5000$, $n = 10$, and with the autocovariance functions $\sigma : \mathbb{Z} \to \mathbb{R}$

(1) of a polynomial decay $\sigma_\tau = 1.44(1 + |\tau|)^{-5.1}$,

(2) of an ARMA$(2,2)$ model with coefficients $(0.7, -0.4, -0.2, 0.2)$ and innovations variance 1.44,

(3) 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, $\mathrm{Var}(Y_i) = 1.44$ in all three examples. Figure 1 shows the spectral densities and the corresponding autocovariance functions for the three examples.

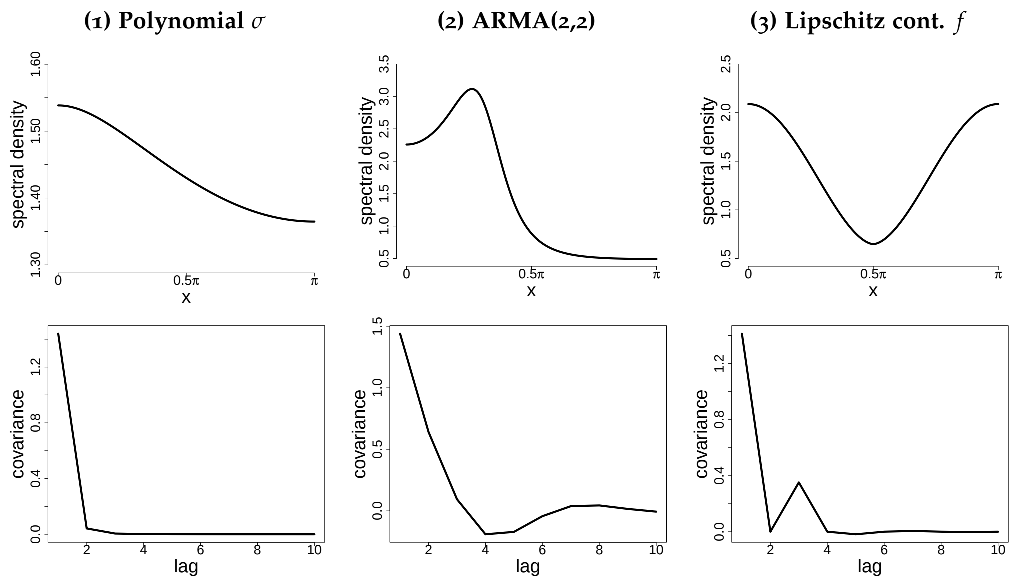

Figure 1: Spectral density functions (first row) and autocovariance functions (second row) for examples **1**, **2**, **3**.

A Monte Carlo simulation with 100 iterations is performed using R (version 4.1.2, seed 42). For our VST-DCT estimator, we use a cubic periodic spline, i.e., $q = 2$ is set in Equation (2.4). The binning parameters are set to $T = 500$ bins with $m = 10$ points for **(A)** and $T = 500$ bins with $m = 100$ points for both **(B)** and **(C)**.

The sample covariance matrix $\tilde{\boldsymbol{\Sigma}}$ is define as $(n-1)^{-1}\sum_{i=1}^{n}\mathbf{Y}_i\mathbf{Y}_i^t$ with averaged diagonals to obtain Toeplitz structure. To select the regularization parameter for our estimator, we implemented the restricted maximum likelihood (ML) method, generalized cross-validation (GCV) and the corresponding oracle versions, i.e., as if $\boldsymbol{\Sigma}$ were known. The tapering parameter can be selected using cross-validation (Bickel and Levina, 2008b), 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 $n_1 = 2/3n$ and a test set of size $n_2 = n/3$. Let $\boldsymbol{\Sigma}_1^{\nu}$ and $\boldsymbol{\Sigma}_2^{\nu}$ be the sample covariance matrices from the ν-th split. The tapering parameter k is then estimated as

$$\hat{k}_{\text{cv}} = \min_{k=2,3,\ldots,p/2} \frac{1}{30}\sum_{\nu=1}^{30} \|\text{Tap}_k(\boldsymbol{\Sigma}_1^{\nu}) - \boldsymbol{\Sigma}_2^{\nu}\|,$$

where $\text{Tap}_k(\boldsymbol{\Sigma}_2^{\nu})$ is the tapering estimator of Cai et al. (2013) with parameter k. If $n = 1$, that is, under scenario **(A)**, Wu and Pourahmadi (2009) suggest to split the time series $\mathbf{Y}$ into l non-overlapping subseries of length p/l and then proceed as before to select the tuning parameter k. To the best of our knowledge, there is no data-driven method for selecting this parameter l. Using the true covariance matrix $\mathbf{\Sigma}$, we selected $l = 30$ subseries for the example **1** and $l = 15$ subseries for the examples **2** and **3**. Then, the parameter k can then be chosen by 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, i.e., $\hat{k}_{\text{or}} = \arg\min_{k=2,3,\dots,p/2} \|\text{Tap}_k(\tilde{\mathbf{\Sigma}}) - \mathbf{\Sigma}\|$. To speed up the computation, one can replace the spectral norm by the ℓ_1 norm, as suggested by Bickel and Levina (2008b).

In Tables 1, 2 and 3, the errors of the Toeplitz covariance estimators with respect to the spectral norm and the computation time for one Monte Carlo iteration are given 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.

	(1) polynomial σ		(2) ARMA$(2,2)$		(3) Lipschitz cont. f		time
	$\|\hat{\Sigma}-\Sigma\|^2$	$\|\hat{f}-f\|_2^2$	$\|\hat{\Sigma}-\Sigma\|^2$	$\|\hat{f}-f\|_2^2$	$\|\hat{\Sigma}-\Sigma\|^2$	$\|\hat{f}-f\|_2^2$	in sec
VST-DCT (GCV)	0.0069	0.0026	0.0817	0.0107	0.0340	0.0061	4.3658
VST-DCT (ML)	0.0059	0.0022	0.0700	0.0097	0.0375	0.0063	4.3607
tapering (semi-oracle)	0.0056	0.0022	0.0973	0.0152	0.0355	0.0098	4.9373
sample covariance	162.4089	38.1068	484.2556	48.3138	244.3804	38.0949	0.3744
VST-DCT (GCV-oracle)	0.0042	0.0017	0.0662	0.0094	0.0332	0.0057	–
VST-DCT (ML-oracle)	0.0046	0.0019	0.0663	0.0092	0.0378	0.0062	–
tapering (oracle)	0.0042	0.0017	0.0578	0.0085	0.0155	0.0037	–

Table 1: **Scenario (A) with $\mathbf{p = 5000, n = 1}$**: Errors of the Toeplitz covariance matrix and the spectral density estimators with respect to the spectral and L_2 norm, respectively, as well as the average computation time of the covariance estimators in seconds for one Monte Carlo sample (last column).

The results show that the tapering and VST-DCT estimator perform overall similar in terms of the spectral norm risk. This is not surprising as both estimators are proved to be rate-optimal. Moreover, both the tapering and VST-DCT estimators are clearly superior to the inconsistent sample covariance matrix. A closer look at the numbers shows that the VST-DCT method has better constants, i.e., VST-DCT estimators have

	(1) polynomial σ		(2) ARMA(2,2)		(3) Lipschitz cont. f		time
	$\|\widehat{\Sigma}-\Sigma\|^2$	$\|\hat{f}-f\|_2^2$	$\|\widehat{\Sigma}-\Sigma\|^2$	$\|\hat{f}-f\|_2^2$	$\|\widehat{\Sigma}-\Sigma\|^2$	$\|\hat{f}-f\|_2^2$	in sec
VST-DCT (GCV)	0.0010	0.0003	0.0123	0.0015	0.0053	0.0008	35.8043
VST-DCT (ML)	0.0008	0.0002	0.0123	0.0014	0.0061	0.0009	35.6555
tapering (CV)	0.0026	0.0012	0.0220	0.0031	0.0079	0.0018	22.8628
sample covariance	0.8833	0.5867	2.7060	0.8482	1.4063	0.6362	0.1200
VST-DCT (GCV-oracle)	0.0006	0.0002	0.0112	0.0014	0.0046	0.0007	–
VST-DCT (ML-oracle)	0.0007	0.0002	0.0121	0.0014	0.0060	0.0009	–
tapering (oracle)	0.0019	0.0011	0.0168	0.0027	0.0065	0.0016	–

Table 2: **Scenario (B)** with $\mathbf{p = 1000, n = 50}$: Errors of the Toeplitz covariance matrix and the spectral density estimators with respect to the spectral and L_2 norm, respectively, as well as the average computation time of the covariance estimators in seconds for one Monte Carlo sample (last column).

	(1) polynomial σ		(2) ARMA(2,2)		(3) Lipschitz cont. f		time
	$\|\widehat{\Sigma}-\Sigma\|^2$	$\|\hat{f}-f\|_2^2$	$\|\widehat{\Sigma}-\Sigma\|^2$	$\|\hat{f}-f\|_2^2$	$\|\widehat{\Sigma}-\Sigma\|^2$	$\|\hat{f}-f\|_2^2$	in sec
VST-DCT (GCV)	0.0009	0.0003	0.0114	0.0014	0.0059	0.0008	4.2319
VST-DCT (ML)	0.0008	0.0002	0.0118	0.0014	0.0068	0.0009	4.2376
tapering (CV)	0.0342	0.0260	0.1606	0.0401	0.0759	0.0287	622.9578
sample covariance	9.2735	4.6053	29.5994	6.3001	12519.7295	4.6666	1.1721
VST-DCT (GCV-oracle)	0.0006	0.0002	0.0102	0.0013	0.0050	0.0007	–
VST-DCT (ML-oracle)	0.0007	0.0002	0.0115	0.0013	0.0066	0.0008	–
tapering (oracle)	0.0308	0.0259	0.1435	0.0394	0.0697	0.0280	–

Table 3: **Scenario (C)** with $\mathbf{p = 5000, n = 10}$: Errors of the Toeplitz covariance matrix and the spectral density estimators with respect to the spectral and L_2 norm, respectively, as well as the average computation time of the covariance estimators in seconds for one Monte Carlo sample (last column).

somewhat smaller errors in the spectral norm than the tapering estimators across all examples, but especially under scenario **(C)**. The oracle estimators show similar behavior, but are slightly less variable compared to the data-driven estimators. We also experimented with different choices for the bin number T in the available interval given by Theorem 2.1. Omitting the extrema near the interval boundaries, the simulation results are quite stable against different choices for T. In general, both the tapering and VST-DCT estimators perform best for example 1, second best for example 3 and worst for example 2, which traces back to functions complexity.

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 expected that for larger p the tapering estimator is much more computationally intensive than the corresponding VST-DCT estimator.

To test how robust our approach is to deviations from the Gaussian assumption, we simulated the data from gamma and uniform distributions and conducted a simulation study for the same scenarios and examples. The results are very similar to those of the Gaussian distribution, see Section 2.8.5 for the details.

2.6 Application to protein dynamics

We revisit the data analysis of protein dynamics performed in Krivobokova et al. (2012) and Singer et al. (2016). We consider data generated by the molecular dynamics (MD) simulations for the yeast aquaporin (Aqy1) – the gated water channel of the yeast *Pichi pastoris*. MD 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 Aqy1 observed in a 100 nanosecond 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 centers 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

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

where $A_{t,i} \in \mathbb{R}^3$, $i = 1, \dots, 783$ denotes the i-th atom of the protein at time t, $B_j \in \mathbb{R}^3$, $j = 1, 2, 3, 4$, is the j-th 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 center of mass. It can therefore be concluded that a linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ holds, where $\mathbf{Y} = (y_1, \dots, y_{20\,000})^t \in \mathbb{R}^{20\,000}$, $\mathbf{X} = (X_1^t, \dots, X_{20\,000}^t)^t \in \mathbb{R}^{20\,000 \times (4 \cdot 783)}$, $\boldsymbol{\beta} \in \mathbb{R}^{4 \cdot 783}$, $\boldsymbol{\epsilon} \in \mathbb{R}^{20\,000}$. This linear model has two specific features which are intrinsic to the problem: first,

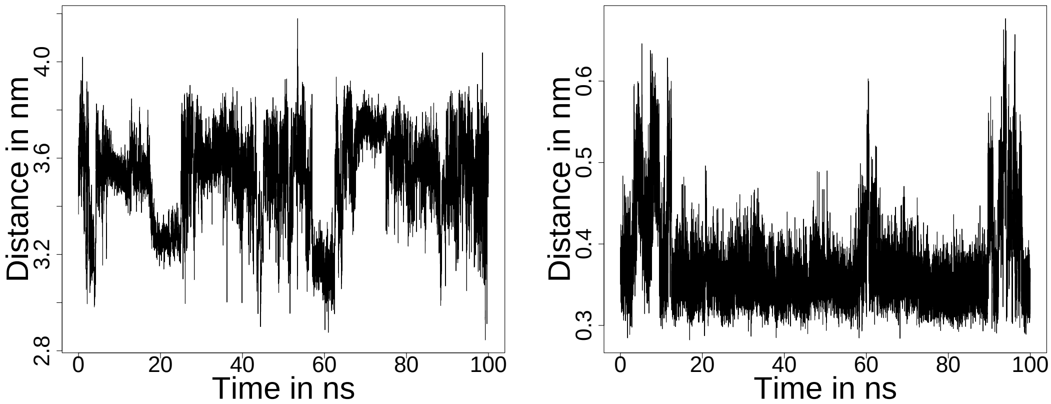

Figure 2: Distance between the first atom and the first center of mass of aquaporin (left) and the opening diameter y_t over time t (right).

the observations are not independent over time and second, X_t is high-dimensional at each t and only few columns of $\mathbf{X}$ are relevant for $\mathbf{Y}$. 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 $\mathbf{Y}$. 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 $\mathbf{Y}$ is a linear combination of columns of $\mathbf{X}$, it can be assumed that $\mathbf{Y}$ and all columns of $\mathbf{X}$ have the same correlation structure. Hence, it is sufficient to estimate $\mathbf{\Sigma} = \text{cov}(\mathbf{Y})$ to decorrelate the data for the PLS algorithm, i.e., $\mathbf{\Sigma}^{-1/2}\mathbf{Y} = \mathbf{\Sigma}^{-1/2}\mathbf{X}\boldsymbol{\beta} + \mathbf{\Sigma}^{-1/2}\boldsymbol{\epsilon}$ results in a standard linear regression with independent errors.

Our goal now is to estimate $\mathbf{\Sigma}$ and compare the performance of the PLS algorithm on original and decorrelated data. For this purpose, we divided the data set into a training and 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 $\mathbf{Y}$ with a p-value< 0.01. The Hurst exponent of $\mathbf{Y}$ is 0.85, indicating moderate long-range dependence supported by a rather slow decay of the sample autocovariances (see grey line in the left plot of Figure 3). Therefore, we set $q = 1$ for the VST-DCT estimator to match the low smoothness of the corresponding spectral density. Moreover, the smoothing parameter is selected with the restricted maximum likelihood method and $T = 550$ bins are used. Figure 3 (black line in the left plot) confirms that the covariance matrix estimated with our VST-DCT method almost completely decorrelates the channel diameter $\mathbf{Y}$ on the training data set. Next, we estimated the regression coefficients $\boldsymbol{\beta}$ with the usual PLS algorithm, ignoring the dependence in the data. Finally, we estimated $\boldsymbol{\beta}$ with PLS that takes into account dependence using our covariance estimator $\hat{\boldsymbol{\Sigma}}$. Based on these regression coefficient estimators, the prediction on the test set was calculated.

The plot on the right side of Figure 2 shows the Pearson correlation between the true channel diameter on the test set and the prediction on the same test set based on raw (grey) and decorrelated data (black). Obviously, the performance of the PLS algorithm on the decorrelated data is significantly better for the small number of components. In particular, with just one PLS 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 0.54, while it is close to zero for PLS that ignores the dependence in the data. Krivobokova et al. (2012) showed that the estimator of $\boldsymbol{\beta}$ based on one PLS component is exactly the ensemble-weighted maximally correlated mode (ewMCM), which is defined as the collective mode of atoms that has the highest probability to achieve a specific alteration of the response $\mathbf{Y}$. 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.

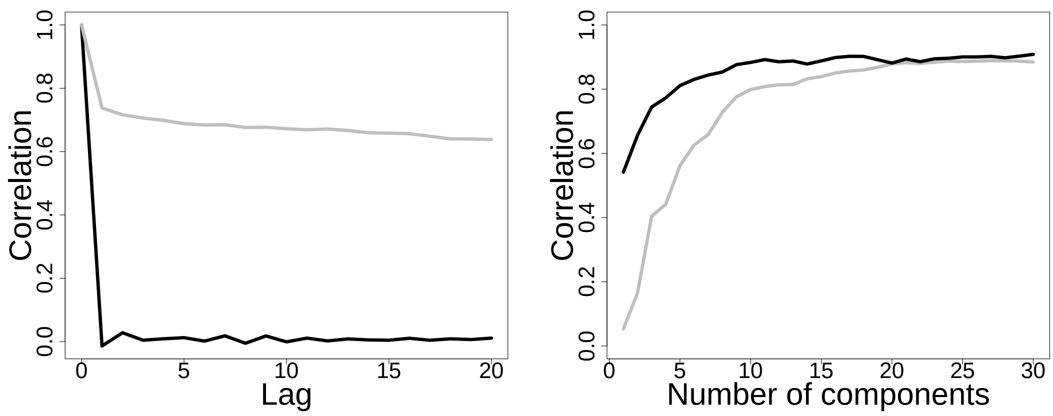

Figure 3: On the left, the autocorrelation function of $\mathbf{Y}$ (grey) and of $\hat{\boldsymbol{\Sigma}}^{-1/2}\mathbf{Y}$ (black), where $\hat{\boldsymbol{\Sigma}}$ is estimated with the VST-DCT method; On the right, correlation between the true values of the test data set and the predicted values, where the prediction is based on partial least squares (grey) and corrected partial least squares (black).

Estimating $\mathbf{\Sigma}$ with a tapered covariance estimator has two practical problems. First, since we only have a single realization $\mathbf{Y}$, i.e. $n = 1$, there is no data-driven method for selecting the tapering parameter. Second, the tapering estimator turned out not to be positive semidefinite for the given dataset. To solve the second problem, we truncated the corresponding spectral density estimator $\hat{f}_{\text{tap}}$ to a small positive value, i.e., $\hat{f}^{+}_{\text{tap}} = \max\{\hat{f}_{\text{tap}}, 1/\log(p)\}$ (see Cai et al., 2013; McMurry and Politis, 2010). 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 PLS component, respectively.

Altogether, our proposed estimator is fully data-driven, fast even for large sample sizes, automatically positive definite and can handle certain long-range processes. In contrast, the tapering estimator is not data-driven and must be manipulated to become positive definite. Our method is implemented in the R package *vstdct*, which is available on CRAN.

2.7 Discussion

In this paper, we proposed a simple, fast, fully data-driven, automatically positive definite and minimax optimal estimator of Toeplitz covariance matrices from a large class that also includes covariance matrices of certain long-range processes.

Our estimator is derived under the assumption that the data are Gaussian. However, simulations show that the suggested approach yields robust estimators even when the data are not normally distributed. In the context of spectral density estimation and Whittle likelihood, several papers investigated the effects of violating the Gaussian assumption on the asymptotic distribution of the periodogram and hence on the Whittle likelihood. In particular, it has been shown that the periodogram distribution $I_j \sim \Gamma\{1, f(2\pi j/p) + o(1)\}, j = 1, \dots, \lfloor p/2 \rfloor$ holds asymptotically for linear time series with i.i.d. innovations (Brockwell and Davis, 2009, Thm. 10.3.2), for mixing processes (Rosenblatt, 2012, Thm. 5.3), as well as for non-linear processes (Shao and Wu, 2007). Since DFT and DCT matrices are closely related, we expect that Equation (2.3) also holds asymptotically for these non-Gaussian time series, but consider a rigorous analysis to be beyond the scope of this paper.

In fact, our numerical experiments have even shown that if the spectral density is estimated from $W_j = f(\pi x_j) + \epsilon_j$, that is, as if W_j were Gaussian instead of gamma distributed, then the resulting spectral density estimator has almost the same L_∞ risk (and hence the corresponding covariance matrix has almost the same spectral norm). Of course, such an estimator would lead to a wrong inference about $f(\pi x_j)$, since the growing variance of W_j would be ignored.

Since our approach translates Toeplitz covariance matrix estimation into an approximate Gaussian nonparametric regression for mean estimation, all approaches developed in the context of Gaussian nonparametric regression, such as (locally) adaptive estimation, as well as the corresponding (simultaneous) inference, can be directly applied. Bayesian tools for adaptive estimation and inference in Gaussian nonparametric regression as proposed in Serra et al. (2017) can also be employed.

2.8 Proofs and auxiliary results

Throughout this section, 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.

2.8.1 Proof of Lemma 2.1

We embed the p-dimensional Toeplitz matrix $\mathbf{\Sigma} = \text{toep}(\sigma_0, \ldots, \sigma_{p-1})$ in a $(2p-2)$-dimensional circulant matrix $\tilde{\mathbf{\Sigma}} = \text{toep}(\sigma_0, \ldots, \sigma_{p-1}, \sigma_{p-2}, \ldots, \sigma_1)$. Then, $\tilde{\mathbf{\Sigma}} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^*$, where $\mathbf{U} = (2p-2)^{-1/2}[\exp\{\pi \mathrm{i}(k-1)(l-1)/(p-1)\}]_{k,l=1}^{2p-2}$ with the conjugate transpose $\mathbf{U}^*$, and $\mathbf{\Lambda}$ is a diagonal matrix with the k-th diagonal value for $k = 1, ..., p$ given by

$$\lambda_k = \sum_{l=1}^{p} \sigma_{l-1} \exp\{\pi \mathrm{i} x_k (l-1)\} + \sum_{l=p+1}^{2p-2} \sigma_{2p-1-l} \exp\{\pi \mathrm{i} x_k (l-1)\},$$

where $x_k = (k-1)/(p-1)$. See e.g., Gray et al. (2006) for properties of circulant matrices.

By symmetry, $\lambda_k = \lambda_{2p-k}$ for $k = p+1, ..., 2p-2$. Furthermore, $\mathbf{\Sigma} = \mathbf{V}^*\mathbf{\Lambda}\mathbf{V}$, where $\mathbf{V} \in \mathbb{C}^{(2p-2)\times p}$ contains the first p columns of $\mathbf{U}$. Then, for $i, j = 1, \ldots, p$

$$(\mathbf{D\Sigma D})_{i,j} = (\mathbf{DV}^*\mathbf{\Lambda}\mathbf{VD})_{i,j} = \sum_{r=1}^{2p-2} \lambda_r \sum_{q=1}^{p} \mathbf{D}_{iq}\mathbf{V}^*_{qr} \sum_{s=1}^{p} \mathbf{V}_{rs}\mathbf{D}_{sj}.$$

Define the scaling factor $a_s = 1/\sqrt{2}$ for $s = 1, p$ and $a_s = 1$ for $s = 2, ..., p-2$, appearing in the definition of the Discrete Cosine Transform I matrix. Then,

$$\begin{aligned}
\sum_{s=1}^{p} \mathbf{V}_{rs}\mathbf{D}_{sj} &= (p-1)^{-1} \sum_{s=1}^{p} a_s a_j \exp\{\pi \mathrm{i}(r-1)x_s\} \cos\{\pi(j-1)x_s\} \\
&\quad \pm \frac{a_j}{p-1} \exp(0)\cos(0) \pm \frac{a_j}{p-1} \exp\{\pi \mathrm{i}(r-1)\}) \cos\{\pi(j-1)\} \\
&= a_j \frac{\sqrt{2}-2}{p-1} \mathbb{1}_{\{|j-r| \text{ even}\}} + \frac{a_j}{p-1} \sum_{s=1}^{p} \exp\{\pi \mathrm{i}(r-1)x_s\} \cos\{\pi(j-1)x_s\} \\
&= a_j b(j,r) + a_j \mathrm{i} c(j,r),
\end{aligned}$$

where $\mathbb{1}$ is the indicator function and

$$\begin{aligned}
b(j,r) &= \frac{\sqrt{2}-2}{p-1} \mathbb{1}_{\{|j-r| \text{ even}\}} + \frac{1}{p-1} \sum_{s=1}^{p} \cos\{\pi(r-1)x_s\} \cos\{\pi(j-1)x_s\}, \\
c(j,r) &= \frac{1}{p-1} \sum_{s=1}^{p} \sin\{\pi(r-1)x_s\} \cos\{\pi(j-1)x_s\}.
\end{aligned}$$

By the symmetry properties of cosine and sine, $b(j,r) = b(j, 2p-r)$ and $c(j,r) = -c(j, 2p-r)$ for $r = p+1, \ldots, 2p-2$. Using $(\sum_{s=1}^{p} \mathbf{V}_{rs}\mathbf{D}_{sj})^* = \sum_{q=1}^{p} \mathbf{D}_{jq}\mathbf{V}^*_{qr}$, we obtain

$$(\mathbf{D\Sigma D})_{i,j} = \begin{cases} a_j^2 \sum_{r=1}^{2p-2} \lambda_r \left\{ b(j,r)^2 + c(j,r)^2 \right\}, & i{=}j \\ a_i a_j \sum_{r=1}^{2p-2} \lambda_r (b(i,r) + \mathrm{i}c(i,r))(b(j,r) - \mathrm{i}c(j,r)), & i{\neq}j. \end{cases} \tag{2.7}$$

Some calculations show that for $r = 1, \ldots p$

$$\frac{1}{p-1} \sum_{s=1}^{p} \cos\{\pi(r-1)x_s\} \cos\{\pi(j-1)x_s\} = \begin{cases} \frac{1}{p-1}, & r{\neq}j, r{-}j \text{ even} \\ 0, & r{-}j \text{ odd} \\ \frac{p+1}{2p-2}, & r{=}j{\neq}1, p \\ \frac{p}{p-1}, & r{=}j{=}1, p \end{cases} \tag{2.8}$$

and

$$
\begin{aligned}
&\frac{1}{p-1}\sum_{s=1}^{p}\sin\{\pi(r-1)x_s\}\cos\{\pi(j-1)x_s\}\\
&=\frac{\{1-(-1)^{r-j}\}}{4(p-1)}\left[\cot\{\pi(r-j)/(2p-2)\}+\cot\{\pi(j+r-2)/(2p-2)\}\right]. \qquad (2.9)
\end{aligned}
$$

Using the Taylor expansion of $\cot(x)$, i.e., $\cot(x)=x^{-1}-x/3-x^3/45+...$, for $0<|x|<\pi$, one obtains for $r=1,\dots p$

$$
c(j,r)=\frac{(r-1)\{1-(-1)^{r-j}\}}{\pi\{(r-1)-(j-1)\}\{(r-1)+(j-1)\}}+\mathcal{O}\left\{\frac{2\pi(2r-2)}{3(2p-2)^2}\right\}, \qquad (2.10)
$$

where the $\mathcal{O}$ term does not depend on j and the hidden constant does not depend on r,p.

Case $i=j$

If $i=j$, equations (2.7) - (2.9) imply

$$
\begin{aligned}
(\mathbf{D\Sigma D})_{j,j} =& a_j^2\sum_{\substack{r=1\\ r\neq j,\\ r\neq 2p-j}}^{2p-2}\lambda_r\left[\frac{(\sqrt{2}-2)^2+1+2(\sqrt{2}-2)(p-1)}{(p-1)^2}\mathbb{1}_{\{|r-j|\text{ even}\}}+c(j,r)^2\right]\\
&+a_j^2\lambda_j\cdot\begin{cases}2\left(\frac{\sqrt{2}-2}{p-1}+\frac{p+1}{2p-2}\right)^2, & j\neq 1,p\\ \left(\frac{\sqrt{2}-2}{p-1}+\frac{p}{p-1}\right)^2, & j=1,p\end{cases}\\
=&a_j^2\sum_{\substack{r=1\\ r\neq j,\\ r\neq 2p-j}}^{2p-2}\lambda_r\left[\frac{(\sqrt{2}-1)^2}{(p-1)^2}\mathbb{1}_{\{|r-j|\text{ even}\}}+c(j,r)^2\right]+\lambda_j\cdot\begin{cases}\frac{2(\sqrt{2}-3/2+p/2)^2}{(p-1)^2}, & j\neq 1,p\\ \frac{(\sqrt{2}-2+p)^2}{2(p-1)^2}, & j=1,p\end{cases}\\
=&\mathcal{O}(p^{-1})+2a_j^2\sum_{r=1}^{p-1}\lambda_r c(j,r)^2+\lambda_j\{0.5+\mathcal{O}(p^{-1})\},
\end{aligned}
$$

where we used the symmetry properties of λ_r and that by assumption $\|f\|<M_0$. Note that the $\mathcal{O}$ terms do not depend on j. Since the complex exponential function is Lipschitz continuous with Lipschitz constant $L=1$, it holds $\lambda_r=\lambda_j+L_{r,j}|r-j|(p-1)^{-1}$ where $-1\leq L_{r,j}\leq 1$ is a constant depending on r,j.

Then,

$$\sum_{r=1}^{p-1} \lambda_r c(j,r)^2 = \lambda_j \sum_{r=1}^{p-1} c(j,r)^2 + (p-1)^{-1} \sum_{r=1}^{p-1} L_{r,j}|r-j|c(r,j)^2. \tag{2.11}$$

Since $\sum_{r=1}^{p-1} \lambda_r c(1,r)^2 = -\sum_{r=1}^{p-1} \lambda_r c(p,r)^2$, it is sufficient to consider $j = 1, ..., p-1$. We begin with first sum in (2.11). For a shorter notation, we use $k := r-1$ and $l := j-1$ in the following. Then, summing the squares of the first term of $c(j,r)$ in (2.10), for $l = 0, ..., p-2$, yields

$$\frac{1}{\pi^2}\sum_{k=0}^{p-2} \frac{k^2\{1-(-1)^{k-l}\}^2}{(k-l)^2(k+l)^2} = \frac{4}{\pi^2}\begin{cases} \sum_{x=1}^{\infty} \frac{1}{(2x-1)^2} - R_1(l,p) = \frac{\pi^2}{8} - R_1(l,p), & l = 0 \\ \sum_{x=0}^{\infty} \frac{(2x)^2}{(2x-l)^2(2x+l)^2} - R_2(l,p) = \frac{\pi^2}{16} - R_2(l,p), & l \text{ odd} \\ \sum_{x=1}^{\infty} \frac{(2x-1)^2}{(2x-1-l)^2(2x-1+l)^2} - R_2(l,p) = \frac{\pi^2}{16} - R_3(l,p), & \text{else} \end{cases}$$

see Chapter 23.2 of Abramowitz and Stegun (1964) on sums of reciprocal powers. If p is even, then the residual terms are given by

$$\begin{aligned} R_1(l,p) &= \sum_{x=(p-2)/2}^{\infty} \frac{1}{(2x-1)^2} = \frac{\phi^{(1)}\{(p-3)/2\}}{4} \\ R_2(l,p) &= \sum_{x=p/2}^{\infty} \frac{(2x)^2}{(2x-l)^2(2x+l)^2} \\ &= \frac{2\phi\{(p+l)/2\} - 2\phi\{(p-l)/2\} + l[\phi^{(1)}\{(p+l)/2\} + \phi^{(1)}\{(p-l)/2\}]}{16l} \\ R_3(l,p) &= \sum_{x=p/2}^{\infty} \frac{(2x-1)^2}{(2x-1-l)^2(2x-1+l)^2} \\ &= \frac{2\phi\{(p-1+l)/2\} - 2\phi\{(p-1-l)/2\} + l[\phi^{(1)}\{(p-1+l)/2\} + \phi^{(1)}\{(p-1-l)/2\}]}{16l}, \end{aligned}$$

where ϕ and $\phi^{(1)}$ denote the digamma function and its derivative. If p is odd, similar remainder terms can be derived. To see that $R_i(l,p) = \mathcal{O}(p^{-1})$ for $i = 1,2,3$ and uniformly in l we use that asymptotically $\phi(x) \sim \log(x) - 1/(2x)$ and $\phi^{(1)}(x) \sim 1/x + 1/(2x^2)$. The mixed term $\frac{4}{3}\sum_{k=0}^{p-2} \frac{k^2\{1-(-1)^{k-l}\}^2}{(k-l)(k+l)(2p-2)^2}$ and the squared $\mathcal{O}$ term $\frac{16\pi^2}{9}\sum_{k=0}^{p-2}\frac{k^2}{(2p-2)^4}$ are both of the order p^{-1}. Furthermore,

$$\frac{1}{p-1}\sum_{r=1}^{p-1} L_{r,j}|r-j|c(r,j)^2 = \frac{1}{p-1}\sum_{k=0}^{p-2}\frac{k^2\{1-(-1)^{k-l}\}^2 L_{k,l}|k-l|}{\pi^2(k-l)^2(k+l)^2} = \mathcal{O}\{\log(p)p^{-1}\},$$

since the harmonic sum diverges at a rate of $\log(p)$. Finally, $\lambda_j = f(x_j) + \mathcal{O}\{\log(p)p^{-\beta}\}$ by the uniform approximation properties of the discrete Fourier series for Hölder continuous functions (Zygmund, 2002). All together, we have shown that $(\mathbf{D\Sigma D})_{j,j} = f(x_j) + \mathcal{O}\{p^{-1} + \log(p)p^{-\beta}\}$, where the $\mathcal{O}$ terms are uniform over $j = 1, ..., p$.

Case $i \neq j$ and $|i-j|$ is even

In this case, $(\mathbf{D\Sigma D})_{i,j} = a_i a_j \sum_{r=1}^{2p-2} \lambda_r \{b(i,r)b(j,r) + c(i,r)c(j,r)\}$ since $|r-j|$, and $|r-i|$ are both either odd or even. Note that $b(i,r)b(j,r) = \mathcal{O}(p^{-2})$ for $r \neq i, j$ and $b(i,r)b(j,r) = \mathcal{O}(p^{-1})$ if $i = r$ or $j = r$, where the $\mathcal{O}(\cdot)$ terms are uniformly in i, j. Since $b(i,r) \geq 0$ and $\|f\|_\infty < M_0$, it follows $a_i a_j \sum_{r=1}^{2p-2} \lambda_r b(i,r)b(j,r) = \mathcal{O}(p^{-1})$ uniformly in i, j. To show that $a_i a_j \sum_{r=1}^{2p-2} \lambda_r c(i,r)c(j,r) = \mathcal{O}(p^{-1})$, we proceed similarly as before. Setting $k{=}r{-}1$, $l{=}j{-}1$, $m{=}i{-}1$ and using that $l{\neq}m$ and $|l{-}m|$ is even, one obtains

$$\frac{1}{\pi^2} \sum_{k=0}^{p-2} \frac{k^2\{1-(-1)^{k-l}\}\{(-1)^{k-m}-1\}}{(k-l)(k+l)(k-m)(k+m)}$$
$$= \frac{4}{\pi^2} \begin{cases} \sum_{x=0}^{\infty} \frac{(2x)^2}{(2x-l)(2x+l)(2x-m)(2x+m)} - R_1(l,m,p) = -R_1(l,m,p), & l, m \text{ odd} \\ \sum_{x=1}^{\infty} \frac{(2x-1)^2}{(2x-1-l)(2x-1+l)(2x-1-m)(2x-1+m)} - R_2(l,m,p) = -R_2(l,m,p), & l, m \text{ even}, \end{cases}$$

where for even p the residual terms are given by

$$R_1(l,m,p) = \sum_{x=p/2}^{\infty} \frac{(2x)^2}{(2x-l)(2x+l)(2x-m)(2x+m)}$$
$$R_2(l,m,p) = \sum_{x=p/2}^{\infty} \frac{(2x-1)^2}{(2x-1-l)(2x-1+l)(2x-1-m)(2x-1+m)}.$$

Similarly as before, one can show that $R_1(l,m,p)$ and $R_2(l,m,p)$ are of the order $\mathcal{O}(p^{-1})$ where the hidden constant does not depend on l, m. If p is odd, analogous residual terms can be derived. Using similar techniques as before, one can show that the two residual terms and the remaining mixed and square terms vanish at a rate of the order $\mathcal{O}(p^{-1})$ and uniformly in i, j.

Case $i \neq j$ and $|i-j|$ is odd

$|r-i|$ and $|r-j|$ are either odd and even, or even and odd. Without loss of generality, assume that $|r-i|$ is even. Then, $(\mathbf{D\Sigma D})_{i,j} = a_i a_j \sum_{r=1}^{2p-2} \lambda_r b(i,r)c(j,r)$. Since $b(i,\cdot)$ is an even function, $c(j,\cdot)$ is an odd function and $\lambda_r = \lambda_{2p-r}$, it follows $(\mathbf{D\Sigma D})_{i,j} = 0$. □

2.8.2 Periodic smoothing splines

In this section, we recapitulate the results obtained in Schwarz and Krivobokova (2016) for the equivalent kernels of periodic smoothing spline estimators. Given data points $\{(x_i, Y_i)\}_{i=1}^N$ with true relationship $Y_i = f(x_i) + \epsilon_i$, where $x_i = i/N$ and the residuals ϵ_i satisfy standard assumptions, the periodic smoothing spline estimator $\hat{f}$ is the solution to

$$\min_{s \in S_{\text{per}}(2q-1, \underline{x}_N)} \left[\frac{1}{N} \sum_{i=1}^N \{Y_i - s(x_i)\}^2 + h^{2q} \int_0^1 \{s^{(q)}(x)\}^2 \, \mathrm{d}x \right],$$

where $h > 0$ is the smoothing parameter and $S_{\text{per}}(2q-1, \underline{x}_N)$ the periodic spline space of degree $2q-1$ based on knots $\underline{x}_N = (x_i)_{i=1}^N = (i/N)_{i=0}^N$. Schwarz and Krivobokova (2016) have shown that $\hat{f}(x) = N^{-1} \sum_{i=1}^N W(x, x_i) Y_i$, where $W(x, x_i)$ is known as effective kernel of a periodic smoothing spline estimator and its explicit expression is given by

$$W(x, x_i) = \sum_{j=1}^N \frac{\Phi\{Nx + q, \exp(-2\pi \mathrm{i} j/N)\} \overline{\Phi\{Nx_i + q, \exp(-2\pi \mathrm{i} j/N)\}}}{Q_{2q-2}(j/N) + h^{2q}(2\pi j)^{2q} \operatorname{sinc}(\pi j/N)^{2q}},$$

where $\Phi(\cdot, \cdot)$ denotes exponential splines (Schoenberg, 1973) and polynomials $Q_{2q-2}(x)$ are formally defined as $Q_{2q-2} = \sum_{l=-\infty}^{\infty} \operatorname{sinc}\{\pi(x+l)\}^{2q}$; for the explicit expressions see Section 4.1.1 in Schwarz (2012). In particular, $W(\cdot, x_i) \in S_{\text{per}}(2q-1, \underline{x}_N)$ see, Schwarz (2012, p.61). A scaled version $K(x, t)$ of $W(x, t)$ in terms of the bandwidth parameter h is defined as $K(x, t) = hW(hx, ht)$. We denote by $K_h(x, t) = K(x/h, t/h) = hW(x, t)$. Also, we use that $W(x, t) = \sum_{l=-\infty}^{\infty} \mathcal{W}(x, t + l)$ (Schwarz, 2012, Lemma 14), where

$$\mathcal{W}(x, t) = \int_0^N \frac{\Phi\{Nx + q, \exp(-2\pi \mathrm{i} u)\} \overline{\Phi\{Nx_i + q, \exp(-2\pi \mathrm{i} u)\}}}{Q_{2q-2}(u) + h^{2q}(2\pi j)^{2q} \operatorname{sinc}(\pi u)^{2q}} \, \mathrm{d}u$$

is the asymptotic equivalent kernel of smoothing spline estimators on $\mathbb{R}$. Finally, we denote $\mathcal{K}_h(x, t) = \mathcal{K}(x/h, t/h) = h\mathcal{W}(x, t)$.

2.8.3 Proof of Theorem 2.1

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 $\mathbb{E}\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\| \leq \mathbb{E}\|\hat{f} - f\|_\infty^2$ the first claim of the theorem follows. Finally, we prove the second statement on the precision matrices. For the sake of clarity, some technical lemmas used in the proof are listed separately in Section 2.8.4.

I. An upper bound on $\mathbb{E}\|\widehat{H(f)} - H(f)\|_\infty^2$

Proposition 2.1 *Let $\mathbf{\Sigma} \in \mathcal{F}_\beta$ such that $\beta > 1/2$. If $h > 0$ such that $h \to 0$ and $hT \to \infty$, then with $T = \lfloor p^v \rfloor$ for any $v \in ((4 - 2\min\{1, \beta\})/3, 1)$, the estimator $\widehat{H(f)}$ described in Section 2.3 with $q = \max\{1, \gamma\}$ satisfies*

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

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

$$\mathbb{E}\|\widehat{H(f)} - H(f)\|_\infty^2 \leq 2\mathbb{E}\|\widehat{H(f)} - \mathbb{E}\{\widehat{H(f)}\}\|_\infty^2 + 2\|\mathbb{E}\{\widehat{H(f)}\} - H(f)\|_\infty^2. \tag{2.12}$$

Set $\tilde{T} = 2T - 2$ and $x_k = (k-1)/\tilde{T}$ for $k = 1, \ldots, \tilde{T}$. Using Lemma 2.4, we can write

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

where $x \in [0,1]$, $|\epsilon_k| \lesssim (np)^{-1} + \log(p)(np)^{-\beta}$, $\zeta_k \sim \mathcal{N}(0, m^{-1})$ such that for $k \neq l$ holds $\mathrm{Cov}(\zeta_k, \zeta_l) = \mathcal{O}\{p^{-2} + \log(p)^2 p^{-2\beta}\}$, and ξ_k is a mean zero random variable satisfying $\mathbb{E}|\xi_k|^\ell \lesssim \log^{2\ell}(m)[m^{-\ell} + \left\{T^{-1} + T^{-1}\log(p)p^{1-\beta}\right\}^\ell]$ for each integer $\ell > 1$. Mirroring and renumerating $\zeta_k, \eta_k, \epsilon_k$ is similar as for Y_k^*, $k = 1, \ldots, \tilde{T}$.

Upper bound on the variance

Using the above representation, one can write

$$\mathbb{E}\|\widehat{H(f)} - \mathbb{E}\{\widehat{H(f)}\}\|_\infty^2 = \mathbb{E}\left\|\frac{1}{\tilde{T}} \sum_{k=1}^{\tilde{T}} W(x, x_k)(\zeta_k + \xi_k)\right\|_\infty^2. \tag{2.13}$$

First, we reduce the supremum to a maximum over a finite number of points. Recall that $W(\cdot, x_i) \in S_{\mathrm{per}}(2q - 1, \underline{x}_N)$, see Section 2.8.2. Thus, for $q > 1$, the function $W(\cdot, x_k)$ is differentiable and in particular Lipschitz continuous with some Lipschitz constant $L > 0$.

Then, it holds almost surely that

$$
\begin{aligned}
&\sup_{x\in[0,1]}\left|\frac{1}{\tilde{T}}\sum_{k=1}^{\tilde{T}} W(x,x_k)(\zeta_k+\xi_k)\right|^2 \\
&\le \left[\max_{1\le j\le \tilde{T}}\left|\frac{1}{\tilde{T}}\sum_{k=1}^{\tilde{T}} W(x_j,x_k)(\zeta_k+\xi_k)\right| + \sup_{\substack{x,x'\in[0,1],\\ |x-x'|<1/\tilde{T}}}\left|\frac{1}{\tilde{T}}\sum_{k=1}^{\tilde{T}}\{W(x,x_k)-W(x',x_k)\}(\zeta_k+\xi_k)\right|\right]^2 \\
&\le 2\max_{1\le j\le \tilde{T}}\left|\frac{1}{\tilde{T}}\sum_{k=1}^{\tilde{T}} W(x_j,x_k)(\zeta_k+\xi_k)\right|^2 + \frac{2L^2}{\tilde{T}^4}\left(\sum_{k=1}^{\tilde{T}}|\zeta_k+\xi_k|\right)^2.
\end{aligned}
$$

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

$$
(2.13) \lesssim \mathbb{E}\left\{\max_{1\le j\le \tilde{T}}\left|\frac{1}{\tilde{T}}\sum_{k=1}^{\tilde{T}} W(x_j,x_k)(\zeta_k+\xi_k)\right|^2\right\} + o\{(np)^{-1}\}. \tag{2.14}
$$

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 $x_j = j/\tilde{T}$, $j=1,...,\tilde{T}$, for any realization of the stochastic weights. This implies that (2.14) is also valid for $q=1$. Using the inequality $(a+b)^2 \le 2a^2+2b^2$ once more, we then obtain

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

We start with bounding

$$
T_1 = \mathbb{E}\left\{\max_{1\le j\le \tilde{T}}\left|\frac{1}{\tilde{T}h}\sum_{k=1}^{\tilde{T}} K_h(x_j,x_k)\zeta_k\right|^2\right\}
$$

with Lemma 1.6 of Tsybakov (2009). This requires a bound on $\|\frac{1}{\tilde{T}h}\sum_{k=1}^{\tilde{T}} K_h(x_j,x_k)\zeta_k\|_{\psi_2}^2$, where $\|\cdot\|_{\psi_2}$ denotes the sub-Gaussian norm. In case of a Gaussian random variable the norm equals the variance. See Vershynin (2018, Ch. 2.5) for a definition of $\|\cdot\|_{\psi_2}$ and further properties.

Thus, with Lemma 2.2 and Lemma 2.4, we obtain

$$\begin{aligned}&\left\|\frac{1}{\tilde{T}h}\sum_{k=1}^{\tilde{T}}K_h(x_j,x_k)\zeta_k\right\|_{\psi_2}^2=\mathbb{V}\left\{\frac{1}{\tilde{T}h}\sum_{k=1}^{\tilde{T}}K_h(x_j,x_k)\zeta_k\right\}\\&=\frac{1}{\tilde{T}^2h^2}\sum_{k=1}^{\tilde{T}}K_h(x_j,x_k)^2\mathbb{V}(\zeta_k)+\frac{1}{\tilde{T}^2h^2}\sum_{k=1}^{\tilde{T}}K_h(x_j,x_k)\sum_{l=1}^{\tilde{T}}K_h(x_j,x_l)\mathrm{Cov}(\zeta_k,\zeta_l)\\&\leq C(Thm)^{-1}+C'\{p^{-2}+\log(p)^2p^{-2\beta}\},\end{aligned}$$

where $0<C,C'$ are some constants. Lemma 1.6 of Tsybakov (2009) then yields

$$T_1\lesssim\log(2\tilde{T})[C(Thm)^{-1}+C'\{p^{-2}+\log(p)^2p^{-2\beta}\}]=\mathcal{O}\{(Thm)^{-1}\log(2\tilde{T})+h^{2\beta}\}. \tag{2.15}$$

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

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

as $h\to 0$. Now consider $1/2<\beta\leq 1$. Recall that $T=p^v$ for some fixed $v\in((4-2\min\{1,\beta\})/3,1)$. Using the inequality $\log(x)\leq x^a/a$ one can find constants $x_v,C_v>0$ depending on v but not on n,p such that $\log(2\tilde{T})\log(p)^2p^{-2\beta}T^{2\beta}\leq C_vp^{-x_v}$. This implies $\log(2\tilde{T})\log(p)^2p^{-2\beta}=\mathcal{O}(h^{2\beta})$ since $(Th)^{2\beta}\to\infty$ by assumption.

Next, we establish the asymptotic order for the second term

$$T_2=\mathbb{E}\left\{\max_{1\leq j\leq\tilde{T}}\left|\frac{1}{\tilde{T}h}\sum_{k=1}^{\tilde{T}}K_h(x_j,x_k)\xi_k\right|^2\right\}.$$

The exponential decay property of the kernel K stated in Lemma 2.2 yields

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

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

$$\xi_k^-=|\xi_k|\mathbb{1}\{|\xi_k|\leq R\},\quad\text{and }\xi_k^+=|\xi_k|\mathbb{1}\{|\xi_k|>R\}.$$

Then,

$$T_2 \lesssim \mathbb{E}\left\{\max_{1\leq j\leq \tilde{T}}\left|\frac{1}{\tilde{T}h}\sum_{k=1}^{\tilde{T}}\gamma_h(x_j,x_k)\xi_k^-\right|^2\right\} + \mathbb{E}\left\{\max_{1\leq j\leq \tilde{T}}\left|\frac{1}{\tilde{T}h}\sum_{k=1}^{\tilde{T}}\gamma_h(x_j,x_k)\xi_k^+\right|^2\right\}. \quad (2.16)$$

The first term in (2.16) 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 $|X_i| \leq R$, $i = 1,\ldots,N$, it holds $\|\sum_{i=1}^N a_i X_i\|_{\psi_2}^2 \leq 4R^2\sum_{i=1}^N a_i^2$ 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 $\mathbb{E}[\exp(\lambda\sum_{i=1}^N a_i X_i)] \leq \exp(\lambda^2 R^2\sum_{i=1}^N a_i^2)$ for all $\lambda \in \mathbb{R}$. Thus, for all $t > 0$ holds $\mathbb{P}(|\sum_{i=1}^N a_i X_i| > t) \leq 2\exp(-\lambda t + \lambda^2 R^2\sum_{i=1}^N a_i^2)$. Setting $\lambda = t/2\cdot(R^2\sum_{i=1}^N a_i^2)^{-1}$, it follows that $\sum_{i=1}^N a_i X_i$ has a sub-Gaussian distribution and the sub-Gaussian norm is bounded by $2R(\sum_{i=1}^N a_i^2)^{1/2}$. See Vershynin (2018) for further details on the sub-Gaussian distribution.

Together, we get

$$\left\|\frac{1}{\tilde{T}h}\sum_{k=1}^{\tilde{T}}\gamma_h(x_j,x_k)\xi_k^-\right\|_{\psi_2}^2 \leq \frac{4R^2}{\tilde{T}^2h^2}\sum_{k=1}^{\tilde{T}}\gamma_h(x_j,x_k)^2 \lesssim \frac{R^2}{\tilde{T}h},$$

where for the second inequality Lemma 2.2(*ii*) is used. Applying Lemma 1.6 of Tsybakov (2009), then yields

$$\mathbb{E}\left\{\max_{1\leq j\leq \tilde{T}}\left|\frac{1}{\tilde{T}h}\sum_{k=1}^{\tilde{T}}\gamma_h(x_j,x_k)\xi_k^-\right|^2\right\} \lesssim \frac{R^2}{\tilde{T}h}\log(2\tilde{T}). \quad (2.17)$$

To bound the second term in (2.16), we use the moment bounds for ξ_k derived in Lemma 2.4. Then, for all integers $\ell > 1$

$$\begin{aligned}
&\mathbb{E}\left\{\max_{1\leq j\leq \tilde{T}}\left|\frac{1}{\tilde{T}h}\sum_{k=1}^{\tilde{T}}\gamma_h(x_j,x_k)\xi_k^+\right|^2\right\} = \mathbb{E}\left\{\max_{1\leq j\leq \tilde{T}}\left|\frac{1}{\tilde{T}h}\sum_{k=1}^{\tilde{T}}\gamma_h(x_j,x_k)(\xi_k^+)^\ell(\xi_k^+)^{1-\ell}\right|^2\right\}\\
&\leq R^{2-2\ell}\mathbb{E}\left\{\max_{1\leq j\leq \tilde{T}}\frac{1}{\tilde{T}^2h^2}\sum_{k=1}^{\tilde{T}}\gamma_h(x_j,x_k)^2\cdot\sum_{k=1}^{\tilde{T}}(\xi_k^+)^{2\ell}\right\} \leq R^{2-2\ell}\mathbb{E}\left\{\frac{C_1}{\tilde{T}h}\sum_{k=1}^{\tilde{T}}\xi_k^{2\ell}\right\}\\
&\lesssim h^{-1}R^{2-2\ell}\log^{4\ell}(m)[m^{-2\ell} + \{T^{-1} + T^{-1}\log(p)p^{1-\beta}\}^{2\ell}]. \qquad (2.18)
\end{aligned}$$

Combining the error bounds (2.17) and (2.18) and choosing $R{=}m^{-1/2}$ gives

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

By assumption, $T = p^v$ and $m = np^{(1-v)}$ for some fixed $v \in ((4 - 2\min\{1, \beta\})/3, 1)$. If ℓ is an integer such that $\ell \geq 1/(1-v)$, then

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

where we used $\log(x) \leq x^a/a$ with $a = 1/(4\ell)$. Consider $1/2 < \beta \leq 1$ and let $0 < \chi < 1$ be a constant. Applying $\log(x) \leq x^a/a$ twice with $a = \chi/(2\ell)$ yields

$$m^{\ell-1} \log^{4\ell}(m) \log^{2\ell}(p) p^{2\ell(1-\beta)} T^{-2\ell} \lesssim n^{\ell-1+\chi} p^{(1-v)(\ell-1+\chi)-2\ell v+2\ell(1-\beta)+\chi}.$$

Since $v \in ((4 - 2\min\{1, \beta\})/3, 1)$ it holds

$$\begin{aligned} &(1-v)(\ell-1+\chi) - 2\ell v + 2\ell(1-\beta) + \chi < -(\ell-1+\chi) - 2 \\ \iff &\ell > \{v(1-\chi) + 3\chi\}/(3v + 2\min\{1, \beta\} - 4). \end{aligned} \tag{2.19}$$

For any fixed $v{\in}((4 - 2\min\{1, \beta\})/3, 1)$ one can find an integer ℓ, which is independent of n, p, such that the right side of (2.19) holds. Since $p/n \to c \in (0, \infty]$ and thus $n/p = \mathcal{O}(1)$ and $p^{-1} = \mathcal{O}(n^{-1})$, it follows for ℓ satisfying (2.19) that

$$n^{\ell-1+\chi} p^{(1-v)(\ell-1+\chi)-2\ell v+2\ell(1-\beta)+\chi} \leq (n/p)^{\ell-1+\chi} p^{-2} = \mathcal{O}\{(np)^{-1}\}.$$

In total, choosing an integer $\ell \geq \max[1/(1-v), \{v(1-\chi)+3\chi\}/(3v+2\min\{1,\beta\} - 4)]$ gives

$$\mathbb{E}\|\widehat{H(f)} - \mathbb{E}\{\widehat{H(f)}\}\|_\infty^2 = \mathcal{O}\left\{\frac{\log(np)}{nph} + h^{2\beta}\right\}. \tag{2.20}$$

Upper bound on the bias

Using the representation in Lemma 2.4 once more gives for each $x \in [0, 1]$

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

The bounds on ϵ_k in Lemma 2.4 and the exponential decay of the kernel function in Lemma 2.2 imply

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

Consider the case that $\beta \geq 1$. In particular, $q = \gamma = \lfloor \beta \rfloor$ 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 $\mathcal{K}$ is the asymptotic equivalent kernel defined in Section 2.8.2. Expanding $g(t)$ in a Taylor series around x and using that $h^{-1}\mathcal{K}_h$ is a kernel of order $2q$, see Lemma 2.2(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. By Lemma 2.2(i), we have $|\mathcal{K}_h(x,t)| < C\gamma^{|x-t|/h}$, for some constants $0 < C$ and $0 < \gamma < 1$. Since $g^{(q)}$ is α-Hölder continuous with some constant L, one obtains

$$\begin{aligned}
\left|\frac{1}{\tilde{T}h}\sum_{k=1}^{\tilde{T}} K_h(x,x_k)g(x_k) - g(x)\right| &\leq \frac{CL}{q!}\sum_{l=-\infty}^{\infty}\int_{x+(l-1)h}^{x+lh} \gamma^{\frac{|x-t|}{h}} |x-t|^q \frac{|\xi_{x,t}-x|^\alpha}{hq!}\,\mathrm{d}t + \mathcal{O}(\tilde{T}^{-\beta}) \\
&\leq \frac{CL}{q!}\sum_{l=-\infty}^{\infty}\int_{x+(l-1)h}^{x+lh} \gamma^{\frac{|x-t|}{h}} \frac{|x-t|^\beta}{h}\,\mathrm{d}t + \mathcal{O}(\tilde{T}^{-\beta}) \\
&\leq h^\beta \frac{CL}{q!}\sum_{l=-\infty}^{\infty} \gamma^{|l-1|}|l|^\beta + \mathcal{O}(\tilde{T}^{-\beta}) \\
&= h^\beta \frac{2CL}{q!}\sum_{l=1}^{\infty} \gamma^{l-1} l^\beta + \mathcal{O}(\tilde{T}^{-\beta}) = \mathcal{O}(h^\beta) + \mathcal{O}(\tilde{T}^{-\beta}).
\end{aligned}$$

If $1/2 < \beta \leq 1$, 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 $g = H(f)$ is β-Hölder continuous. Expanding g to the entire line and using that $h^{-1}\int_{-\infty}^{\infty} \mathcal{K}_h(x,t)\,\mathrm{d}x = 1$, see Lemma 2.2($iii$) 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 $\beta > 1/2$, $Th \to \infty$ and $h \to 0$ by assumption. Since the derived bounds are uniform for $x \in [0,1]$ it holds

$$\left\|\mathbb{E}\{\widehat{H(f)}(x)\} - H\{f(x)\}\right\|_{\infty} = \mathcal{O}(h^{\beta}) + \mathcal{O}\{(np)^{-1} + \log(p)(np)^{-\beta}\}. \tag{2.21}$$

Combining the bounds for the variance in (2.20) and for the bias in (2.21) finally gives

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

II. An upper bound on $\mathbb{E}\|\hat{f} - f\|_{\infty}^2$

Proposition 2.2 *Let $\mathbf{\Sigma} \in \mathcal{F}_{\beta}$ such that $\beta > 1/2$. If $h > 0$ such that $h \to 0$ and $hT \to \infty$, then with $T = \lfloor p^{v} \rfloor$ for any $v \in ((4 - 2\min\{1,\beta\})/3, 1)$, the estimator $\hat{f}$ described in Section 2.3 with $q = \max\{1,\gamma\}$ satisfies*

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

Proof: By the mean value theorem, it holds for some function g between $\widehat{H(f)}$ and $H(f)$ that

$$\begin{aligned}\mathbb{E}\|\hat{f} - f\|_{\infty}^2 &= \mathbb{E}\|H^{-1}\{\widehat{H(f)}\} - H^{-1}\{H(f)\}\|_{\infty}^2 = \mathbb{E}\|(H^{-1})'(g)\{\widehat{H(f)} - H(f)\}\|_{\infty}^2 \\ &\lesssim \mathbb{E}\|H^{-1}(g)\{\widehat{H(f)} - H(f)\}\|_{\infty}^2.\end{aligned}$$

Application of the Cauchy-Schwarz inequality with a constant $C > 0$ yields

$$\mathbb{E}\|\hat{f} - f\|_\infty^2 \leq C^2 \mathbb{E}\|\widehat{H(f)} - H(f)\|_\infty^2 + (\mathbb{E}\|\hat{f} - f\|_\infty^4)^{1/2} \cdot \mathbb{P}(\|H^{-1}(g)\|_\infty > C)^{1/2}. \quad (2.22)$$

To show that the second term on the right-hand side of (2.22) is negligible, we use Markov's inequality. In order to do so, we first derive the asymptotic order of the moment generating function of $\widehat{H(f)}$ for $n, p \to \infty$.

Moment generating function of $\|\widehat{H(f)}\|_\infty$

By the exponential decay property of the kernel K stated in Lemma 2.2, it holds almost surely that

$$\|\widehat{H(f)}\|_\infty \lesssim \sup_{x \in [0,1]} \frac{1}{\tilde{T}h} \sum_{k=1}^{\tilde{T}} \gamma_h(x, x_k) \cdot |Y_k^*| \, .$$

First, $\|\widehat{H(f)}\|_\infty$ is bounded with the maximum over a finite number of points. Calculating the derivative of $s : [0,1] \to \mathbb{R}$, $x \mapsto s(x) = \sum_{k=1}^{\tilde{T}} \gamma_h(x, x_k) \left|Y_k^*\right|$ for $x \neq x_k$, $k = 1, ..., \tilde{T}$, gives almost surely

$$\frac{\partial}{\partial x} s(x) = \frac{\log(\gamma)}{h} \sum_{k=1}^{\tilde{T}} \left[\gamma^{|x-t|/h} |x - x_k| + \frac{\gamma^{1/h}}{1 - \gamma^{1/h}} \left\{ \gamma^{(x-t)/h} + \gamma^{(t-x)/h} \right\} \right] |Y_k^*| \, .$$

Since $\frac{\partial}{\partial x} s(x) > 0$ almost surely, for $x \neq x_k$, the extrema occur at x_k, $k = 1, ..., \tilde{T}$. Thus, for $\lambda > 0$, the moment generating function of $\|\widehat{H(f)}\|_\infty$ is bounded by

$$\mathbb{E}[\exp\{\lambda \|\widehat{H(f)}\|_\infty\}] \leq \sum_{j=1}^{\tilde{T}} \mathbb{E} \left[\exp \left\{ \frac{\lambda}{\tilde{T}h} \sum_{k=1}^{\tilde{T}} \gamma_h(x_j, x_k) \, |Y_k^*| \right\} \right] .$$

Let $M_j = (\tilde{T}h)^{-1} \sum_{k=1}^{\tilde{T}} \gamma_h(x_j, x_k)$, which by Lemma 2.2 is bounded uniformly in j by some global constant $M > 0$. By the convexity of the exponential function, we obtain

$$\mathbb{E} \left[\exp \left\{ \frac{\lambda}{\tilde{T}h} \sum_{k=1}^{\tilde{T}} \gamma_h(x_j, x_k) |Y_k^*| \right\} \right] \leq \frac{1}{M_j \tilde{T}h} \sum_{k=1}^{\tilde{T}} \gamma_h(x_j, x_k) \mathbb{E} \left[\exp \left\{ \lambda M_j |Y_k^*| \right\} \right] .$$

Recall that $Y_k^* = \log(Q_k/m)/\sqrt{2}$ and by assumption $0 \leq \delta \leq f \leq M_0$. Using Lemma 2.3, Q_k can be written as a sum of $m = np/T$ independent gamma random variables, i.e., $Q_k = \sum_{i=1}^m A_i$ where $A_i \sim \Gamma(1/2, 2a_i)$ with $a_i \in (c_1\delta, c_2M_0)$ for some constants c_1, $c_2 > 0$. Thus, we can find some (fully correlated) random variables $Q_k^- \sim \Gamma(m/2, \tilde{c}_1\delta)$ and $Q_k^+ \sim \Gamma(m/2, \tilde{c}_2 M_0)$ such that $Q_k^- \leq Q_k \leq Q_k^+$ almost surely and $\tilde{c}_1, \tilde{c}_2 > 0$ are some constants. Then,

$$\mathbb{E}[\exp\{\lambda M_j |Y_k^*|\}] \leq \mathbb{E}\left[\exp\left\{\lambda M_j/\sqrt{2}\max\left\{|\log(Q_k^+/m)|, |\log(Q_k^-/m)|\right\}\right\}\right].$$

The moment generating function of $|\log(X)|$, when X follows a $\Gamma(a,b)$-distribution, is given by

$$\mathbb{E}\left[\exp\{t\,|\log(X)|\}\right] = b^{-t}\frac{\Gamma(a-t) - \gamma(a-t, 1/b) + b^{2t}\gamma(a+t, 1/b)}{\Gamma(a)}, \quad |t| < a,$$

where $\Gamma(a)$ is the gamma function and $\gamma(a,b)$ is the lower incomplete gamma function. In particular, for $0 < t < a$,

$$\mathbb{E}[\exp\{t|\log(X)|\}] \leq \frac{b^{-t}\Gamma(a-t) + b^t\Gamma(a+t)}{\Gamma(a)}. \tag{2.23}$$

In our setting, $X = Q_k^-/m \sim \Gamma(m/2, \tilde{c}_1\delta/m)$ or $X = Q_k^+/m \sim \Gamma(m/2, \tilde{c}_2 M_0/m)$, respectively. To derive the asymptotic order of $\mathbb{E}[\exp\{\lambda\|\widehat{H(f)}\|_\infty\}]$ for $n, p \to \infty$, where $\lambda > 0$, we first establish the asymptotic order of the ratio $\Gamma(a+t)/\Gamma(a)$ for $a \to \infty$. We distinguish the two cases where t is independent of a and where t linearly depends on a. Based on Stirling's approximation, one derives for $a > 1$, $t > 0$

$$\begin{aligned}\frac{\Gamma(a+t)}{\Gamma(a)} &= \frac{\{2\pi/(a+t)\}^{1/2}(a+t)^{a+t}\exp(-a-t)\left[1+\mathcal{O}\{1/(a+t)\}\right]}{\sqrt{2\pi/a}\,a^a\exp(-a)\left\{1+\mathcal{O}(1/a)\right\}} \\ &= \sqrt{a/(a+t)}\,\{(a+t)/a\}^a\,(a+t)^t\exp(-t)\left[\frac{1+\mathcal{O}\{1/(a+t)\}}{1+\mathcal{O}(1/a)}\right].\end{aligned} \tag{2.24}$$

Thus, for $0 < t < a$ and t independent of a, Equation (2.24) implies for $a \to \infty$ that $\Gamma(a+t)/\Gamma(a) = \mathcal{O}(a^t)$. Similarly, it can be seen that $\Gamma(a-t)/\Gamma(a) = \mathcal{O}(a^{-t})$. If $0 < t < a$ and t linearly depends on a, i.e., $t = ca$ for some constant $c \in (0,1)$, then we get $\Gamma(a \pm t)/\Gamma(a) = \mathcal{O}(a^{\pm t}\exp\{a\})$ for $a \to \infty$. Hence, for a fixed λ not depending on n, p and such that $0 < \lambda < m/(\sqrt{2}M_j)$ we get for sufficiently large n, p

$$\begin{aligned}\mathbb{E}\left[\exp\left\{\lambda M_j|Y_k^*|\right\}\right] &= \mathbb{E}\left[\exp\left\{\lambda M_j/\sqrt{2}\,|\log(Q_k/m)|\right\}\right] \\ &\leq c_3 \sum_{b\in\{\tilde{c}_1\delta/m, \tilde{c}_2 M_0/m\}} (bm/2)^{-\frac{\lambda M_j}{\sqrt{2}}} + (bm/2)^{\frac{\lambda M_j}{\sqrt{2}}} = \mathcal{O}(1).\end{aligned}$$

If $\lambda = cm$ such that $0 < \lambda < m/(\sqrt{2}M_j)$, then for sufficiently large n, p

$$\mathbb{E}\left[\exp\left\{\lambda M_j |Y_k^*|\right\}\right] \le c_4 \exp\{m/2\} \sum_{b \in \{\tilde{c}_1\delta/m, \tilde{c}_2 M_0/m\}} (bm/2)^{-\frac{\lambda M_j}{\sqrt{2}}} + (bm/2)^{\frac{\lambda M_j}{\sqrt{2}}} = \mathcal{O}(L^m),$$

for some constant $L > 1$. Set $K = \min_{j=1,\dots,\tilde{T}} 1/(\sqrt{2}M_j)$ which is a constant independent of n, p. Altogether, we showed that for $0 < \lambda < Km$ and $n, p \to \infty$

$$\mathbb{E}[\exp\{\lambda \|\widehat{H(f)}\|_\infty\}] = \mathcal{O}(T), \qquad \text{if } \lambda \text{ does not depend on } n, p, \tag{2.25}$$

$$\mathbb{E}[\exp\{\lambda \|\widehat{H(f)}\|_\infty\}] = \mathcal{O}(TL^m), \qquad \text{if } \lambda = cm \text{ for } c \in (0, K). \tag{2.26}$$

Bounding the right-hand side of (2.22)

Noting that $\|f\|_\infty \le M_0$ and $m/2 \exp\{\phi(-m/2)\} \in [1, 4]$ for $m \ge 1$, (2.25) implies for some constants $c_0, c_1 > 0$ and $n, p \to \infty$

$$\begin{aligned} \mathbb{E}\|\widehat{f} - f\|_\infty^4 &\le c_0 \mathbb{E}\|\widehat{f}\|_\infty^4 + c_1 = c_0 \mathbb{E}\|(m/2)\exp\{\sqrt{2}\widehat{H(f)} - \phi(m/2)\}\|_\infty^4 + c_1 \\ &\le c_0 \left|(m/2)^4 \exp\{-4\phi(m/2)\}\right| \cdot \mathbb{E}[\exp\{4\sqrt{2}\|\widehat{H(f)}\|_\infty\}] + c_1 = \mathcal{O}(T). \end{aligned}$$

Since g lies between $\widehat{H(f)}$ and $H(f)$, and $H^{-1} \propto \exp$, it follows $H^{-1}(g) \le H^{-1}\{\widehat{H(f)}\} + f$ almost surely pointwise. Thus, for $C > \|f\|_\infty = M_0$, it holds

$$\mathbb{P}(\|H^{-1}(g)\|_\infty > C) \le \mathbb{P}(\|H^{-1}\{\widehat{H(f)}\}\|_\infty > C - M_0) \le \mathbb{P}(\|\widehat{H(f)}\|_\infty > c_1)$$

where $c_1 := H(C - M_0)$. Applying Markov's inequality for $t = cm$ with $c \in (0, K)$ and $C = 2L^{4/c} + M_0$, where c, K, L are the constants in (2.26), gives

$$\mathbb{P}(\|H^{-1}(g)\|_\infty > C) \le \exp(-c_1 t)\mathbb{E}[\exp\{t\|\widehat{H(f)}\|_\infty\}] = \mathcal{O}(TL^{-m}).$$

Thus, the right-hand side of (2.22) is asymptotically negligible to the first term in (2.22). Together with Proposition 2.1, it follows that

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

III. An upper bound on $\mathbb{E}\|\hat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}\|^2$

We use the fact that that the spectral norm of a Toeplitz matrix is upper bounded by the L_∞ norm of its spectral density, see e.g., Gray et al. (2006). Since the derived bounds in this section hold for each $\boldsymbol{\Sigma}(f) \in \mathcal{F}_\beta$, we obtain

$$\sup_{\mathcal{F}_\beta} \mathbb{E}\|\hat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}(f)\|^2 \leq \sup_{\mathcal{F}_\beta} \mathbb{E}\|\hat{f}-f\|_\infty^2 = \mathcal{O}\left\{\log(np)/(nph)\right\} + \mathcal{O}(h^{2\beta}). \tag{2.27}$$

IV. An upper bound on $\mathbb{E}\|\hat{\boldsymbol{\Omega}}-\boldsymbol{\Sigma}^{-1}\|^2$

According to the mean value theorem, for a function g between $\widehat{H(f)}$ and $H(f)$, it holds that

$$\begin{aligned}\mathbb{E}\|\hat{\boldsymbol{\Omega}}-\boldsymbol{\Omega}\|^2 &\leq \mathbb{E}\|1/\hat{f}-1/f\|_\infty^2 = \mathbb{E}\|1/H^{-1}\{\widehat{H(f)}\} - 1/H^{-1}\{H(f)\}\|_\infty^2 \\ &= \mathbb{E}\|(1/H^{-1})'(g)\{\widehat{H(f)}-H(f)\}\|_\infty^2 \lesssim \mathbb{E}\|1/H^{-1}(g)\{\widehat{H(f)}-H(f)\}\|_\infty^2\end{aligned}$$

Application of the Cauchy-Schwarz inequality with a constant $C>0$ yields

$$\mathbb{E}\|1/\hat{f}-1/f\|_\infty^2 \lesssim \mathbb{E}\|\widehat{H(f)}-H(f)\|_\infty^2 + (\mathbb{E}\|1/\hat{f}-1/f\|_\infty^4)^{1/2}\mathbb{P}(\|1/\{H^{-1}(g)\}\|_\infty > C)^{1/2}.$$

Note that $\left\|1/H^{-1}(g)\right\|_\infty \leq 2/m\exp\{\sqrt{2}\|g\|_\infty\}\exp\{\phi(m/2)\} \leq c_1\left\|H^{-1}(g)\right\|_\infty$ for some constant $c_1>0$ not depending on n,p. Choosing the same constant C as in Section 2.8.3, it follows

$$\mathbb{P}\left(\left\|1/\{H^{-1}(g)\}\right\|_\infty > C\right) = \mathcal{O}(TL^{-m}).$$

Noting that $\|1/f\|_\infty \leq 1/\delta$ and $2/m\exp\{\phi(m/2)\} \in [0.25,1]$ for $m \geq 1$, the asymptotic property of the moment generating function of $\widehat{H(f)}$ in (2.25) implies for some constants $c_2, c_3 > 0$ and $n,p\to\infty$

$$\begin{aligned}\mathbb{E}\|1/\hat{f}-1/f\|_\infty^4 &\leq c_2|(2/m)^4\exp\{4\phi(m/2)\}|\cdot\mathbb{E}\|\exp\{-4\sqrt{2}\widehat{H(f)}\}\|_\infty + c_3 \\ &\leq c_2|(2/m)^4\exp\{4\phi(m/2)\}|\cdot\mathbb{E}\exp\{4\sqrt{2}\|\widehat{H(f)}\|_\infty\} + c_3 = \mathcal{O}(T).\end{aligned}$$

Since the derived bounds hold for each $\boldsymbol{\Sigma}(f)\in\mathcal{F}_\beta$, we get all together

$$\sup_{\mathcal{F}_\beta} \mathbb{E}[\|\hat{\boldsymbol{\Omega}}-\boldsymbol{\Sigma}^{-1}(f)\|^2] = \mathcal{O}\left\{\log(np)/(nph))\right\} + \mathcal{O}(h^{2\beta}). \tag{2.28}$$

Optimal bandwith parameter h

Minimizing the right side of (2.27) and (2.28) for the bandwidth parameter h yields $h \asymp \{\log(np)/(np)\}^{\frac{1}{2\beta+1}}$. In particular, $Th \geq p^{\frac{2}{2\beta+1}}\{\log(np)/(np)\}^{\frac{1}{2\beta+1}} \to \infty$ and $h \to 0$ since $v > 2/(2\beta+1)$ for $\beta > 1/2$ and $v \in ((4-2\min\{1,\beta\})/3, 1)$. Thus, substituting $h \asymp \{\log(np)/(np)\}^{\frac{1}{2\beta+1}}$ into (2.27) and (2.28) proves the second statement of Theorem 2.1.

$\square$

2.8.4 Auxiliary lemmata for Theorem 2.1

This section states some technical lemmata needed for the proof of Theorem 2.1. The first lemma lists some properties of the kernel K_h and its extension $\mathcal{K}_h$ on the real line. The proof is based on Schwarz and Krivobokova (2016) and (Schwarz, 2012).

Lemma 2.2 *Let $h > 0$ be the bandwith parameter depending on N.*

(i) There are constants $0 < C < \infty$ and $0 < \gamma < 1$ such that for all for $x, t \in [0,1]$

$$|\mathcal{K}_h(x,t)| < C\gamma^{|x-t|/h},$$
$$|K_h(x,t)| < C[\gamma^{|x-t|/h} + \gamma^{1/h}\{\gamma^{(x-t)/h} + \gamma^{(t-x)/h}\}/(1-\gamma^{1/h})].$$

(ii) Let $t_i = (i-1)/(N-1)$, $i = 1, \ldots, N$. If $Nh \to \infty$ for $N \to \infty$, then $\frac{1}{Nh}\sum_{i=1}^{N}\gamma_h(x,t_i) = \mathcal{O}(1)$ and $\frac{1}{Nh}\sum_{i=1}^{N}\gamma_h^2(x,t_i) = \mathcal{O}(1)$ uniformly for $x \in [0,1]$, where $\gamma_h(x) := \gamma^{|x-t|/h} + \gamma^{1/h}\{\gamma^{(x-t)/h} + \gamma^{(t-x)/h}\}/(1-\gamma^{1/h})$.

(iii) It holds

$$h^{-1}\int_{-\infty}^{\infty}(x-t)^m \mathcal{K}_h(x,t)\,dt = \delta_{m,0}, \qquad \text{for } m = 0, \ldots, 2q-1,$$
$$h^{-1}\int_{-\infty}^{\infty}(x-t)^m \mathcal{K}_h(x,t)\,dt \neq 0, \qquad \text{for } m = 2q.$$

Proof: (i) See Lemma 16 of Schwarz (2012) for the proof of the first statement. Furthermore, for $x, t \in [0,1]$ holds

$$|K_h(x,t)| \leq \sum_{l=-\infty}^{\infty}|\mathcal{K}_h(x,t+l)|$$
$$\leq C\sum_{l=-\infty}^{\infty}\gamma^{|x-t-l|/h} = C\left\{\gamma^{|x-t|/h} + \sum_{l>0}\gamma^{(x-t+l)/h} + \sum_{l>0}\gamma^{(-x+t+l)/h}\right\}$$

$$= C[\gamma^{|x-t|/h} + \gamma^{1/h}\{\gamma^{(x-t)/h} + \gamma^{(t-x)/h}\}/(1-\gamma^{1/h})].$$

(ii) Since $\gamma_h(x,\cdot)$ is differentiable everywhere except at x, it holds

$$\frac{1}{Nh}\sum_{i=1}^{N}\gamma_h(x,t_i) = h^{-1}\int_0^1 \gamma_h(x,t)\,\mathrm{d}t + \mathcal{O}\{(Nh)^{-1}\}$$
$$= h^{-1}\int_0^1 \gamma^{|x-t|/h}\,\mathrm{d}t + h^{-1}\frac{\gamma^{1/h}}{1-\gamma^{1/h}}\int_0^1 \gamma^{(x-t)/h} + \gamma^{(t-x)/h}\,\mathrm{d}t + \mathcal{O}\{(Nh)^{-1}\}.$$

Note that

$$h^{-1}\int_0^1 \gamma^{|x-t|/h}\,\mathrm{d}t = \{\gamma^{(1-x)/h} + \gamma^{x/h} - 2\}/\log(\gamma)$$

and $h^{-1}\int_0^1 \gamma^{(x-t)/h} + \gamma^{(t-x)/h}\,\mathrm{d}x = \gamma^{-t/h}(\gamma^{1/h}-1)/\log(\gamma) - \gamma^{t/h}(\gamma^{-1/h}-1)/\log(\gamma)$.

In particular, for some constants $C_1, C_2 > 0$ depending on $\gamma \in (0,1)$ but not on h and x, it holds $h^{-1}\int_0^1 \gamma^{|x-t|/h}\,\mathrm{d}t \leq C_1$ and $h^{-1}\frac{\gamma^{1/h}}{1-\gamma^{1/h}}\int_0^1 \gamma^{(x-t)/h} + \gamma^{(t-x)/h}\,\mathrm{d}x \leq C_2$. The proof to show $h^{-1}\int_0^1 \gamma_h(x,t)^2\,\mathrm{d}t \leq C_3$ for some constant $C_3 > 0$ is similar. By assumption $(Nh)^{-1} = o(1)$ for $N \to \infty$ which concludes the proof.

(iii) See Lemma 15 of Schwarz (2012) with $p = 2q-1$. □

Lemma 2.3 states that the sum of the correlated gamma random variables in each bin can be rewritten as a sum of independent gamma random variables.

Lemma 2.3 *If $\mathbf{\Sigma} \in \mathcal{F}_\beta$, then $Q_k = \sum_{j=(k-1)p/T+1}^{kp/T}\sum_{i=1}^{n}\tilde{W}_{i,j}$, $k = 1,\ldots,T$, where*

$$\tilde{W}_{i,j} \overset{indep.}{\sim} \Gamma\left(1/2, 2f(x_j) + \mathcal{O}\{T^{-1} + T^{-1}\log(p)p^{1-\beta}\}\right)$$

for $i = 1,\ldots,n$ and $j = (k-1)m+1,\ldots,km$, and $x_j = (j-1)/(2p-2)$.

Proof: It is sufficient to show the statement for $n = 1$ by the independence of the $\mathbf{Y}_i$. Then, the number of points per bin is $m = p/T$. For simplicity, the index i is skipped in the following. First, we write Q_k as a matrix-vector product and refactor it so that it corresponds to a sum of independent scaled chi-squared distributed random variables.

In the second step, we calculate the scaling factors. Let $\mathbf{E}^{(km)}$ be a diagonal matrix with ones on the $(k-1)m+1, ..., km$-th entries and otherwise zero diagonal elements. Then,

$$Q_k = \sum_{j=(k-1)m+1}^{km} W_j = \mathbf{Y}^t \mathbf{D} \mathbf{E}^{(km)} \mathbf{D} \mathbf{Y}.$$

Define $\mathbf{X} = \mathbf{\Sigma}^{-1/2}\mathbf{Y} \sim \mathcal{N}_p(0_p, \mathbf{I})$ and let $\mathbf{U}^t \mathbf{\Gamma} \mathbf{U}$ be the eigendecomposition of $\mathbf{\Sigma}^{1/2}\mathbf{D}\mathbf{E}^{(jm)}\mathbf{D}\mathbf{\Sigma}^{1/2}$ with eigenvalues $(\lambda_j)_{j=1}^p$. Note that $\mathbf{\Sigma}^{1/2}$ is invertible since $\mathbf{\Sigma}$ invertible. Then,

$$Q_k = \mathbf{X}^t \mathbf{\Sigma}^{1/2} \mathbf{D} \mathbf{E}^{(km)} \mathbf{D} \mathbf{\Sigma}^{1/2} \mathbf{X} = \mathbf{X}^t \mathbf{U}^t \mathbf{\Gamma} \mathbf{U} \mathbf{X} = \tilde{\mathbf{X}}^t \mathbf{\Gamma} \tilde{\mathbf{X}} = \sum_{j=1}^p \lambda_j \tilde{\mathbf{X}}_j^2,$$

where $\tilde{\mathbf{X}} := \mathbf{U}\mathbf{X}$. Since $\mathbf{U}$ is an orthogonal matrix, it follows that $\tilde{\mathbf{X}} \sim \mathcal{N}_p(0_p, \mathbf{I})$. In particular, Q_k is a sum of independent scaled chi-squared distributed random variables $\tilde{W}_j := \lambda_j \tilde{X}_j^2 \overset{indep.}{\sim} \Gamma(1/2, 2\lambda_j)$ where the scaling factors are the eigenvalues λ_j. It remains to calculate the λ_j. Note that the matrix $\mathbf{\Sigma}^{1/2}\mathbf{D}\mathbf{E}^{(km)}\mathbf{D}\mathbf{\Sigma}^{1/2}$ is similar to the matrix

$$\mathbf{A} := \mathbf{D}\mathbf{\Sigma}^{-1/2} \left(\mathbf{\Sigma}^{1/2}\mathbf{D}\mathbf{E}^{(km)}\mathbf{D}\mathbf{\Sigma}^{1/2} \right) \mathbf{\Sigma}^{1/2}\mathbf{D} = \mathbf{E}^{(km)}\mathbf{D}\mathbf{\Sigma}\mathbf{D}.$$

In particular, $\mathbf{\Sigma}^{1/2}\mathbf{D}\mathbf{E}^{(km)}\mathbf{D}\mathbf{\Sigma}^{1/2}$ has the same eigenvalues as $\mathbf{A}$. The rows of $\mathbf{A}$ are zero except for the $(k-1)m+1, ..., km$ -th rows, which equal the $(k-1)m+1, ..., km$ -th rows of $\mathbf{D}\mathbf{\Sigma}\mathbf{D}$. By Lemma 2.1,

$$(\mathbf{D}\mathbf{\Sigma}\mathbf{D})_{i,j} = f(x_i)\delta_{i,j} + \frac{1+(-1)^{|i-j|}}{2} \mathcal{O}\{p^{-1} + \log(p)p^{-\beta}\}.$$

Define $\mathbf{B} := (\mathbf{E}^{(km)}\mathbf{D}\mathbf{\Sigma}\mathbf{D}\mathbf{E}^{(km)})_{r,s=(k-1)m+1}^{km}$. Then, $\lambda(\mathbf{A}) = \lambda(\mathbf{B}) \cup \{0\}$ and $\mathbf{A}$ has an eigenvalue zero with multiplicity at least $p-m$. Since $\mathbf{B}$ is symmetric all eigenvalues are real-valued. Application of the Gershgorin circle theorem to $\mathbf{B}$ yields for the remaining m eigenvalues $\lambda_{(k-1)m+1}, ..., \lambda_{km}$ are contained in the following set

$$\bigcup_{j=(k-1)m+1}^{km} \left\{ z \in \mathbb{R} : \ |f(x_j) + \mathcal{O}\{p^{-1} + \log(p)p^{-\beta}\} - z| \quad < \mathcal{O}\{mp^{-1} + m\log(p)p^{-\beta}\} \right\}.$$

Since f is $\min\{\beta, 1\}$-Hölder continuous, we have $\lambda_j = f(x_j) + \mathcal{O}\{mp^{-1} + m\log(p)p^{-\beta}\}$, for $j = (k-1)m+1, ..., km$. Noting that $T^{-1} = mp^{-1}$ the statement follows. □

Finally, Lemma 2.4 quantifies the difference to an exact Gaussian regression setting by giving explicit bounds for the stochastic and deterministic errors of the variance stabilizing transform. This result is a generalization of Theorem 1 of Cai and Zhou (2010) adapted to our setting with $n \geq 1$ observations and correlated observations.

Lemma 2.4 *If $\mathbf{\Sigma} \in \mathcal{F}_\beta$, then $Y_k^* = \log(Q_k/m)/\sqrt{2}$ can be written as*

$$Y_k^* = H\left\{f\left(x_k\right)\right\} + \epsilon_k + m^{-1/2} Z_k + \xi_k, \qquad x_k = (k-1)/(2T-2),\, k = 1, \ldots, T.$$

where $|\epsilon_k| \lesssim (np)^{-1} + \log(p)(np)^{-\beta}$, $Z_k \sim \mathcal{N}(0,1)$, and $\mathbb{E}[\xi_k] = 0$. Furthermore,

$$\begin{aligned} &\mathbb{E}|\xi_k|^\ell \lesssim \log^{2\ell}(m)[m^{-\ell} + \left\{T^{-1} + T^{-1}\log(p)p^{1-\beta}\right\}^\ell], && \ell \in \mathbb{N}_{>1}, \\ &Cov(Z_k, Z_l) = \mathcal{O}\{mp^{-2} + m\log(p)^2 p^{-2\beta}\}, && k \neq l = 1, \ldots, T. \end{aligned}$$

Proof: In the regression setting of Cai and Zhou (2010) and Brown et al. (2010) the gamma random variables in each bin are independent. Using Lemma 2.3, we can rewrite our setting into one with independent observations per bin. Thus, the first step of the proof is to apply the results of Cai and Zhou (2010) and Brown et al. (2010) on the approximation of Y_k^* by a Gaussian random variable and a non-Gaussian remainder ξ_k. Since in our setting the observations of different bins $i \neq j$ are correlated, the covariance between the Gaussian random variables is derived in the second part of the proof.

By Lemma 2.3, Q_k can be written as sum of $m = np/T$ independent gamma random variables with mean function $\tilde{f}(x) := f(x) + S(x)$, where $S(x) = \mathcal{O}\{T^{-1} + T^{-1}\log(p)p^{1-\beta}\}$. It follows from Lemma 2.1 and Lemma 2.3 that the S term depends continuously on $x \in [0,1]$ and that a global constant hidden in the $\mathcal{O}$ term exists, which is independent of the x. Furthermore, $f \in [\delta, M_0] \subset \mathbb{R}_{>0}$ by assumption. As $\tilde{f}$ is the mean of gamma random variables, it holds that $\tilde{f} > 0$. Thus, $\tilde{f}$ is continuous and takes values in a compact set $[\epsilon, v] \subset \mathbb{R}_{>0}$ of the support of the gamma distribution. By Theorem 1 of Cai and Zhou (2010) for the gamma distribution, it follows

$$Y_k^* = H\left\{\tilde{f}\left(x_k^*\right)\right\} + m^{-1/2} Z_k + \xi_k, \quad k = 1, \ldots, T,$$

where $(k-1)/(2T-2) \leq x_k^* \leq k/(2T-2)$, $Z_k \sim \mathcal{N}(0,1)$, and ξ_k is a mean zero random variable. Z_k is defined via quantile coupling. It is the Gaussian approximation to the standardized version of Q_k where the W_{ij} in the k-th bin are assumed to be i.i.d. In particular, Z_k approximates

$$\bar{Q}_k = \frac{\sum_{j=(k-1)p/T+1}^{kp/T} \sum_{i=1}^{n} \bar{W}_{i,j} - m\tilde{f}\left(x_k^*\right)}{\sqrt{2m}\tilde{f}\left(x_k^*\right)}$$

where $\bar{W}_{i,j} \overset{\text{i.i.d.}}{\sim} \Gamma(1/2, 2\tilde{f}\left(x_k^*\right))$ and such that $\mathrm{Cov}(\bar{W}_{i,j}, \bar{W}_{i,h}) = \mathrm{Cov}(W_{i,j}, W_{i,h})$ for $j = (k-1)p/T+1, ..., kp/T$ and $h \in \{1, ..., p\} \setminus \{(k-1)p/T+1, ..., kp/T\}$.

Let θ be the maximum difference of the observations' means in each bin. Then,

$$\theta = \max_{k=1,...,T} \max_{j,l=(k-1)p/T+1,...,kp/T} \left|\mathbb{E}[\tilde{W}_{1,j}] - \mathbb{E}[\tilde{W}_{1,l}]\right| \leq c_1 \left\{T^{-1} + T^{-1}\log(p)p^{1-\beta}\right\}.$$

Application of Lemma 4 and Theorem 1 of Brown et al. (2010) yield that ξ_k satisfies

$$\mathbb{E}[|\xi_k|^\ell] \lesssim \log^{2\ell}(m) \left[m^{-\ell} + \left\{T^{-1} + T^{-1}\log(p)p^{1-\beta}\right\}^\ell\right]$$

for any constant integer $\ell > 0$ and where the hidden constant depends on ℓ.

Next,

$$Y_k^* = H\left\{f\left(\frac{k-1}{2T-2}\right)\right\} + \epsilon_k + m^{-1/2}Z_k + \xi_k, \quad k = 1, ..., T,$$

where $\epsilon_k = H[f\{(k-1)/(2T-2)\}] - H\{\tilde{f}\ (x_k^*)\}$ and $|\epsilon_k| \lesssim (np)^{-1} + \log(p)(np)^{-\beta}$ by the Lipschitz continuity of the logarithm and the α-Hölder continuity of f.

To compute $\mathrm{Cov}(Z_k, Z_l)$ for $k \neq l$ we use Hoeffding's covariance identity. In particular, we show that $\mathrm{Cov}(Z_k, Z_l) \lesssim Cov(\bar{Q}_k, \bar{Q}_l)$. The claim then follows from noting that

$$0 \leq \mathrm{Cov}(\bar{Q}_k, \bar{Q}_l) = \frac{m\mathcal{O}\{\mathrm{Cov}(W_{i,j}, W_{i,h})\}}{2\tilde{f}\left(x_k^*\right)\tilde{f}\left(x_l^*\right)} = \mathcal{O}\left\{mp^{-2} + m\log(p)p^{-2\beta}\right\},$$

where $j = (k-1)p/T+1, ..., kp/T$ and $h = (l-1)p/T+1, ..., lp/T$.

Let Φ and $F_{\bar{Q}}$ denote the cumulative distribution functions of Z_k, Z_l, $\bar{Q}_k$ and $\bar{Q}_l$. Since the Z_k are defined via quantile coupling, it holds $Z_k = \Phi^{-1}\{F_{\bar{Q}}(\bar{Q}_k)\}$ (see Komlós et al., 1975; Mason and Zhou, 2012). Furthermore, define the uniform random variables $U_k = \Phi(Z_k) = F_{\bar{Q}}(\bar{Q}_k)$ and $U_l = \Phi(Z_l) = F_{\bar{Q}}(\bar{Q}_l)$ with cumulative distribution function F_U. The corresponding joint cumulative distribution functions are denoted by $F_{Z,Z}$, $F_{\bar{Q},\bar{Q}}$, $F_{U,U}$. By Hoeffding's covariance identity

$$\operatorname{Cov}(\bar{Q}_k, \bar{Q}_l) = \int_{\mathbb{R}} \int_{\mathbb{R}} F_{\bar{Q},\bar{Q}}(q,r) - F_{\bar{Q}}(q) F_{\bar{Q}}(r) \, \mathrm{d}q \, \mathrm{d}r \tag{2.29}$$

$$= \int_{\mathbb{R}} \int_{\mathbb{R}} \{F_{Z,Z}(x,y) - \Phi(x)\Phi(y)\} \frac{\phi(x)}{f_{\bar{Q}}\{F_{\bar{Q}}^{-1}(x)\}} \frac{\phi(y)}{f_{\bar{Q}}\{F_{\bar{Q}}^{-1}(y)\}} \, \mathrm{d}x \, \mathrm{d}y. \tag{2.30}$$

Let $\rho = \operatorname{Cov}(Z_k, Z_l)$. Then, the identity

$$F_{ZZ}(x,y) = \frac{1}{2\pi} \int_0^{\rho} \frac{1}{\sqrt{(1-r^2)}} \exp\left\{ -\frac{x^2 - 2rxy + y^2}{2(1-r^2)} \right\} \mathrm{d}r + \Phi(x)\Phi(y)$$

implies $F_{Z,Z}(x,y) - \Phi(x)\Phi(y) \geq 0$ for all $x, y \in \mathbb{R} \iff \rho \geq 0$, (see Patel and Read, 1996, Eq. 9.3.1.3). Furthermore, the ratios of the two density function in (2.29) are non-negative. Together with the fact that $\operatorname{Cov}(\bar{Q}_k, \bar{Q}_l) \geq 0$, it hence follows $\operatorname{Cov}(Z_k, Z_l) \geq 0$.

Furthermore,

$$\begin{aligned}
\operatorname{Cov}(Z_l, Z_k) &= \int_{\mathbb{R}} \int_{\mathbb{R}} F_{Z,Z}(x,y) - \Phi(x)\Phi(y) \, \mathrm{d}x \, \mathrm{d}y \\
&= \int_0^1 \int_0^1 \{F_{UU}(u,v) - F_U(u)F_U(v)\} \frac{1}{\phi\{\Phi^{-1}(u)\}} \frac{1}{\phi\{\Phi^{-1}(v)\}} \, \mathrm{d}u \, \mathrm{d}v \\
&\leq \int_{1/2}^1 \int_{1/2}^1 \{F_{UU}(u,v) - F_U(u)F_U(v)\} \frac{1}{\phi\{\Phi^{-1}(u)\}} \frac{1}{\phi\{\Phi^{-1}(v)\}} \, \mathrm{d}u \, \mathrm{d}v \\
&+ \int_0^1 \int_0^{1/2} \{F_{UU}(u,v) - F_U(u)F_U(v)\} \frac{2}{\phi\{\Phi^{-1}(u)\}} \frac{1}{\phi\{\Phi^{-1}(v)\}} \, \mathrm{d}u \, \mathrm{d}v.
\end{aligned}$$

The last integral corresponds to $\mathbb{E}[Z_k Z_l^-] = \operatorname{Cov}(Z_k, Z_l)/2$ where $Z_l^- = Z_l \mathbb{1}\{Z_l \leq 0\}$. It remains to show that it exists a $c > 0$ such that $\frac{f_{\bar{Q}}\{F_{\bar{Q}}^{-1}(u)\}}{\phi\{\Phi^{-1}(u)\}} \leq c$ for all $u \in (1/2, 1)$. Then, it follows

$$\begin{aligned}
\operatorname{Cov}(Z_l, Z_k) &\leq 2c^2 \int_0^1 \int_0^1 \{F_{UU}(u,v) - F_U(u)F_U(v)\} \frac{1}{f_{\bar{Q}}\{F_{\bar{Q}}^{-1}(u)\}} \frac{1}{f_{\bar{Q}}\{F_{\bar{Q}}^{-1}(v)\}} \, \mathrm{d}u \, \mathrm{d}v \\
&= 2c^2 \operatorname{Cov}(\bar{Q}_k, \bar{Q}_l).
\end{aligned}$$

Note that for $u \in (1/2, 1)$

$$\frac{\partial}{\partial u} \frac{f_{\bar{Q}}\{F_{\bar{Q}}^{-1}(u)\}}{\phi\{\Phi^{-1}(u)\}} = \frac{f'_{\bar{Q}}\{F_{\bar{Q}}^{-1}(u)\}}{f_{\bar{Q}}\{F_{\bar{Q}}^{-1}(u)\}\phi\{\Phi^{-1}(u)\}} - \frac{\Phi^{-1}(u) f_{\bar{Q}}\{F_{\bar{Q}}^{-1}(u)\}}{\phi\{\Phi^{-1}(u)\}^2}.$$

Since $\Phi^{-1}(u)>0$ for $u \in (1/2,1)$, the second term is positive. To see that the first term is negative for $u \in (1/2,1)$ we rewrite $f_{\bar{Q}}$ and $F_{\bar{Q}}^{-1}$ with respect to the non-normalized gamma random variable $X = \sum_{j=(k-1)p/T+1}^{kp/T} \sum_{i=1}^{n} \bar{W}_{i,j} \sim \Gamma(m/2, 2\tilde{f}(x_k^*))$. Hence,

$$f_{\bar{Q}}(x) = \sigma f_X(x\sigma + \mu), \qquad F_{\bar{Q}}^{-1}(u) = \frac{F_X^{-1}(u) - \mu}{\sigma}$$

where σ and μ are the standard deviation and the mean of X. Note that the mode of $A \sim \Gamma(a,b)$ is at $b(a-1)$. Since $x\sigma + \mu = 2\tilde{f}(x_k)(\frac{m}{2} - 1)$ implies $x = -\sqrt{2/m}$, it follows that $f_{\bar{Q}}(x)$ is monotone decreasing for $x \geq -\sqrt{2/m}$. Furthermore, $F_{\bar{Q}}(-\sqrt{m/2}) \leq 0.5$ for all $m \in \mathbb{N}$ as $f_{\bar{Q}}(x)$ is right-skewed. In particular, $-\sqrt{m/2} \leq F_{\bar{Q}}^{-1}(1/2)$ for all $m \in \mathbb{N}$. Finally, since $f_{\bar{Q}}(-\sqrt{2/m}) \to \phi(0)$ for $m \to \infty$ there is a constant $c > 0$ not depending on m such that

$$\frac{f_{\bar{Q}}\{F_{\bar{Q}}^{-1}(u)\}}{\phi\{\Phi^{-1}(u)\}} \leq \frac{f_{\bar{Q}}(-\sqrt{2/m})}{\phi(0)} \leq c. \qquad \square$$

2.8.5 Simulation study with non-Gaussian data

The simulation study in Section 2.5 is performed in the same way, but with the uniform and the gamma distribution instead of the Gaussian distribution.

Uniform distribution

The observations follow a uniform distribution with covariance matrices $\mathbf{\Sigma}_1$, $\mathbf{\Sigma}_2$, $\mathbf{\Sigma}_3$ of examples **1, 2, 3**, i.e., $\mathbf{Y}_i = \mathbf{\Sigma}_j^{1/2} X_i$, $j = 1,3$, with $X_1, ..., X_n \overset{\text{i.i.d.}}{\sim} \text{Unif}[-\sqrt{3}, \sqrt{3}]$. For example **2**, the parameter *innov* of the R function *arima.sim* is used to pass the innovations $X_1, ..., X_n \overset{\text{i.i.d.}}{\sim} \text{Unif}[-\sqrt{3}, \sqrt{3}]$. The resulting process is multiplied by $\sqrt{1.2}$. Table 4, 5 and 6 show respectively the results for **(A)** $p = 5000$, $n = 1$, **(B)** $p = 1000$, $n = 50$ and **(C)** $p = 5000$, $n = 10$.

Gamma distribution

The observations follow a centered gamma distribution with covariance matrices $\mathbf{\Sigma}_1$, $\mathbf{\Sigma}_2$, $\mathbf{\Sigma}_3$ of examples **1, 2, 3**, i.e., $\mathbf{Y}_i = \mathbf{\Sigma}_j^{1/2} X_i$, $j = 1,3$, with $X_1 + \sqrt{2}, ..., X_n + \sqrt{2} \overset{\text{i.i.d.}}{\sim} \Gamma(2, 1/\sqrt{2})$. For example **2**, the parameter *innov* of the R function *arima.sim* is used to pass the innovations $X_1 + \sqrt{2}, ..., X_n + \sqrt{2} \overset{\text{i.i.d.}}{\sim} \Gamma(2, 1/\sqrt{2})$. The resulting process is multiplied by $\sqrt{1.2}$. Table 7, 8 and 9 show respectively the results for **(A)** $p = 5000$, $n = 1$, **(B)** $p = 1000$, $n = 50$ and **(C)** $p = 5000$, $n = 10$.

	(1) polynomial σ		(2) ARMA$(2,2)$		(3) Lipschitz cont. f		time
	$\|\widehat{\Sigma}-\Sigma\|^2$	$\|\hat{f}-f\|_2^2$	$\|\widehat{\Sigma}-\Sigma\|^2$	$\|\hat{f}-f\|_2^2$	$\|\widehat{\Sigma}-\Sigma\|^2$	$\|\hat{f}-f\|_2^2$	in sec
VST-DCT (GCV)	0.0074	0.0023	0.1886	0.0330	0.0401	0.0062	3.9605
VST-DCT (ML)	0.0057	0.0019	0.1844	0.0326	0.0399	0.0061	3.9633
tapering (semi-oracle)	0.0049	0.0016	0.1726	0.0370	0.0284	0.0077	4.7148
sample covariance	171.3200	35.9913	566.5580	59.6976	269.6001	37.7044	0.3431
VST-DCT (GCV-oracle)	0.0039	0.0014	0.1711	0.0313	0.0328	0.0054	–
VST-DCT (ML-oracle)	0.0042	0.0015	0.1818	0.0322	0.0390	0.0060	–
tapering (oracle)	0.0037	0.0013	0.1583	0.0311	0.0146	0.0034	–

Table 4: **Scenario (A)** with **p = 5000, n = 1**, Uniform distribution: Errors of the Toeplitz covariance matrix and the spectral density estimators with respect to the spectral norm and the L_2 norm, respectively. Average computation time of the covariance estimators in seconds for one Monte Carlo sample (last column).

	(1) polynomial σ		(2) ARMA$(2,2)$		(3) Lipschitz cont. f		time
	$\|\widehat{\Sigma}-\Sigma\|^2$	$\|\hat{f}-f\|_2^2$	$\|\widehat{\Sigma}-\Sigma\|^2$	$\|\hat{f}-f\|_2^2$	$\|\widehat{\Sigma}-\Sigma\|^2$	$\|\hat{f}-f\|_2^2$	in sec
VST-DCT (GCV)	0.0009	0.0002	0.0888	0.0220	0.0068	0.0009	23.3553
VST-DCT (ML)	0.0008	0.0002	0.0959	0.0226	0.0077	0.0010	23.2826
tapering (CV)	0.0030	0.0012	0.1400	0.0364	0.0081	0.0017	24.4997
sample covariance	0.8885	0.5917	4.0965	0.9704	1.3657	0.6137	0.1525
VST-DCT (GCV-oracle)	0.0006	0.0002	0.0871	0.0219	0.0057	0.0008	–
VST-DCT (ML-oracle)	0.0007	0.0002	0.0957	0.0226	0.0076	0.0009	–
tapering (oracle)	0.0018	0.0011	0.1306	0.0360	0.0062	0.0015	–

Table 5: **Scenario (B)** with **p = 1000, n = 50**, Uniform distribution: Errors of the Toeplitz covariance matrix and the spectral density estimators with respect to the spectral norm and the L_2 norm, respectively. Average computation time of the covariance estimators in seconds for one Monte Carlo sample (last column).

	(1) polynomial σ		(2) ARMA$(2,2)$		(3) Lipschitz cont. f		time
	$\lVert\widehat{\Sigma}-\Sigma\rVert^2$	$\lVert\hat{f}-f\rVert_2^2$	$\lVert\widehat{\Sigma}-\Sigma\rVert^2$	$\lVert\hat{f}-f\rVert_2^2$	$\lVert\widehat{\Sigma}-\Sigma\rVert^2$	$\lVert\hat{f}-f\rVert_2^2$	in sec
VST-DCT (GCV)	0.0009	0.0002	0.0931	0.0220	0.0057	0.0008	4.2534
VST-DCT (ML)	0.0007	0.0002	0.0998	0.0227	0.0064	0.0009	4.2722
tapering (CV)	0.0332	0.0259	0.4620	0.1300	0.0743	0.0285	649.1202
sample covariance	9.1367	4.5427	35.9029	7.2986	12958.1671	4.7971	1.1976
VST-DCT (GCV-oracle)	0.0006	0.0002	0.0901	0.0219	0.0048	0.0007	–
VST-DCT (ML-oracle)	0.0007	0.0002	0.0993	0.0227	0.0063	0.0008	–
tapering (oracle)	0.0301	0.0258	0.4254	0.1292	0.0681	0.0278	–

Table 6: **Scenario (C)** with $\mathbf{p = 5000, n = 10}$, Uniform distribution: Errors of the Toeplitz covariance matrix and the spectral density estimators with respect to the spectral norm and the L_2 norm, respectively. Average computation time of the covariance estimators in seconds for one Monte Carlo sample (last column).

	(1) polynomial σ		(2) ARMA$(2,2)$		(3) Lipschitz cont. f		time
	$\lVert\widehat{\Sigma}-\Sigma\rVert^2$	$\lVert\hat{f}-f\rVert_2^2$	$\lVert\widehat{\Sigma}-\Sigma\rVert^2$	$\lVert\hat{f}-f\rVert_2^2$	$\lVert\widehat{\Sigma}-\Sigma\rVert^2$	$\lVert\hat{f}-f\rVert_2^2$	in sec
VST-DCT (GCV)	0.0118	0.0043	0.2035	0.0323	0.0518	0.0088	3.8888
VST-DCT (ML)	0.0097	0.0039	0.1625	0.0306	0.0479	0.0082	3.8831
tapering (semi-oracle)	0.0095	0.0037	0.1744	0.0374	0.0416	0.0115	4.6909
sample covariance	167.5072	33.8064	545.1911	62.8606	248.2843	35.9759	0.3418
VST-DCT (GCV-oracle)	0.0079	0.0035	0.1511	0.0293	0.0417	0.0076	–
VST-DCT (ML-oracle)	0.0082	0.0036	0.1610	0.0302	0.0479	0.0081	–
tapering (oracle)	0.0076	0.0033	0.1505	0.0312	0.0199	0.0054	–

Table 7: **Scenario (A)** with $\mathbf{p = 5000, n = 1}$, Gamma distribution: Errors of the Toeplitz covariance matrix and the spectral density estimators with respect to the spectral norm and the L_2 norm, respectively. Average computation time of the covariance estimators in seconds for one Monte Carlo sample (last column).

	(1) polynomial σ		(2) ARMA$(2,2)$		(3) Lipschitz cont. f		time
	$\Vert\widehat{\Sigma}-\Sigma\Vert^2$	$\Vert\hat{f}-f\Vert_2^2$	$\Vert\widehat{\Sigma}-\Sigma\Vert^2$	$\Vert\hat{f}-f\Vert_2^2$	$\Vert\widehat{\Sigma}-\Sigma\Vert^2$	$\Vert\hat{f}-f\Vert_2^2$	in sec
VST-DCT (GCV)	0.0013	0.0004	0.0862	0.0218	0.0059	0.0010	23.3559
VST-DCT (ML)	0.0011	0.0004	0.0928	0.0225	0.0074	0.0011	23.2890
tapering (CV)	0.0029	0.0014	0.1338	0.0359	0.0085	0.0020	24.1672
sample covariance	0.8944	0.6033	4.0165	1.0358	1.4070	0.6303	0.1460
VST-DCT (GCV-oracle)	0.0009	0.0003	0.0818	0.0216	0.0055	0.0009	–
VST-DCT (ML-oracle)	0.0009	0.0004	0.0921	0.0224	0.0073	0.0011	–
tapering (oracle)	0.0022	0.0013	0.1224	0.0355	0.0072	0.0018	–

Table 8: **Scenario (B)** with $\mathbf{p = 1000, n = 50}$, Gamma distribution:: Errors of the Toeplitz covariance matrix and the spectral density estimators with respect to the spectral norm and the L_2 norm, respectively. Average computation time of the covariance estimators in seconds for one Monte Carlo sample (last column).

	(1) polynomial σ		(2) ARMA$(2,2)$		(3) Lipschitz cont. f		time
	$\Vert\widehat{\Sigma}-\Sigma\Vert^2$	$\Vert\hat{f}-f\Vert_2^2$	$\Vert\widehat{\Sigma}-\Sigma\Vert^2$	$\Vert\hat{f}-f\Vert_2^2$	$\Vert\widehat{\Sigma}-\Sigma\Vert^2$	$\Vert\hat{f}-f\Vert_2^2$	in sec
VST-DCT (GCV)	0.0014	0.0004	0.0910	0.0220	0.0063	0.0010	4.4351
VST-DCT (ML)	0.0011	0.0004	0.0963	0.0226	0.0070	0.0011	4.4039
tapering (CV)	0.0337	0.0259	0.4376	0.1287	0.0777	0.0284	681.1327
sample covariance	8.9878	4.3716	36.2328	7.2028	13099.5464	4.7331	1.2680
VST-DCT (GCV-oracle)	0.0009	0.0003	0.0865	0.0218	0.0056	0.0009	–
VST-DCT (ML-oracle)	0.0010	0.0003	0.0958	0.0226	0.0070	0.0010	–
tapering (oracle)	0.0306	0.0258	0.4085	0.1281	0.0675	0.0276	–

Table 9: **Scenario (C)** with $\mathbf{p = 5000, n = 10}$, Gamma distribution: Errors of the Toeplitz covariance matrix and the spectral density estimators with respect to the spectral norm and the L_2 norm, respectively. Average computation time of the covariance estimators in seconds for one Monte Carlo sample (last column).

3 On second-order statistics of the log-average periodogram

This chapter is based on:
Klockmann, K. and Krivobokova, T. (2023b). On second-order statistics of the log-average periodogram

Abstract

We present an approximate expression for the covariance of the log-average periodogram for a zero mean stationary Gaussian process. Our findings extend the work of Ephraim and Roberts (2005) on the covariance of the log-periodogram by additionally taking averaging over local frequencies into account. Moreover, we provide a simple expression for the non-integer moments of a non-central chi-squared distribution.

Keywords: Log-periodogram, smoothing, spectral density

3.1 Introduction

This paper investigates the covariance of the log-average periodogram. The estimator is obtained from a real-valued time series $y_1, y_2 \dots, y_p$, which is a realization of a stationary Gaussian process with zero mean. The spectral density function of the process is assumed to be Hölder continuous. The log-average periodogram consists of averaging local frequencies of the periodogram before applying the logarithm. We use the discrete cosine transform to define the periodogram. Since the time series is real-valued, this provides estimates of the spectral density function on a grid twice as fine as the periodogram defined by the discrete Fourier transform (Klockmann and Krivobokova, 2023a, Lem. 1). Let

$$Y_i = \frac{1}{\sqrt{p-1}}\left[\frac{y_1 + (-1)^i y_p}{\sqrt{2}} + \sum_{j=2}^{p-1} y_j \cos\left\{\pi(j-1)\frac{i-1}{p-1}\right\}\right] \tag{3.1}$$

denote the discrete cosine I transform of the time series at frequency $0 \leq \pi(i-1)/(p-1) \leq \pi$ for $i = 2, ..., p-1$. For $i = 1, p$, the right-hand side of (3.1) is additionally multiplied with a factor $1/\sqrt{2}$. The Y_i are also called spectral components. The periodogram and log-periodogram of the time series are given by Y_i^2 and $\log\{Y_i^2\}$, $i = 1, ..., p$, where log denotes the natural logarithm. We define the log-average periodogram by

$$Y_j^* = \log\left(\frac{1}{m}\sum_{i=(j-1)m+1}^{jm} Y_i^2\right)$$

where $j = 1, \ldots, \lfloor p/m \rfloor$ and $1 \leq m \leq p$ is the number of averaged frequencies. The number of bins $\lfloor p/m \rfloor$ will be denoted by T in the following. If p/m is not an integer, the T-th bin may contain more than m frequencies. For finite p, the Y_i, $i = 1, ..., p$, are in general correlated. We are interested in an expression for the covariance of the log-average periodogram

$$\operatorname{Cov}\left(Y_j^*, Y_{j'}^*\right), \quad j, j' = 1, ..., T.$$

The log-average periodogram is a crucial quantity in various applications such as estimating the long memory parameter (Robinson, 1995), cepstral coefficients (Gunter, 2001; Sandberg and Hansson-Sandsten, 2012), and the spectral density function (Klockmann and Krivobokova, 2023a). Averaging adjacent frequencies of the periodogram has been a well-known technique to reduce variance, with various kernel functions proposed for this purpose such as Bartletts's window, Welch's window, the triangular kernel, see e.g., Welch (1967) and Bartlett (1950). Numerical studies demonstrate that the log-average periodogram has superior performance in estimating the Hurst parameter compared to the non-averaged log-periodogram (Karagiannis et al., 2006).

Estimation and inference for the periodogram and estimators based on it are usually performed under the assumption of independence of the spectral components. However, the spectral components are independent only asymptotically (for $p \to \infty$) and for time series with a fast enough decaying autocovariance function. Taking into account this dependence explicitly leads to improved finite sample performance and more accurate inference for periodogram-based estimators. For example, Ephraim and Roberts (2005) derived the covariance of the log-periodogram and demonstrated enhanced performance in cepstral coefficient estimation when accounting for the covariance structure. However, no formulas for the covariance matrix of the log-periodogram with averaged frequencies are available, to the best of our knowledge.

In this paper, the covariance structure of the log-average periodogram for a stationary Gaussian process with a smooth spectral density function is derived. The obtained covariance expression extends the work of Ephraim and Roberts (2005) by accounting for the averaging of local frequencies and complements the work of Gunter (2001), which computes only the bias and variance of the log-average periodogram.

3.2 Main result

We consider a real-valued stationary Gaussian process with zero mean, covariance matrix $\mathbf{\Sigma} = (\sigma_{|i-j|})_{i,j=1}^{p} \in \mathbb{R}^{p\times p}$ and corresponding spectral density function $f : [-\pi, \pi] \to [0, \infty)$. The spectral components Y_i, $i = 1, ..., p$, of the time series are defined as in (3.1). In particular, the vector $(Y_1, ..., Y_p)$ is Gaussian distributed with zero mean and covariance matrix $\mathbf{\Omega} = (\omega_{st})_{s,t=1}^{p} = \mathbf{D\Sigma D}$ where $\mathbf{D}$ is the Discrete Cosine Transform I matrix defined by

$$\mathbf{D} = \sqrt{2}(p-1)^{-1/2}\left[\cos\left\{\pi(i-1)\frac{j-1}{p-1}\right\}\right]_{i,j=1}^{p} \quad \text{divided by } \sqrt{2} \text{ when } i,j \in \{1,p\}.$$

Let $\mathbf{\Omega}$ be partitioned as follows in $(m \times m)$-dimensional submatrices

$$\mathbf{\Omega} = \begin{bmatrix} \mathbf{\Omega}_{1,1} & \cdots & \mathbf{\Omega}_{1,T} \\ \cdots & \cdots & \cdots \\ \mathbf{\Omega}_{T,1} & \cdots & \mathbf{\Omega}_{T,T} \end{bmatrix}.$$

Note that the matrices $\mathbf{\Omega}_{T,i}$ can have a different dimension if p/m is not an integer. Our result is expressed in terms of the m-dimensional correlation matrices, which for $j, j' = 1, ..., T$, are defined as

$$\bar{\mathbf{\Omega}}_{j,j'} = \mathbf{\Omega}_{j,j}^{-1}\mathbf{\Omega}_{j',j}\mathbf{\Omega}_{j,j'}\mathbf{\Omega}_{j',j'}^{-1}.$$

Our main result is derived under the following assumptions:

$$f \text{ is at least } \alpha - \text{Hölder continuous for some } \alpha \in (0,1], \tag{A1}$$

$$0 < \epsilon \le f \le M < \infty \text{ for some } 0 < \epsilon < M. \tag{A2}$$

Assumption (A1) implies that $\omega_{st} = f\{\pi(s-1)/(p-1)\} + \mathcal{O}\{\log(p)p^{-\alpha}\}$ for $s = t$ and $\omega_{st} = \mathcal{O}\{\log(p)p^{-\alpha}\}$ for $s \neq t = 1, ..., p$, where the $\mathcal{O}$ term is uniform over s, t and for $p \to \infty$, see Lemma 1 of Klockmann and Krivobokova (2023a). Assumption (A2) implies that the covariance matrix $\boldsymbol{\Sigma}$ is invertible. Let $(a)_n := \prod_{i=1}^{n}(a+i-1)$ denote the Pochhammer symbol and let

$$ {}_3F_2(a,b,c,d,e;z) = \sum_{n=0}^{\infty} \frac{(a)_n(b)_n(c)_n}{(d)_n(e)_n} \frac{z^n}{n!} $$

denote the generalized hypergeometric series with $a, b, c, d, e, f \in \mathbb{R}$ such that $d, e \neq 0, -1, -2, ...$ The series converges absolutely for $|z| < 1$ and for $|z| = 1$ when $d + e - a - b - c > 0$ (see Lebedev et al., 1965, Eq. 9.14.1).
Under the Assumptions (A1) and (A2), our main result is an approximate expression for the covariance of the log-average periodogram and is given by

$$ \begin{aligned} \operatorname{Cov}(Y_j^*, Y_{j'}^*) &= \sum_{n=1}^{\infty} \frac{n!\operatorname{tr}(\bar{\boldsymbol{\Omega}}_{j,j'}/m)^n}{(m/2)_n n^2}\{1+o(1)\} \\ &= \frac{2\operatorname{tr}(\bar{\boldsymbol{\Omega}}_{j,j'}/m)}{m}\, {}_3F_2(1,1,1,2,m/2+1;\operatorname{tr}(\bar{\boldsymbol{\Omega}}_{j,j'}/m))\{1+o(1)\}, \end{aligned} \tag{3.2} $$

where the o term is with respect to $p \to \infty$ and m fixed.

3.3 Derivation of the covariance of the log-average periodogram

Without loss of generality, we consider $j = 1$ and $j' = 2$. The desired covariance expression (3.2) is obtained from the second order derivative of the moment generating function

$$ \begin{aligned} M_{12}(\mu, \eta) &= \mathbb{E}\left[\exp\{\mu \log(Y_1^2 + \cdots + Y_m^2) + \eta \log(Y_{m+1}^2 + \cdots + Y_{2m}^2)\}\right] \\ &= \mathbb{E}\left\{(Y_1^2 + \cdots + Y_m^2)^{\mu}(Y_{m+1}^2 + \cdots + Y_{2m}^2)^{\eta}\right\} \end{aligned} $$

with respect to μ and η at $\mu = \eta = 0$. In particular, $M_{12}(\mu, \eta) \geq 0$. Let $\mathbf{Y}_j = (Y_{(j-1)m+1}, ..., Y_{jm})^t$ for $j = 1, 2$. The moment generating function is evaluated as

$$ M_{12}(\mu, \eta) = \mathbb{E}[(Y_{m+1}^2 + \cdots + Y_{2m}^2)^{\eta}\mathbb{E}\{(Y_1^2 + \cdots + Y_m^2)^{\mu}|\mathbf{Y}_2\}]. \tag{3.3} $$

The conditional distribution of $\mathbf{Y}_1$ given $\mathbf{Y}_2$ is Gaussian with mean $\boldsymbol{\kappa}_{12} = \boldsymbol{\Omega}_{1,2}\boldsymbol{\Omega}_{2,2}^{-1}\mathbf{Y}_2$ and covariance matrix $\boldsymbol{\Lambda}_{1,2} = \boldsymbol{\Omega}_{1,1} - \boldsymbol{\Omega}_{1,2}\boldsymbol{\Omega}_{2,2}^{-1}\boldsymbol{\Omega}_{2,1}$. Let $\mathbf{z} = (z_1, ..., z_m)^t$. Then,

$$\mathbb{E}\{(Y_1^2 + \cdots + Y_m^2)^\mu | \mathbf{Y}_2\} = \int_{\mathbb{R}^m} (\mathbf{z}^t\mathbf{z})^\mu \psi_{\boldsymbol{\kappa}_{12},\boldsymbol{\Lambda}_{12}}(\mathbf{z})\, \mathrm{d}z_1 \cdots \mathrm{d}z_m, \tag{3.4}$$

where

$$\psi_{\boldsymbol{\kappa},\boldsymbol{\Lambda}}(\mathbf{z}) = (2\pi)^{-m/2}|\boldsymbol{\Lambda}|^{-1/2} \exp\{-(\mathbf{z}-\boldsymbol{\kappa})^t\boldsymbol{\Lambda}^{-1}(\mathbf{z}-\boldsymbol{\kappa})/2\}$$

is the density function of the multivariate normal distribution with mean vector $\boldsymbol{\kappa}$ and covariance matrix $\boldsymbol{\Lambda}$. Let $\mathbf{U}^t\mathbf{S}\mathbf{U}$ be the eigendecomposition of $\boldsymbol{\Lambda}_{1,2}$ with $\mathbf{S} = \mathrm{diag}(s_1, ..., s_m)$, where $\mathbf{U}$ is an orthogonal matrix and $s_1, ..., s_m > 0$ are the eigenvalues. Define $\lambda_{12} = \mathrm{tr}(\boldsymbol{\Lambda}_{1,2})/m$ and let $\mathbf{I}$ denote the $(m \times m)$-dimensional identity matrix. We expand the density function $\psi_{\boldsymbol{\kappa}_{12},\boldsymbol{\Lambda}_{1,2}}$ with respect to the vector of eigenvalues $\mathbf{s} = (s_1, ..., s_m)$ at the point $\boldsymbol{\lambda} = (\lambda_{12}, ..., \lambda_{12})$. Using the product rule, the first-order Taylor expansion is given by

$$\psi_{\boldsymbol{\kappa}_{12},\boldsymbol{\Lambda}_{1,2}}(\mathbf{z}) = (2\pi)^{-m/2}\prod_{i=1}^m s_i^{-1/2} \exp\left[-\frac{1}{2}\sum_{i=1}^m \frac{\{\mathbf{U}(\mathbf{z}-\boldsymbol{\kappa})\}_i^2}{s_i}\right] \tag{3.5}$$

$$= \psi_{\boldsymbol{\kappa}_{12},\lambda_{12}\mathbf{I}}(\mathbf{z}) + \frac{\psi_{\boldsymbol{\kappa}_{12},\boldsymbol{\Lambda}_{12}^*}(\mathbf{z})}{2}\sum_{i=1}^m \left[\frac{\{\mathbf{U}(\mathbf{z}-\boldsymbol{\kappa}_{12})\}_i^2}{(s_i^*)^2} - \frac{1}{s_i^*}\right](s_i - \lambda_{12}), \tag{3.6}$$

where $\boldsymbol{\Lambda}_{12}^* = \mathbf{U}^t\mathrm{diag}(s_1^*, ..., s_m^*)\mathbf{U}$ for some point $\mathbf{s}^* = (s_1^*, ..., s_m^*)$ between $\mathbf{s}$ and $\boldsymbol{\lambda}$, and the subscript i denotes the i-th component. By Assumption (A1), the spectral density function f is α-Hölder continuous. Thus, by Lemma 1 of Klockmann and Krivobokova (2023a) and the Greshgorin theorem it holds for the set of eigenvalues $\lambda(\boldsymbol{\Omega}_{1,1})$ of $\boldsymbol{\Omega}_{1,1}$ that for some constant $C > 0$

$$\lambda(\boldsymbol{\Omega}_{1,1}) \subset \bigcup_{i=1}^m B[f\{\pi(i-1)/(p-1)\}, C\log(p)mp^{-\alpha}],$$

where $B[a, b]$ is a closed disc centered at a with radius b. Since $\boldsymbol{\Lambda}_{1,2}$ is the Schur complement of $\boldsymbol{\Omega}_{1,1}$, it follows that $\lambda_{\min}(\boldsymbol{\Omega}_{1,1}) \leq \lambda_{\min}(\boldsymbol{\Lambda}_{1,2}) \leq \lambda_{\max}(\boldsymbol{\Lambda}_{1,2}) \leq \lambda_{\max}(\boldsymbol{\Omega}_{1,1})$. Together with the α-Hölder continuity of f this implies $|s_i - \lambda_{12}| = \mathcal{O}\{\log(p)p^{-\alpha}\}$ uniformly over i for $p \to \infty$ and m fixed. Furthermore, $\boldsymbol{\Omega}_{1,1}$ is a submatrix of $\boldsymbol{\Omega}$ which is similar to the Toeplitz covariance matrix $\boldsymbol{\Sigma}$ of the process. Together with Assumption (A2), this implies $\lambda(\boldsymbol{\Lambda}_{1,2}) \subset [\epsilon, M]$ and $s_i^* \in [\epsilon, M]$.

Let $\Gamma(\cdot)$ denote the Gamma function and let

$$ {}_0F_1(c;z) = \sum_{n=0}^{\infty} \frac{1}{(c)_n} \frac{1}{n!} z^n, \qquad {}_1F_1(b,c;z) = \sum_{n=0}^{\infty} \frac{(b)_n}{(c)_n} \frac{1}{n!} z^n $$

denote the hypergeometric series for $b, c \in \mathbb{R}$ such that $c \neq 0, -1, ...$ The convergence region for both series is for $|z| < \infty$ (see Lebedev et al., 1965, Eq. 9.14.1). To calculate the conditional expectation (3.4), we first consider the expectation with respect to the density $\psi_{\boldsymbol{\kappa}_{12},\lambda_{12}\mathbf{I}}(\mathbf{z})$. Recall that the density function of a non-central chi-squared distribution with non-centrality parameter $\bar{\kappa}$ and m degrees of freedom is given by

$$ f(x;m,\bar{\kappa}) = \frac{\exp(-\bar{\kappa}/2)}{2^{m/2}\Gamma(m/2)} x^{m/2-1} \exp(-x/2)\, {}_0F_1(m/2; \bar{\kappa}x/4), $$

see e.g., Stuart and Ord (2010, Eq. 22.18). Setting $\bar{\kappa} = \boldsymbol{\kappa}_{12}^t \boldsymbol{\kappa}_{12}$, we can thus perform a substitution and rewrite the expectation with respect to the density of a non-central chi-squared distribution $\psi_{\boldsymbol{\kappa}_{12},\lambda_{12}\mathbf{I}}(\mathbf{z})$, i.e., conditioned on $\mathbf{Y}_2$, it holds almost surely that

$$ \begin{aligned} &\int_{\mathbb{R}^m} (\mathbf{z}^t\mathbf{z})^{\mu} \psi_{\boldsymbol{\kappa}_{12},\lambda_{12}\mathbf{I}}(\mathbf{z})\, \mathrm{d}z_1 \cdots \mathrm{d}z_m \\ &= \frac{\lambda_{12}^{\mu} \exp\{-\bar{\kappa}/(2\lambda_{12})\}}{2^{m/2}\Gamma(m/2)} \int_0^{\infty} x^{\mu+m/2-1} \exp(-x/2)\, {}_0F_1(m/2; \bar{\kappa}x/(4\lambda_{12}))\, \mathrm{d}x \\ &= \frac{\lambda_{12}^{\mu} \exp\{-\bar{\kappa}/(2\lambda_{12})\}}{2^{m/2}\Gamma(m/2)} \sum_{n=0}^{\infty} \frac{\{\bar{\kappa}/(4\lambda_{12})\}^n}{(m/2)_n n!} \int_0^{\infty} x^{\mu+m/2-1+n} \exp(-x/2)\, \mathrm{d}x. \end{aligned} $$

Next, we apply the identity

$$ \int_0^{\infty} x^{a-1} \exp(-bx)\, \mathrm{d}x = b^{-a}\Gamma(a) $$

for $a, b > 0$, see Gradshteyn and Ryzhik (2014, Eq. 3.381). Then, it holds for $\mu > -m/2$

$$ \begin{aligned} &\frac{1}{\Gamma(m/2)} \sum_{n=0}^{\infty} \frac{\{\bar{\kappa}/(4\lambda_{12})\}^n}{(m/2)_n n!} \int_0^{\infty} x^{\mu+m/2-1+n} \exp(-x/2)\, \mathrm{d}x \\ &= \frac{1}{\Gamma(m/2)} \sum_{n=0}^{\infty} \frac{\{\bar{\kappa}/(4\lambda_{12})\}^n}{(m/2)_n n!} (1/2)^{-(\mu+m/2+n)} \Gamma(\mu+m/2+n) \\ &= 2^{\mu+m/2} \frac{\Gamma(\mu+m/2)}{\Gamma(m/2)} \sum_{n=0}^{\infty} \frac{(\mu+m/2)_n \{\bar{\kappa}/(2\lambda_{12})\}^n}{(m/2)_n n!} \\ &= 2^{\mu+m/2} \frac{\Gamma(\mu+m/2)}{\Gamma(m/2)}\, {}_1F_1(m/2+\mu, m/2; \bar{\kappa}/(2\lambda_{12})), \end{aligned} $$

where we used that $\Gamma(a+n)/\Gamma(a) = (a)_n$. Finally, we apply the identity

$$\exp(z)\,{}_1F_1(a,c;-z) = {}_1F_1(c-a,c;z),$$

which holds for $a, c \in \mathbb{R}$ such that $c \neq 0, -1, \ldots$, see Lebedev et al. (1965, Eq. 9.11.2). Together, this gives for $\mu > -m/2$

$$\int_{\mathbb{R}^m} (\mathbf{z}^t\mathbf{z})^\mu \psi_{\boldsymbol{\kappa}_{12},\lambda_{12}\mathbf{I}}(\mathbf{z})\,\mathrm{d}z_1\cdots\mathrm{d}z_m = 2^\mu\lambda_{12}^\mu \frac{\Gamma(\mu+m/2)}{\Gamma(m/2)}\,{}_1F_1(-\mu, m/2; -\bar{\kappa}/(2\lambda_{12})). \quad (3.7)$$

In particular, (3.7) gives a simple formula for the non-integer moments of a scaled non-central chi-squared distribution. For computational purposes this formula is more convenient than the recently obtained formula by Rujivan et al. (2023, Cor. 2). Additionally, our formula holds for a range of negative exponents.

Next, we consider the expectation with respect to the first-order remainder term of the Taylor expansion. Let $c_1, c_2, c_3 > 0$ denote some constants not depending on m, p. Then, conditioned on $\mathbf{Y}_2$, it holds almost surely that

$$\begin{aligned}
&\int_{\mathbb{R}^m} (\mathbf{z}^t\mathbf{z})^\mu \frac{\psi_{\boldsymbol{\kappa}_{12},\boldsymbol{\Lambda}_{12}^*}(\mathbf{z})}{2} \sum_{i=1}^m \left[\frac{\{\mathbf{U}(\mathbf{z}-\boldsymbol{\kappa}_{12})\}_i^2}{(s_i^*)^2} + \frac{1}{s_i^*}\right](s_i - \lambda_{12})\,\mathrm{d}z_1\cdots\mathrm{d}z_m \qquad (3.8)\\
\leq& c_1 \max_{i=1,\ldots,m} |s_i - \lambda_{12}| \int_{\mathbb{R}^m} (\mathbf{z}^t\mathbf{z})^\mu \psi_{\boldsymbol{\kappa}_{12},\boldsymbol{\Lambda}_{12}^*}(\mathbf{z}) \sum_{i=1}^m \left[\frac{(\mathbf{U}\mathbf{z})_i^2 + (\mathbf{U}\boldsymbol{\kappa}_{12})_i^2}{(s_i^*)^2} + \frac{1}{s_i^*}\right]\mathrm{d}z_1\cdots\mathrm{d}z_m\\
\leq& c_2 \max_{i=1,\ldots,m} |s_i - \lambda_{12}| \int_{\mathbb{R}^m} \{(\mathbf{z}^t\mathbf{z})^{\mu+1} + \bar{\kappa}(\mathbf{z}^t\mathbf{z})^\mu + (\mathbf{z}^t\mathbf{z})^\mu\}\psi_{\boldsymbol{\kappa}_{12},\boldsymbol{\Lambda}_{12}^*}(\mathbf{z})\,\mathrm{d}z_1\cdots\mathrm{d}z_m\\
\leq& c_2 \max_{i=1,\ldots,m} |s_i - \lambda_{12}| \int_{\mathbb{R}^m} \{(\mathbf{z}^t\boldsymbol{\Lambda}_{12}^*\mathbf{z})^{\mu+1} + \bar{\kappa}(\mathbf{z}^t\boldsymbol{\Lambda}_{12}^*\mathbf{z})^\mu + (\mathbf{z}^t\boldsymbol{\Lambda}_{12}^*\mathbf{z})^\mu\}\psi_{(\boldsymbol{\Lambda}_{12}^*)^{-1/2}\boldsymbol{\kappa}_{12},\mathbf{I}}(\mathbf{z})\,\mathrm{d}z_1\cdots\mathrm{d}z_m\\
\leq& c_3 \max_{i=1,\ldots,m} |s_i - \lambda_{12}| \int_{\mathbb{R}^m} \{(\mathbf{z}^t\mathbf{z})^{\mu+1} + \bar{\kappa}(\mathbf{z}^t\mathbf{z})^\mu + (\mathbf{z}^t\mathbf{z})^\mu\}\psi_{(\boldsymbol{\Lambda}_{12}^*)^{-1/2}\boldsymbol{\kappa}_{12},\mathbf{I}}(\mathbf{z})\,\mathrm{d}z_1\cdots\mathrm{d}z_m,
\end{aligned}$$

where we used $(a-b)^2 \leq 2a^2 + 2b^2$, $s_i^* \in [\epsilon, M]$ and that $\bar{\kappa} \geq 0$ almost surely. The integrals in the last line can be solved with Equation (3.7). Substituting (3.7) and (3.8) in (3.3), we obtain for $\mu > -m/2$

$$\begin{aligned}
&M_{12}(\mu,\eta)\\
&= 2^\mu\lambda_{12}^\mu \frac{\Gamma(\mu+m/2)}{\Gamma(m/2)} \int_{\mathbb{R}^m} (\mathbf{z}^t\mathbf{z})^\eta\,{}_1F_1(-\mu, m/2; -\mathbf{z}^t\mathbf{A}_{1,2}\mathbf{z}/(2\lambda_{12}))\psi_{0,\boldsymbol{\Omega}_{2,2}}(\mathbf{z})\,\mathrm{d}z_1\cdots\mathrm{d}z_m\\
&\quad\times [1 + \mathcal{O}\{\log(p)p^{-\alpha}\}], \qquad (3.9)
\end{aligned}$$

where $\mathbf{A}_{1,2} = \boldsymbol{\Omega}_{2,2}^{-1}\boldsymbol{\Omega}_{2,1}\boldsymbol{\Omega}_{1,2}\boldsymbol{\Omega}_{2,2}^{-1}$ and the $\mathcal{O}$ term is with respect to $p \to \infty$ and m fixed. In particular, we used that the integrals

$$\int_{\mathbb{R}^m} (\mathbf{z}^t\mathbf{z})^\eta \, {}_1F_1(-\mu, m/2; -\mathbf{z}^t\tilde{\mathbf{A}}_{1,2}\mathbf{z}/(2\lambda_{12}))\psi_{0,\boldsymbol{\Omega}_{2,2}}(\mathbf{z})\,\mathrm{d}z_1\cdots\mathrm{d}z_m$$
$$\int_{\mathbb{R}^m} (\mathbf{z}^t\mathbf{z})^\eta \mathbf{z}^t\mathbf{A}_{1,2}\mathbf{z}\, {}_1F_1(-\mu, m/2; -\mathbf{z}^t\tilde{\mathbf{A}}_{1,2}\mathbf{z}/(2\lambda_{12}))\psi_{0,\boldsymbol{\Omega}_{2,2}}(\mathbf{z})\,\mathrm{d}z_1\cdots\mathrm{d}z_m,$$

where $\tilde{\mathbf{A}} = \boldsymbol{\Omega}_{2,2}^{-1}\boldsymbol{\Omega}_{2,1}(\boldsymbol{\Lambda}_{12}^*)^{-1}\boldsymbol{\Omega}_{1,2}\boldsymbol{\Omega}_{2,2}^{-1}$, are both of the order $\mathcal{O}(1)$ for $p \to \infty$ and m fixed. As discussed before, $\max_{i=1,\ldots,m}|s_i - \lambda_{12}| = \mathcal{O}\{\log(p)p^{-\alpha}\}$. Next, we insert the integral form of ${}_1F_1(b, c; z)$ into Equation (3.9), i.e.,

$${}_1F_1(b, c; z) = \frac{\Gamma(c)}{\Gamma(b)\Gamma(c-b)}\int_0^1 \exp(zt)t^{b-1}(1-t)^{c-b-1}\,\mathrm{d}t$$

which holds for $z, b, c \in \mathbb{R}$ such that $c > b > 0$ (see Lebedev et al., 1965, Eq. 9.11.1). Temporarily, we assume that $0 < -\mu < m/2$. Define $\mathbf{B}_{1,2} = \boldsymbol{\Omega}_{2,2}^{1/2}\mathbf{A}_{1,2}\boldsymbol{\Omega}_{2,2}^{1/2}$. Then,

$$\begin{aligned}
&\int_{\mathbb{R}^m} (\mathbf{z}^t\mathbf{z})^\eta \, {}_1F_1(-\mu, m/2; -\mathbf{z}^t\mathbf{A}_{1,2}\mathbf{z}/(2\lambda_{12}))\psi_{0,\boldsymbol{\Omega}_{2,2}}(\mathbf{z})\,\mathrm{d}z_1\cdots\mathrm{d}z_m \\
&= \frac{\Gamma(m/2)}{\Gamma(-\mu)\Gamma(m/2+\mu)}\int_0^1 t^{-\mu-1}(1-t)^{m/2+\mu-1} \\
&\quad \times \int_{\mathbb{R}^m} \frac{(\mathbf{z}^t\boldsymbol{\Omega}_{2,2}\mathbf{z})^\eta}{(2\pi)^{m/2}} \exp\left\{-\frac{\mathbf{z}^t\left(\mathbf{I} + \mathbf{B}_{1,2}t/\lambda_{12}\right)\mathbf{z}}{2}\right\}\mathrm{d}z_1\cdots\mathrm{d}z_m\,\mathrm{d}t. \qquad (3.10)
\end{aligned}$$

To compute the integral in (3.10), we perform two more Taylor expansions. Let $\mathbf{U}^t\mathbf{S}\mathbf{U}$ be the eigendecomposition of $\boldsymbol{\Omega}_{2,2}$ with $\mathbf{S} = \mathrm{diag}(s_1, ..., s_m)$. A first order Taylor expansion of the function $g(\mathbf{s}) = (\mathbf{z}^t\mathbf{U}^t\mathrm{diag}(s_1, ..., s_m)\mathbf{U}\mathbf{z})^\eta$ at $\boldsymbol{\omega} = (\omega_2, ..., \omega_2)$ with $\omega_2 = \mathrm{tr}(\boldsymbol{\Omega}_{2,2})/m$ gives

$$g(\mathbf{s}) = \omega_2^\eta(\mathbf{z}^t\mathbf{z})^\eta + \eta\{\mathbf{z}^t\mathbf{U}^t\mathrm{diag}(s_1^*, ..., s_m^*)\mathbf{U}\mathbf{z}\}^{\eta-1}\sum_{i=1}^m(\mathbf{U}\mathbf{z})_i^2(s_i - \omega_2),$$

where $\mathbf{s}^* = (s_1^*, ..., s_m^*)$ is between $\mathbf{s}$ and $\boldsymbol{\omega}$. By similar arguments as before it holds $\lambda(\boldsymbol{\Omega}_{2,2}) \subset [\epsilon, M]$ and $|s_i - \omega_2| = \mathcal{O}\{\log(p)p^{-\alpha}\}$ for $p \to \infty$ and uniformly over i.

Let $\mathbf{V}^t\mathbf{R}\mathbf{V}$ be the eigendecomposition of $\mathbf{B}_{1,2}$ where $\mathbf{R} = \mathrm{diag}(r_1, ..., r_m)$ with eigenvalues $r_1, ..., r_m \geq 0$ since $\mathbf{B}_{1,2}$ is positive semidefinite. A first order Taylor expansion of the function $h(\mathbf{r}) = \exp\{-\mathbf{z}^t\mathbf{V}^t\text{diag}(r_1,...,r_m)\mathbf{V}\mathbf{z}t/(2\lambda_{12})\}$ at the point $\boldsymbol{\gamma} = (\gamma_{12},...,\gamma_{12})$ with $\gamma_{12} = \text{tr}(\mathbf{B}_{1,2})/m$ gives

$$h(\mathbf{r}) = \exp\left\{-\mathbf{z}^t\mathbf{z}\frac{\gamma_{12}}{2\lambda_{12}}t\right\} - \frac{h(\mathbf{r}^*)t}{2\lambda_{12}}\sum_{i=1}^m(\mathbf{V}\mathbf{z})_i^2(r_i-\gamma_{12}),$$

where $\mathbf{r}^* = (r_1^*,...,r_m^*)$ is between $\mathbf{r}$ and $\boldsymbol{\gamma}$. The eigenvalues of $\mathbf{B}_{1,2}$ are non-negative and the largest eigenvalue is bounded by $c\{\log(p)mp^{-\alpha}\}^2$ for some constant $c > 0$, as the entries of $\boldsymbol{\Omega}_{1,2}\boldsymbol{\Omega}_{2,1}$ are of the order $\mathcal{O}[\{\log(p)p^{-\alpha}\}^2]$. In particular, it follows $|r_i - \gamma_{12}| = \mathcal{O}[\{\log(p)p^{-\alpha}\}^2]$ uniformly in i. The resulting integrals with respect to $z_1,...,z_m$ without the remainder terms of the Taylor expansion can be solved similar to the integral before in Equation (3.7), i.e., for $\eta > -m/2$

$$\int_{\mathbb{R}^m}\frac{\omega_2^\eta(\mathbf{z}^t\mathbf{z})^\eta}{(2\pi)^{m/2}}\exp\left\{-\frac{\mathbf{z}^t\mathbf{z}(1+\frac{\gamma_{12}}{\lambda_{12}}t)}{2}\right\}\mathrm{d}z_1\cdots\mathrm{d}z_m = \frac{\Gamma(\eta+m/2)}{\Gamma(m/2)}\frac{2^\eta\omega_2^\eta}{(1+\gamma_{12}t/\lambda_{12})^{\eta+m/2}}. \tag{3.11}$$

Next, we consider the integrals with respect to $z_1,...,z_m$ involving the remainder terms of the Taylor expansions which are given by

$$A{:=}\int_{\mathbb{R}^m}\frac{\eta\{\mathbf{z}^t\mathbf{U}^t\text{diag}(s_1^*,...,s_m^*)\mathbf{U}\mathbf{z}\}^{\eta-1}}{(2\pi)^{m/2}}\sum_{i=1}^m(\mathbf{U}\mathbf{z})_i^2(s_i-\omega_2)\exp\left\{-\frac{\mathbf{z}^t\mathbf{z}(1+\frac{\gamma_{12}}{\lambda_{12}}t)}{2}\right\}\mathrm{d}z_1\cdots\mathrm{d}z_m$$
$$B{:=}-\int_{\mathbb{R}^m}\frac{(\mathbf{z}^t\boldsymbol{\Omega}_{2,2}\mathbf{z})^\eta}{(2\pi)^{m/2}}\frac{h(\mathbf{r}^*)t}{2\lambda_{12}}\sum_{i=1}^m(\mathbf{V}\mathbf{z})_i^2(r_i-\gamma_{12})\exp\left\{-\frac{\mathbf{z}^t\mathbf{z}}{2}\right\}\mathrm{d}z_1\cdots\mathrm{d}z_m.$$

The integrals can be bounded as follows

$$\begin{aligned}
|A| &\leq \max_{i=1,...,m}|s_i-\omega_2|\cdot\max\{\epsilon^{\eta-1},M^{\eta-1}\}\int_{\mathbb{R}^m}\frac{\eta(\mathbf{z}^t\mathbf{z})^\eta}{(2\pi)^{m/2}}\exp\left\{-\frac{\mathbf{z}^t\mathbf{z}(1+\frac{\gamma_{12}}{\lambda_{12}}t)}{2}\right\}\mathrm{d}z_1\cdots\mathrm{d}z_m\\
|B| &\leq \max_{i=1,...,m}|r_i-\gamma_{12}|\cdot\max\{\epsilon^\eta,M^\eta\}\\
&\quad\times\int_{\mathbb{R}^m}\frac{(\mathbf{z}^t\mathbf{z})^{\eta+1}}{(2\pi)^{m/2}}\frac{1}{2\lambda_{12}}\exp\left\{-\frac{\mathbf{z}^t\mathbf{V}^t\text{diag}\,(\lambda_{12}+r_1^*t,...,\lambda_{12}+r_m^*t)\,\mathbf{V}\mathbf{z}}{2\lambda_{12}}\right\}\mathrm{d}z_1\cdots\mathrm{d}z_m\\
&\leq \max_{i=1,...,m}|r_i-\gamma_{12}|\max\{\epsilon^\eta,M^\eta\}\int_{\mathbb{R}^m}\frac{(\mathbf{z}^t\mathbf{z})^{\eta+1}}{(2\pi)^{m/2}}\frac{1}{2\lambda_{12}}\exp\left\{-\epsilon\frac{\mathbf{z}^t\mathbf{z}t}{2\lambda_{12}}\right\}\mathrm{d}z_1\cdots\mathrm{d}z_m
\end{aligned}$$

since $t \in [0,1]$ and $\lambda_{12} \in [\epsilon, M]$. In particular, $|A|, |B| = \mathcal{O}\{\log(p)p^{-\alpha}\}$ for $p \to \infty$ and m fixed. For $a, b, c \in \mathbb{R}$ such that $c \neq 0, -1, \ldots$, let

$$ {}_2F_1(a,b,c;z) = \sum_{n=0}^{\infty} \frac{(a)_n (b)_n}{(c)_n} \frac{z^n}{n!} $$

be the hypergeometric series, which converges absolutely for $|z| < 1$ and for $z = 1$ if $c - a - b > 0$ (see Lebedev et al., 1965, Eq. 9.1.2). The function ${}_2F_1(a,b,c;z)$ can be analytically extended to $z \in \mathbb{C} \setminus [1, \infty)$. An integral version for $z < 1$ and $a, b, c \in \mathbb{R}$ such that $c > b > 0$ is given by

$$ {}_2F_1(a,b,c;z) = \frac{\Gamma(c)}{\Gamma(b)\Gamma(c-b)} \int_0^1 \frac{t^{b-1}(1-t)^{c-b-1}}{(1-tz)^a} dt, $$

see Lebedev et al. (1965, Eq. 9.1.6). This implies

$$ \int_0^1 \frac{t^{-\mu-1}(1-t)^{m/2+\mu-1}}{(1+\gamma_{12}t/\lambda_{12})^{\eta+m/2}} \mathrm{d}t = \frac{\Gamma(m/2)}{\Gamma(-\mu)\Gamma(\mu+m/2)} {}_2F_1(\eta+m/2, -\mu, m/2; -\gamma_{12}/\lambda_{12}). $$

Together with equations (3.11), (3.10) and the bounds on the Taylor terms, one obtains

$$ \begin{aligned} M_{12}(\mu,\eta) = {} & 2^{\mu+\eta} \lambda_{12}^{\mu} \omega_2^{\eta} \frac{\Gamma(\mu+m/2)}{\Gamma(m/2)} \frac{\Gamma(\eta+m/2)}{\Gamma(m/2)} {}_2F_1(\eta+m/2, -\mu, m/2; -\gamma_{12}/\lambda_{12}) \\ & \times [1 + \mathcal{O}\{\log(p)p^{-\alpha}\}]. \end{aligned} \tag{3.12} $$

Note that $-\gamma_{12}/\lambda_{12} < 0$ since $\gamma_{12}, \lambda_{12} > 0$. Formula (3.12) was proven under the temporary assumption that $m/2 > -\mu > 0$. According to Chapter 9.4 of Lebedev et al. (1965), it is known that ${}_2F_1(a,b,c;z)/\Gamma(c)$ is an entire function of the parameters a, b, c. Hence, by the principle of analytic continuation, (3.12) can be analytically extended for $\mu > -m/2$. Let $\omega_1 = \mathrm{tr}(\mathbf{\Omega}_{1,1}/m)$. Next, we apply the linear transformation (Lebedev et al., 1965, Eq. 9.5.2)

$$ {}_2F_1(a,b,c;z) = (1-z)^{-b} {}_2F_1(c-a, b, c; z/(z-1)) $$

with

$$ \frac{z}{z-1} = -\frac{\gamma_{12}}{\lambda_{12}} = -\frac{\mathrm{tr}(\mathbf{\Omega}_{1,1}\bar{\mathbf{\Omega}}_{1,2}/m)}{\mathrm{tr}(\mathbf{\Omega}_{1,1}/m) - \mathrm{tr}(\mathbf{\Omega}_{1,1}\bar{\mathbf{\Omega}}_{1,2}/m)} = \frac{\mathrm{tr}(\mathbf{\Omega}_{1,1}\bar{\mathbf{\Omega}}_{1,2}/m)/\omega_1}{\mathrm{tr}(\mathbf{\Omega}_{1,1}\bar{\mathbf{\Omega}}_{1,2}/m)/\omega_1 - 1}, $$

where $\omega_1 = \mathrm{tr}(\mathbf{\Omega}_{1,1}/m)$. In particular, $z = \mathrm{tr}(\mathbf{\Omega}_{1,1}\bar{\mathbf{\Omega}}_{1,2}/m)/\omega_1$ and $\lambda_{12} = \omega_1(1-z)$. Together, this yields

$$M_{12}(\mu,\eta) = 2^{\mu+\eta}\omega_1^{\mu}\omega_2^{\eta}\frac{\Gamma(\mu+m/2)}{\Gamma(m/2)}\frac{\Gamma(\eta+m/2)}{\Gamma(m/2)}\,{}_2F_1(-\eta,-\mu,m/2;\mathrm{tr}(\mathbf{\Omega}_{1,1}\bar{\mathbf{\Omega}}_{1,2}/m)/\omega_1) \times [1+\mathcal{O}\{\log(p)p^{-\alpha}\}].$$

The Lipschitz continuity of ${}_2F_1(a,b,c;z)$ with respect to the argument z implies

$$\begin{aligned}&{}_2F_1(-\eta,-\mu,m/2;\mathrm{tr}(\mathbf{\Omega}_{1,1}\bar{\mathbf{\Omega}}_{1,2}/m)/\omega_1)\\ &= {}_2F_1(-\eta,-\mu,m/2;\mathrm{tr}(\bar{\mathbf{\Omega}}_{1,2}/m))[1+\mathcal{O}\{\log(p)p^{-\alpha}\}]\end{aligned}$$

because

$$|\mathrm{tr}(\mathbf{\Omega}_{1,1}\bar{\mathbf{\Omega}}_{1,2} - \omega_1\bar{\mathbf{\Omega}}_{1,2})| \leq \max_{i=1,\dots,m}|s_i-\omega_1|\mathrm{tr}(\bar{\mathbf{\Omega}}_{1,2}),$$

where $s_1,\dots,s_m$ are the eigenvalues of $\mathbf{\Omega}_{1,1}$. Note that $0 \leq \lambda_{12} = \omega_1 - \mathrm{tr}(\mathbf{\Omega}_{1,1}\bar{\mathbf{\Omega}}_{1,2}/m)$ implies $\mathrm{tr}(\mathbf{\Omega}_{1,1}\bar{\mathbf{\Omega}}_{1,2}/m)/\omega_1 \leq 1$. Similarly, $0 \leq \mathrm{tr}(\mathbf{\Omega}_{1,1}^{-1}\mathbf{\Lambda}_{12}/m)$ implies $\mathrm{tr}(\bar{\mathbf{\Omega}}_{1,2}/m) \leq 1$. All together we get

$$M_{12}(\mu,\eta) = 2^{\mu+\eta}\omega_1^{\mu}\omega_2^{\eta}\frac{\Gamma(\mu+m/2)}{\Gamma(m/2)}\frac{\Gamma(\eta+m/2)}{\Gamma(m/2)}\,{}_2F_1(-\eta,-\mu,m/2;\mathrm{tr}(\bar{\mathbf{\Omega}}_{1,2}/m)) \times [1+\mathcal{O}\{\log(p)p^{-\alpha}\}].$$

The second order derivative of $M_{12}(\mu,\eta)$ at $\mu=\eta=0$ can be obtained by using $\partial(\mu)_n/\partial\mu = -(n-1)!$ at $\mu = 0$ for $n > 0$. Since $0 \leq \mathrm{tr}(\bar{\mathbf{\Omega}}_{1,2}/m) \leq 1$, we can utilize the series representation ${}_2F_1(a,b,c;z) = \sum_{n=0}^{\infty}\frac{(a)_n(b)_n}{(c)_n}\frac{z^n}{n!}$ to compute the derivative. Since all integrals are finite, the moment generating function exists and is infinitely differentiable with respect to its parameters. Thus, the derivative of the $\mathcal{O}$ term evaluated at $\mu=\eta=0$ goes to zero at least as fast as $\log(p)p^{-\alpha}$ for $p\to\infty$. Some calculations using the product rule lead to

$$\begin{aligned}&\left.\frac{\partial^2 M_{12}(\mu,\eta)}{\partial\eta\partial\mu}\right|_{\substack{\mu=0\\ \eta=0}} - \left.\frac{\partial M_{12}(\mu,\eta)}{\partial\eta}\right|_{\substack{\mu=0\\ \eta=0}}\left.\frac{\partial M_{12}(\mu,\eta)}{\partial\mu}\right|_{\substack{\mu=0\\ \eta=0}}\\ &= \sum_{n=1}^{\infty}\left.\frac{\frac{\partial}{\partial\eta}(-\eta)_n\frac{\partial}{\partial\mu}(-\mu)_n\mathrm{tr}(\bar{\mathbf{\Omega}}_{1,2}/m)^n}{(m/2)_n n!}\right|_{\substack{\mu=0\\ \eta=0}}[1+\mathcal{O}\{\log(p)p^{-\alpha}\}]\\ &= \sum_{n=1}^{\infty}\frac{n!\,\mathrm{tr}(\bar{\mathbf{\Omega}}_{1,2}/m)^n}{(m/2)_n n^2}[1+\mathcal{O}\{\log(p)p^{-\alpha}\}],\end{aligned}$$

where we used that ${}_2F_1(a, 0, c; z) = {}_2F_1(0, b, c; z) = 1$. Finally,

$$\sum_{n=1}^{\infty} \frac{n! \mathrm{tr}(\bar{\mathbf{\Omega}}_{1,2}/m)^n}{(m/2)_n n^2} = \frac{2}{m} \sum_{n=1}^{\infty} \frac{(1)_{n-1}(1)_{n-1}(1)_{n-1}}{(m/2+1)_{n-1}(2)_{n-1}} \frac{\mathrm{tr}(\bar{\mathbf{\Omega}}_{1,2}/m)^n}{(n-1)!}$$
$$= \frac{2\mathrm{tr}(\bar{\mathbf{\Omega}}_{1,2}/m)}{m} {}_3F_2(1, 1, 1, 2, m/2+1; \mathrm{tr}(\bar{\mathbf{\Omega}}_{1,2}/m)),$$

since it holds $(n-1)! = (1)_{n-1}$, $n = (2)_{n-1}/(1)_{n-1}$ and $(m/2)_n = (m/2) \cdot (m/2+1)_{n-1}$ for $n \geq 1$.

3.4 Application

In a simulation study, we investigate the effect of the covariance formula (3.2) to improve the spectral density estimator of Klockmann and Krivobokova (2023a) which is based on the log-average periodogram. We consider a Gaussian time series of length $p = 60$ and with the autocovariance functions $\sigma : \mathbb{Z} \to \mathbb{R}$

(1) of an ARMA$(2,2)$ model with coefficients $(0.7, -0.6, -0.2, 0.2)$ and innovations variance 1.44,

(2) of a polynomial decay $\sigma(\tau) = 1.44(1 + 2|\tau|)^{-2.5}$.

Figure 4 shows the spectral densities for the two examples. A Monte Carlo simulation

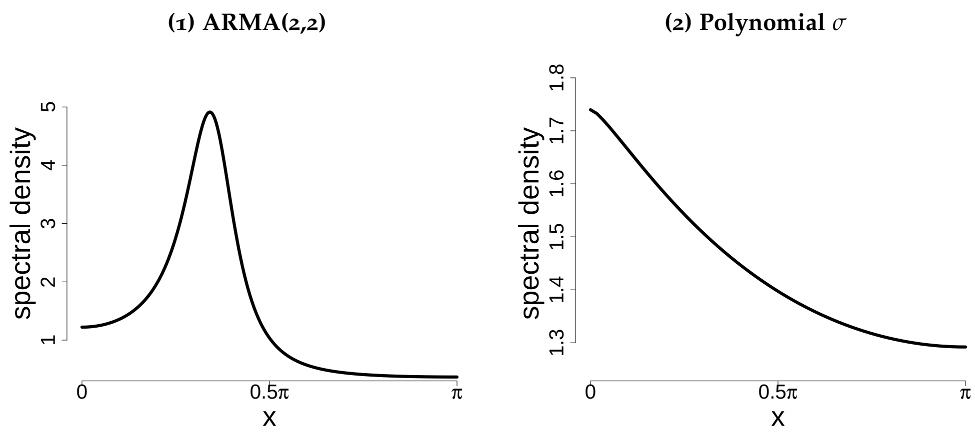

Figure 4: Spectral density functions for examples 1 and 2.

with 300 iterations is performed using R (version 4.1.2). We compare the unsmoothed periodogram, the original spectral density estimator of Klockmann and Krivobokova (2023a), and a variant which takes the covariance formula (3.2) without the $\mathcal{O}$ term into account. For the estimator of Klockmann and Krivobokova (2023a) cubic splines are used and the smoothing parameter is selected with generalized cross-validation. The number of bins is chosen to be $T = 12$. In Tables 10 and 11, the average L_∞ and L_2 estimation errors for the three different spectral density estimators are given for each example. In addition, the estimation error for the covariance matrix estimator $\hat{\boldsymbol{\Sigma}}$, obtained with the discrete Fourier transform form of the spectral density estimator, is shown with respect to the spectral norm. For the implementation we used the R package *vstdct*, see Klockmann and Krivobokova (2023c). We see a reduction between 3% and 6% in the L_2 norm, and between 7% and 13% in the L_∞ norm and the spectral norm, respectively. These findings offer promising insights for improving other estimators based on the log-average periodogram, particularly for scenarios involving smaller signal lengths where the derived covariance formula becomes more relevant.

	$\|\hat{f} - f\|_\infty$	$\|\hat{f} - f\|_2$	$\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_2$
raw periodogram	96.6029	7.2323	42.5832
smoothed, without decor.	11.2931	1.2896	10.4717
smoothed, with decor.	9.7487	1.2144	9.1225

Table 10: **Example (1)**: Errors of the spectral density estimators with respect to the L_2 norm and L_∞ norm, and estimation error for the covariance matrix $\boldsymbol{\Sigma}$ with respect to the spectral norm.

	$\|\hat{f} - f\|_\infty$	$\|\hat{f} - f\|_2$	$\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_2$
raw periodogram	31.0121	2.3038	9.7694
smoothed, without decor.	0.4602	0.1605	0.4546
smoothed, with decor.	0.4249	0.1556	0.4206

Table 11: **Example (2)**: Errors of the spectral density estimators with respect to the L_2 norm and L_∞ norm, and estimation error for the covariance matrix $\boldsymbol{\Sigma}$ with respect to the spectral norm.

3.5 Discussion

We have derived an approximate expression for the covariance of the log-average periodogram for stationary Gaussian time series that have an invertible covariance matrix and a bounded, α-Hölder continuous spectral density function. The main challenge in deriving the covariance expression (3.2) is computing the non-integer moments of a weighted sum of non-central chi-squared distributed random variable in (3.4) and the mixed moments of a weighted sum of chi-squared distributed random variables in (3.10). Recently, Rujivan et al. (2023) obtained an analytic expression for the non-integer moments of a sum of weighted non-central chi-squared distributed random variables, but this formula is only valid for positive exponents. For (3.10), there is currently no known analytic expression. However, by using a Taylor expansion, we can simplify (3.4) and (3.10) to an unweighted version. In particular, we derive formula (3.7) for the non-integer moments of the non-central chi-squared distribution with m degrees of freedom. This formula is simpler than the expression given by Corrollary 2 ofRujivan et al. (2023) with equal weights, for computational purposes more convenient and also holds for negative exponents $\mu > -m/2$.

The obtained covariance formula (3.2) reveals that the variance of the log-average periodogram is inversely proportional to the number of averaged frequencies, i.e., $\mathrm{Cov}(Y_j^*, Y_j^*) = \phi^{(1)}(m/2)\{1 + o(1)\} \approx 2/m$, where $\phi^{(1)}$ denotes the trigamma function. For $j \neq j'$, the covariance of two components of the log-average periodogram is of the order $\mathrm{Cov}(Y_j^*, Y_{j'}^*) = \mathcal{O}\{\mathrm{tr}(\bar{\boldsymbol{\Omega}}_{j,j'}/m^2)\} = \mathcal{O}\{\log(p)^2 p^{-2\alpha}\}$ for $p \to \infty$. For $m = 1$, the Taylor expansions are not necessary and our results correspond to the exact covariance formula for the log-periodogram derived by Ephraim and Roberts (2005) up to a factor of $1/2$. Ephraim and Roberts (2005) used the discrete Fourier transform to define the periodogram, which implies that the spectral components are complex Gaussian distributed. When the discrete Fourier transform is used, the spectral components are real-valued Gaussian distributed, resulting in simpler calculations and the additional factor. The derivation of the covariance expression (3.2) can also be performed using the discrete Fourier transform, resulting in the same formula but without the $1/2$ factor. Finally, we conjecture that a similar formula could be derived for Gaussian time series with only positive semidefinite covariance matrices by carefully handling the quadratic forms involving the degenerate Gaussian distribution and using generalized matrix inversions.

4 Joint nonparametric estimation of mean and autocovariances for Gaussian processes

This chapter is based on:
Krivobokova, T., Serra, P., Rosales, F., and Klockmann, K. (2022). Joint non-parametric estimation of mean and auto-covariances for gaussian processes. *Comput. Statist. Data Anal.*, 173:107519

Abstract

Gaussian processes that can be decomposed into a smooth mean function and a stationary autocorrelated noise process are considered. A fully automatic nonparametric method for simultaneous estimation of the mean and autocovariance functions of such processes is developed. The proposed empirical Bayes approach is data-driven, numerically efficient, and allows for the construction of confidence sets for the mean function. The performance of the method is demonstrated through simulations and real data analysis. The method is implemented in the R package *eBsc*[1].

Keywords: Demmler-Reinsch basis, empirical Bayes, spectral density, stationary process.

4.1 Introduction

Consider the following observations from a fixed design, nonparametric regression model

$$
\begin{aligned}
Y_i = \mu(t_i) + \sigma\epsilon_i,\ \mathbb{E}(\epsilon_i) = 0,\ \sigma > 0,\ \ i = 1, \ldots, p, \\
\mathrm{Cor}(\epsilon_i, \epsilon_j) = r(i-j) = r_{|i-j|} \in [-1, 1],\ \ i, j = 1, \ldots, p,
\end{aligned}
\tag{4.1}
$$

where $r(\cdot)$ denotes the autocorrelation function of the underlying noise process, and $r_{|i|}$ denotes autocorrelation at lag i. The design points $t_i \in \mathbb{R}$ are equidistant and

1 The *eBsc* package is available on *CRAN*.

represent time points. The functions μ and r are unknown, but it is assumed that the noise terms $(\epsilon_i)_{i=1}^p$ are sampled from a stationary noise process and that μ is a smooth function. (In this paper, the Gaussian process is used as a template for such data, but the methodology can be applied to other processes as well.) The observations $(Y_i)_{i=1}^p$ might represent measures of some experimental quantity observed with a time dependent measurement error. In this case, estimation of μ is of interest, while r is considered as a nuisance parameter. Alternatively, $(Y_i)_{i=1}^p$ could be measurements from a stochastic process indexed by time or space with a seasonal or other deterministic effect described by μ. In this case the focus is on the estimation of the autocorrelation r.

Nonparametric estimation of the covariance structure of a vector of observations with *known* mean is already a challenging problem. This problem is important in time series analysis, such as for prediction, as well as in multivariate analysis for problems like clustering, principal component analysis (PCA), linear and quadratic discriminant analysis, and regression analysis. Two frameworks are usually considered when the mean is known: either there are p independent observations of a p-dimensional vector with correlated components (Bickel and Levina, 2008a,b), or there is one observation of a p-dimensional vector sampled from a stationary process (Xiao and Wu, 2012). The second framework is the one relevant for the present setting. Regardless of the framework, natural (moment or sample autocovariances) estimators for the covariance matrix of the observed vectors are well known to be inconsistent, e.g., in the operator norm. Some form of regularization, such as banding, tapering, thresholding, is necessary to ensure positive definiteness and consistency of the estimator. There is no simple data-driven approach to select the regularization parameter in this context, and thus fully taking the correlation structure into account (Purahmadi, 2011). The minimax rates of estimators for the correlation structure depend on the decay of the autocorrelation function as the lag increases or the smoothness of the corresponding spectral density (Yang et al., 2001; Cai et al., 2010; Purahmadi, 2011; Xiao and Wu, 2012; Fan et al., 2016). Therefore, nonparametric estimation of the covariance structure of the process is of great importance.

When the mean function is unknown, then a natural approach to covariance estimation is to first obtain a consistent estimator for the mean function. Subsequently, methods for (nonparametric) covariance estimation with a known mean can be applied. Unfortunately, ignoring the dependence on the error process hampers the accurate estimation of the mean function, denoted as μ.

To exemplify the problem, consider a time series of hourly loads (kW) for a US utility (depicted as grey line in Figure 5), which is described in detail in Section 4.4. This process has clear seasonal effects over years and possibly over weeks and days. These deterministic effects can be captured by a function μ. However, when no parametric assumptions are made about μ and the errors are treated as independent and identically distributed (i.i.d.), the standard nonparametric estimators of μ are significantly affected by the neglected dependence in the errors. In particular, automatic selectors of the smoothing parameter, such as cross-validation, tend to choose a biased smoothing parameter, resulting in an almost interpolating estimator in case of positively correlated $(\epsilon_i)_{i=1}^{p}$. This is illustrated by the black line in the left plot of Figure 5. This problem has been known for several decades, for an overview see Opsomer et al. (2001).

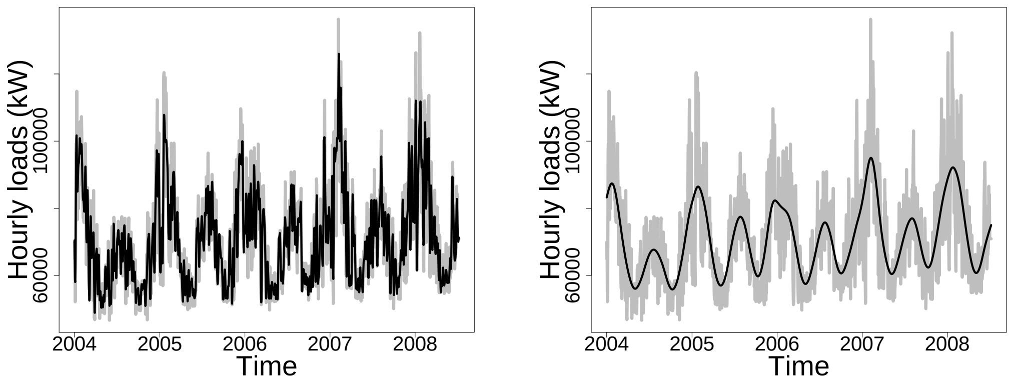

Figure 5: Hourly loads (kW) for a US utility: Estimation ignoring dependence in the data (left) and assuming that errors follow an autoregressive process of order one (right). The grey line shows the data, the black lines show the estimators. The estimators are obtained using the functions `gam` (left) and `gamm` (right) of the R package *mgcv*.

Hence, there is a need for an approach that enables accurate smoothing of the mean function across a wide range of possible autocovariance functions. Existing methods often rely on explicit parametric assumptions regarding the correlation structure, such as assuming that $(\epsilon_i)_{i=1}^{p}$ follows an ARMA(p,q) process. Once the autocorrelation function r is parametrized, the usual smoothing parameter selection criteria are adjusted to incorporate r, allowing simultaneous estimation of both μ and r.

For example, Hart (1994) introduce time series cross-validation for estimating μ using kernel estimators, assuming that the errors follow an AR(p) process, see also Altman (1990) and Hall and Keilegom (2003). Kohn et al. (1992) use spline smoothing for estimating μ and a general ARMA(p, q) model for the errors, estimating all parameters through generalized cross-validation or maximum likelihood. Similar ideas are employed in Wang (1998). While this approach allows for simultaneous estimation of μ and r, it is limited to parametric models for r. However, the true correlation structure may be more complex than any parametric assumption, and if the parametric model for the error process is misspecified, it significantly impacts the estimator of μ. Moreover, in practice it is difficult to select and verify a parametric model for r when the mean function μ is unknown.

Smoothing with low-rank splines and an ARMA(p, q) model for the residuals is implemented in the `gamm` function of the R package *mgcv*. This function has been used to estimate hourly loads in the right plot of Figure 5, assuming that the errors follow an AR(1) process. However, verifying the appropriateness of this model for the given data can be challenging. For a more detailed analysis of this dataset, please refer to Section 4.4.

Another group of methods for estimating μ focuses on mitigating the influence of correlation on the smoothing parameter, treating r as a nuisance parameter. For example, Chu and Marron (1991) and Hall et al. (1995) study the modified cross-validation criterion obtained by excluding a whole block of $2l + 1$ observations around each observation. Chiu (1989) and Hurvich and Zeger (1990) study Mallow's C_p and cross-validation in the frequency domain. Herrmann et al. (1992) and Lee et al. (2010) take a different approach by incorporating some sample estimators of the autocovariance function, which depend on unknown parameters linked to assumptions on the error process, into the smoothing parameter selection criteria. These methods require careful selection of tuning parameters, for which no data-driven methods are currently available.

In this paper, we present a likelihood-based method for estimating a regression function μ observed with correlated additive noise. We also develop estimators for the noise level σ and the autocorrelation function r. Moreover, the proposed empirical Bayesian framework provides a computationally efficient approach for constructing confidence sets for the regression function that take the correlation structure into account. There are several novel points in this work. Unlike other approaches in the literature, the proposed method is completely automatic so that no tuning parameters need to be set by the user. Additionally, it is fully nonparametric. Our approach to estimating the autocorrelations is also novel. We employ spline smoothers to estimate the spectral density of the noise process. The autocorrelations are then reconstructed from the the spectral density estimate, rather than through tapering or thresholding some empirical estimate. The resulting covariance matrix is positive definite by construction and consistent in the operator norm. Further investigations into the properties of this estimator require the development of new tools, which will be explored in a separate work.

This paper is structured as follows. In Section 4.2 the proposed estimators are introduced, Section 4.3 contains simulation results, Section 4.4 presents a real data example, and Section 4.5 closes the paper with some conclusions. Section 4.6 gives the proofs and auxiliary results.

4.2 Construction of the estimators

Assume that the regression function μ in model (4.1) belongs to a Sobolev space $\mathcal{W}_\beta$, $\beta > 1/2$, see Section 4.6.1 for more details. The noise terms ϵ_i are sampled from a stationary Gaussian noise process with zero mean and variance $\sigma^2 > 0$. The Gaussian model is used here as a template to construct estimators, but a penalization framework is followed (see Section 4.2) so that the estimators are expected to perform well under other data distributions. To show consistency of the estimators, infill asymptotics, as discussed in Robinson (1989), are employed. Define $\mathbf{Y} = (Y_1, \ldots, Y_p)^t$, $\boldsymbol{\mu} = \mu(\mathbf{t}) = \{\mu(t_1), \ldots, \mu(t_p)\}^t$, $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_p)^t$, where $t_i = (i-1)/(p-1)$, $i = 1, \ldots, p$. The observations in (4.1) are modeled as

$$\mathbf{Y} = \boldsymbol{\mu} + \sigma\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{R}).$$

The symmetric $(p \times p)$-dimensional correlation matrix $\mathbf{R}$ has Toeplitz structure and satisfies the regularity assumptions described in Lemma 4.1. These do imply that the first row of $\mathbf{R}$ is absolutely summable which means that the noise process is modeled as having short-term dependence. However, the assumptions are otherwise quite mild; in particular, a large (nonparametric) collection of covariance structures for the noise terms is considered, for instance ARMA noise models as a very particular case. It is also assumed that for some $0 < \delta < 1$, the eigenvalues of $\mathbf{R}$ lie in the interval $[\delta, 1/\delta]$.

The regression function μ is estimated using a smoothing spline, i.e., $\hat{\mu}$ that solves

$$\min_{\mu \in \mathcal{W}_q} \left[\frac{1}{p} \{\mathbf{Y} - \mu(\mathbf{t})\}^t \mathbf{R}^{-1} \{\mathbf{Y} - \mu(\mathbf{t})\} + \lambda \int_0^1 \{\mu^{(q)}(t)\}^2 \, dt \right], \tag{4.2}$$

for some $q \in \mathbb{N}$, $\lambda > 0$ and some "working" correlation matrix $\mathbf{R}$. The solution is well known to be a natural polynomial smoothing spline of degree $2q - 1$. Subsequently, q, λ, and $\mathbf{R}$ are estimated from the data using the empirical Bayes approach. It is well-known that for given $\mathbf{R}$, q and λ, the resulting estimator $\hat{\mu}$ is a natural spline of degree $2q - 1$ with knots at $\mathbf{t}$ and can be written as $\hat{\mu}(t) = \mathbf{S}(t)\mathbf{Y}$, where $\mathbf{S}$ is a $(p \times p)$-dimensional smoother matrix. To represent $\mathbf{S}$, the so-called Demmler-Reinsch basis of the natural spline space of degree $2q - 1$ is used, which is defined in Section 4.6.1. As such, the smoother matrix $\mathbf{S}$ can be written as

$$\mathbf{S} = \mathbf{S}_{\lambda,q,\mathbf{R}} = \mathbf{\Phi}_q \left\{ \mathbf{\Phi}_q^t \mathbf{R}^{-1} \mathbf{\Phi}_q + \lambda \, \mathrm{diag}(p\boldsymbol{\eta}_q) \right\}^{-1} \mathbf{\Phi}_q^t \mathbf{R}^{-1}, \tag{4.3}$$

where $\mathbf{\Phi}_q$ is an appropriate $(p \times p)$-dimensional basis matrix and $\boldsymbol{\eta}_q \in \mathbb{R}^p$ is a vector of eigenvalues. This representation makes the dependence of $\mathbf{S}$ on the parameters λ, $\mathbf{R}$, and q more explicit. To keep the notation simple, the dependence on these parameters is omitted, unless these are set to a particular value.

4.2.1 Estimation of the remaining parameters

Although a frequentist paradigm is followed, the estimator $\hat{\mu}$ has a Bayesian interpretation which provides us with a convenient way of estimating all of the unknown parameters by employing the empirical Bayes approach. Later, this Bayesian interpretation is used to obtain confidence sets; see Section 4.2.3.

The regression function $\boldsymbol{\mu}$ is endowed with a partly informative prior – given $(\mathbf{t}, \lambda, q, \sigma^2, \mathbf{R})$, the prior on $\boldsymbol{\mu}$ admits a density proportional to

$$\left| \frac{\mathbf{R}^{-1}(\mathbf{S}^{-1} - \mathbf{I}_p)}{2\pi\sigma^2} \right|_+^{1/2} \exp\left\{ -\frac{\boldsymbol{\mu}^t \mathbf{R}^{-1}(\mathbf{S}^{-1} - \mathbf{I}_p)\boldsymbol{\mu}}{2\sigma^2} \right\}, \tag{4.4}$$

where $|\cdot|_+$ denotes the product of non-zero eigenvalues ($\mathbf{S}$ has exactly q eigenvalues equal to 1), and $\mathbf{I}_p$ is the identity matrix. This prior corresponds to a non-informative part on the null-space of $\mathbf{R}^{-1}(\mathbf{S}^{-1} - \mathbf{I}_p)$ and a proper Gaussian prior on the remaining space. (This prior distribution happens to be independent of $\mathbf{R}$. This follows from the identity $\mathbf{R}^{-1}(\mathbf{S}^{-1} - \mathbf{I}_p) = \mathbf{S}_{\mathbf{I}}^{-1} - \mathbf{I}_p$, where $\mathbf{S}_{\mathbf{I}}$ denotes the smoother matrix with $\mathbf{R} = \mathbf{I}_p$, see Section 4.6.2 for the derivation of the identity. The mean of the corresponding posterior distribution for $\boldsymbol{\mu}|(\mathbf{t}, \lambda, \sigma^2, \mathbf{R})$ is the smoothing spline estimator $\hat{\boldsymbol{\mu}} = \mathbf{SY}$, see Speckman and Sun (2003). The variance σ^2 given $(\mathbf{t}, \lambda, q, \mathbf{R})$ is endowed with an inverse-gamma prior $\mathrm{IG}(a, b)$ with $a, b > 0$.

The resulting prior on $(\boldsymbol{\mu}, \sigma^2)|(\lambda, q, \mathbf{R})$ is conjugate for model (4.1) in the sense that the posterior distribution on $(\boldsymbol{\mu}, \sigma^2)|(\lambda, q, \mathbf{R})$ is a known distribution. Namely, the marginal posterior for σ^2 given $(\lambda, q, \mathbf{R})$ is an inverse-gamma distribution with shape parameter $(p - q + 2a)/2$ and scale parameter $\{\mathbf{Y}^t\mathbf{R}^{-1}(\mathbf{I}_p - \mathbf{S})\mathbf{Y} + 2b\}/2$. As for $\boldsymbol{\mu}|(\lambda, q, \mathbf{R})$, its posterior distribution is a multivariate t-distribution (see Kotz and Nadarajah (2004) for the definition of this distribution) with $p + 1$ degrees of freedom, mean $\hat{\boldsymbol{\mu}} = \mathbf{SY}$, and scale $\hat{\sigma}^2\mathbf{SR}$, where

$$\hat{\sigma}^2 = \frac{\mathbf{Y}^t\mathbf{R}^{-1}(\mathbf{I}_p - \mathbf{S})\mathbf{Y} + 1}{p + 1}. \tag{4.5}$$

It remains to estimate λ, q and $\mathbf{R}$ which are parameters of the prior. To do so, the empirical Bayes approach is used: these parameters are estimated from the marginal distribution of $\mathbf{Y}$, given $(\mathbf{t}, \lambda, q, \mathbf{R})$, which is a multivariate t-distribution with density

$$\frac{\Gamma(a + \frac{p-q}{2}) \left|\mathbf{R}^{-1}(\mathbf{I}_p - \mathbf{S})\frac{2a-q}{2b}\right|_+^{1/2}}{\{\pi(2a - q)\}^{p/2}\Gamma(a - q/2)} \left\{1 + \frac{\mathbf{Y}^t\mathbf{R}^{-1}(\mathbf{I}_p - \mathbf{S})\mathbf{Y}}{2b}\right\}^{-(p+2a-q)/2}, \tag{4.6}$$

where $\Gamma(\cdot)$ denotes the Gamma function. It remains to set the parameters a and b. The choice of parameters a and b is irrelevant for the asymptotics (as long as a and b are $\mathcal{O}(p)$ and lead to a proper prior and marginal distributions). To simplify the log-likelihood, these are set to $b = 1/2$ and $a = (q + 1)/2$. This yields

$$\ell_p(\lambda, q, \mathbf{R}) = -\frac{p+1}{2}\log\left\{\mathbf{Y}^t\mathbf{R}^{-1}(\mathbf{I}_p - \mathbf{S})\mathbf{Y} + 1\right\} + \frac{1}{2}\log\left|\mathbf{R}^{-1}(\mathbf{I}_p - \mathbf{S})\right|_+ \tag{4.7}$$

up to an additive constant that is independent of the parameters of interest. With this choice of a and b the posterior mean for σ^2 becomes $\hat{\sigma}^2$ as defined in (4.5) which is the proposed estimator for the variance of the noise.

4.2.2 Estimating equations and algorithm

The log-likelihood (4.7) depends on the unknown parameters via the matrix $\mathbf{R}^{-1}(\mathbf{I}_p - \mathbf{S})$ only and can be represented in a more convenient way as

$$\mathbf{R}^{-1}(\mathbf{I}_p - \mathbf{S}) = \mathbf{\Phi}_q \left\{ \mathbf{I}_p + \operatorname{diag}(\lambda p \boldsymbol{\eta}_q) \mathbf{\Phi}_q^t \mathbf{R} \mathbf{\Phi}_q \right\}^{-1} \operatorname{diag}(\lambda p \boldsymbol{\eta}_q) \mathbf{\Phi}_q^t,$$

using the Demmler-Reinsch basis (see Section 4.6.1). The key result that enable efficient simultaneous estimation of μ and r (or, equivalently, $\mathbf{R}$) is given in the following lemma.

Lemma 4.1 *Let* $\mathbf{R}$ *be a* $(p \times p)$*-dimensional correlation matrix of a stationary process with spectral density function* ρ*. Set* $\rho_j = \rho(\pi t_j)$ *and denote by* $\delta_{i,j}$ *the Kronecker delta.*

(i) If ρ *belongs to the Hölder space* $\mathcal{H}_\beta(M)$*, where* $\beta = \gamma + \alpha, M > 0$ *with* $\gamma \in \mathbb{N}$*,* $0 < \alpha \leq 1$*, i.e.,* $\rho < M$ *and* $\rho^{(\gamma)}$ *is* α*-Hölder continuous, then, for* $i, j = q+1, \ldots, p$*,*

$$(\mathbf{\Phi}_q^t \mathbf{R} \mathbf{\Phi}_q)_{i,j} = \rho_j \delta_{i,j} + \frac{1 + (-1)^{|i-j|}}{2} \cdot \begin{cases} \mathcal{O}\{\log(p) p^{-\beta+1}\}, & 1 < \beta \leq 2, \\ \mathcal{O}(p^{-1}), & 2 < \beta. \end{cases}$$

(ii) If ρ *belongs to the Sobolev space* $\mathcal{W}_\beta(M)$*, then, for* $i, j = q+1, \ldots, p$*,*

$$(\mathbf{\Phi}_q^t \mathbf{R} \mathbf{\Phi}_q)_{i,j} = \rho_j \delta_{i,j} + \frac{1 + (-1)^{|i-j|}}{2} \cdot \begin{cases} \mathcal{O}\{\log(p) p^{-1}\}, & \beta = 2, \\ \mathcal{O}(p^{-1}), & \beta = 3, 4, \ldots. \end{cases}$$

(iii) If the autocorrelation function satisfies $r_k = \mathcal{O}(k^{-\beta})$ *for* $\beta > 1$ *and* $\rho < M$ *for some constant* $M > 0$*, then the same result as in part (i) holds.*

Note that the $\mathcal{O}(\cdot)$ terms are uniform over i, j. The proof can be found in Section 4.6.5. Hence, the Demmler-Reinsch basis $\mathbf{\Phi}_q$ asymptotically diagonalizes $\mathbf{R}$ for any q. As a result, the log-likelihood (4.7) can be represented as

$$\begin{aligned} \ell_p(\lambda, q, \mathbf{R}) = &-\frac{p+1}{2} \log \left[\sum_{i=q+1}^{p} \frac{B_i^2 \lambda p \eta_{q,i}}{1 + \lambda p \eta_{q,i} \rho_i} \{1 + o(1)\} + \mathcal{O}_p(1) \right] \\ &+ \frac{1}{2} \sum_{i=q+1}^{p} \log \left(\frac{\lambda p \eta_{q,i}}{1 + \lambda p \eta_{q,i} \rho_i} \right) \{1 + o(1)\}, \end{aligned} \tag{4.8}$$

where $B_i = \{\mathbf{\Phi}_q^t \mathbf{Y}\}_i$, see Section 4.6.6. The representation in (4.8) should apply in greater generality, but Lemma 4.1 already provides a rather large model for the covariance structure of the noise.

Since $\eta_{q,i} = [\pi\{i-(q+1)/2\}]^{2q}\{1+o(1)\}$ (see Section 4.6.1 for more details), it is straightforward to differentiate the approximate log-likelihood (4.8) with respect to λ, q and ρ_i. After appropriate scaling, the following estimating equations are obtained

$$\begin{aligned}
T_\lambda(\lambda, q, \boldsymbol{\rho}) &= \left\{ \sum_{i=q+1}^{p} \frac{B_i^2 \lambda p \eta_{q,i}}{(1+\lambda p \eta_{q,i}\rho_i)^2} - \hat{\sigma}^2 \sum_{i=q+1}^{p} \frac{1}{1+\lambda p \eta_{q,i}\rho_i} \right\} \{1+o(1)\}, \\
T_q(\lambda, q, \boldsymbol{\rho}) &= \left\{ \sum_{i=q+1}^{p} \frac{B_i^2 \lambda p \eta_{q,i} \log(p\eta_{q,i})}{(1+\lambda p \eta_{q,i}\rho_i)^2} - \hat{\sigma}^2 \sum_{i=q+1}^{p} \frac{\log(p\eta_{q,i})}{1+\lambda p \eta_{q,i}\rho_i} \right\} \{1+o(1)\}, \\
T_{\rho_i}(\lambda, q, \boldsymbol{\rho}) &= \left(\frac{B_i^2 \lambda p \eta_{q,i}\rho_i}{1+\lambda p \eta_{q,i}\rho_i} - \rho_i \right) \{1+o(1)\}, \quad i = 1, \ldots, p,
\end{aligned}$$

where

$$\hat{\sigma}^2 = \frac{1}{p+1} \left(\sum_{i=q+1}^{p} \frac{B_i^2 \lambda p \eta_{q,i}}{1+\lambda p \eta_{q,i}\rho_i} + 1 \right).$$

Joint estimates for all unknown model parameters are obtained by solving these equations simultaneously over $\lambda > 0$, $\rho_i > 0$, $i = 1, \ldots, p$, and $q \in \mathbb{N}$. However, this cannot be done directly due to interdependence and nonlinearities in the equations. In practice, the equations are solved iteratively. For each fixed q, start with an initial guess $\hat{\mathbf{R}}^{(0)}$ (typically the identity matrix), to obtain a preliminary estimate $\hat{\lambda}^{(0)}$. Iterate to get $\hat{\lambda}_q$ and $\hat{\mathbf{R}}_q$. Finally, the value of q is determined by solving $T_q(\hat{\lambda}_q, q, \hat{\mathbf{R}}_q) = 0$.

Since $\rho_i = \rho(\pi t_i)$ are values of a smooth function ρ at given points, the estimation of ρ should be carried out over a space of smooth functions. In the Bayesian framework, this can be achieved by introducing an appropriate prior on $\mathbf{R}$, which acts as a penalty term. However, for simplicity, a post-processing procedure is performed instead. First, $\tilde{\rho}_i$ are obtained as solutions of the corresponding estimating equations. Then, a smoothing step is performed on $\tilde{\rho}_i$ yielding $\hat{\boldsymbol{\rho}} = \mathbf{S}_{\xi,q,\mathbf{I}}\tilde{\boldsymbol{\rho}}$, where $\mathbf{S}_{\xi,q,\mathbf{I}}$ is a smoother matrix (4.3) with parameters ξ, q and $\mathbf{I}_p$. (Note that this is inevitable: the estimates $\tilde{\rho}_i$ are inconsistent without smoothing.)

In summary, for each fixed q, at each step $\hat{\lambda}^{(j)}$ and $\tilde{\rho}_i^{(j)}$ are obtained as solutions of the corresponding estimating equations and $\tilde{\rho}_i^{(j)}$ are smoothed to get $\hat{\rho}_i^{(j)}$. The algorithm is iterated until convergence of $\hat{\lambda}$ and $\hat{\boldsymbol{\rho}}$, which are then used to get $\hat{q}$. Finally, an estimate of the correlation matrix $\hat{\mathbf{R}}$ is obtained from $\hat{\boldsymbol{\rho}}$ by the discrete Fourier transform; for the details see Section 4.6.3. The estimation procedure is summarized in Algorithm 1.

for *q in Q_p* **do**
  set $j = 1$ and $\hat{\boldsymbol{\rho}}^{(0)} \equiv 1$
  while *stopping criterium not met* **do**
    set $\hat{\lambda}^{(j)}$ to a solution of $T_\lambda(\lambda, q, \hat{\boldsymbol{\rho}}^{(j-1)}) = 0$;
    set $\tilde{\rho}_i^{(j)}$ to a solution of $T_{\rho_i}(\lambda^{(j)}, q, \hat{\rho}_i^{(j-1)}) = 0$;
    compute $\hat{\boldsymbol{\rho}}^{(j)}$ by smoothing $\tilde{\boldsymbol{\rho}}^{(j)}$;
    set $j = j + 1$;
  end
  set $\hat{\lambda}_q = \hat{\lambda}^{(j)}$ and $\hat{\boldsymbol{\rho}}_q = \hat{\boldsymbol{\rho}}^{(j)}$;
end
set $\hat{q}$ to a solution of $T_q(\hat{\lambda}_q, q, \hat{\boldsymbol{\rho}}_q) = 0$ over Q_p;
set $\hat{\lambda}$ to $\hat{\lambda}_{\hat{q}}$;
set $\hat{\boldsymbol{\rho}} = \hat{\boldsymbol{\rho}}_{\hat{q}}$;
set $(\hat{\mathbf{R}})_{i,j} = \hat{r}_{|i-j|}$ with $\hat{r}_k = p^{-1} \sum_{l=1}^{p} \cos\{k\pi(l-1)/(p-1)\}\, \hat{\rho}_l$;

Algorithm 1: Estimation procedure.

Here, Q_p denotes the collection of values for q being considered. The stopping rule is standard: after each iteration, the algorithm checks the change in the estimate of λ and the norm of the change in the estimate of $\boldsymbol{\rho}$. If these values fall below a certain threshold, the algorithm terminates. The algorithm is fast and typically converges after just a few iterations.

In summary, the procedure described in Algorithm 1 approximates the optimizers of the likelihood (4.8), while also ensuring that the estimate of the spectral density is smooth. (This additional smoothness property is not achieved by simply maximizing the likelihood.)

The estimators for $\hat{\mu}$ and $\hat{\mathbf{R}}$ are consistent, as shown in Section 4.6.7. However, deriving the convergence rate for $\hat{\mu}$ and $\hat{\mathbf{R}}$ is not a trivial task due to their interdependence and requires a separate analysis. Regarding $\hat{\mu}$, based on Lemma 4.1 and Equation (4.3), it is evident that $\mathbf{R}$ has a scaling effect on the smoother: as p grows $\mathbf{S}_{\mathbf{I},\lambda/\delta} \leq \mathbf{S}_{\mathbf{R},\lambda} \leq \mathbf{S}_{\mathbf{I},\delta\lambda}$ for any $\mathbf{R}$ with eigenvalues on $[\delta, 1/\delta]$, where $\mathbf{A} \leq \mathbf{B}$ means that $\mathbf{B} - \mathbf{A}$ is positive semidefinite. Therefore, the impact of the smoother $\mathbf{S}_{\mathbf{R},\lambda}$ is comparable to that of $\mathbf{S}_{\mathbf{I},\lambda}$ for different choices of λ. It is stressed, however, that the smoother $\mathbf{S}_{\mathbf{R},\lambda}$ is far superior to $\mathbf{S}_{\mathbf{I},\lambda}$ in small sample sizes when dealing with correlated data. The smoother $\mathbf{S}_{\mathbf{I},\lambda}$ is known to underperform, especially in the case of positively correlated data, as it does not account for the correlation structure of the noise. However, by examining the likelihood in (4.8), it becomes evident that the presence of correlation simply results in the presence of the (bounded) ρ_i in the denominators. This suggests that it should be possible to study the asymptotics of the proposed estimator $\hat{\mu}$ in a similar manner to the case without correlation, as in Serra et al. (2017).

The asymptotic behavior of the estimator $\hat{q}$ is expected to be similar to the case without correlation in the data, as discussed in Serra et al. (2017). Regardless, the estimator is expected to perform well even if the value of q is specified by the user rather than selected from the data. A common choice is $q = 2$. The estimator for σ^2 should also be consistent.

In conclusion, the presence of short-range dependent noise is close enough to the independent noise case to result in the same large sample behavior for estimators. However, the small sample behavior of estimators that neglect correlation can be significantly improved by employing the proposed approach.

4.2.3 Confidence sets

Once the estimators $\hat{\lambda}$, $\hat{\mathbf{R}}$, $\hat{q}$, and $\hat{\sigma}^2$ for λ, $\mathbf{R}$, q, and σ^2 are obtained, they can be inserted into the marginal posterior for $\boldsymbol{\mu}$ to get the so called empirical, marginal posterior for $\boldsymbol{\mu}$:

$$\begin{aligned}\hat{\Pi}^{\boldsymbol{\mu}}(\,\cdot \mid \mathbf{Y}) &= \Pi^{\boldsymbol{\mu}}(\,\cdot \mid \mathbf{Y}, \hat{\sigma}^2, \hat{\lambda}, \hat{q}, \hat{\mathbf{R}}) \\ &= \Pi^{\boldsymbol{\mu}}(\,\cdot \mid \mathbf{Y}, \sigma^2, \lambda, q, \mathbf{R})\big|_{(\lambda,q,\mathbf{R},\sigma^2)=(\hat{\lambda},\hat{q},\hat{\mathbf{R}},\hat{\sigma}^2)}.\end{aligned} \tag{4.9}$$

Given $\mathbf{Y}$, this is just a multivariate $t_{p+1}(\hat{\boldsymbol{\mu}}, \hat{\sigma}^2\hat{\mathbf{S}}\hat{\mathbf{R}})$ distribution. Here, $\hat{\boldsymbol{\mu}}$ is the spline estimate, $\hat{\mathbf{S}}$ is the smoother matrix $\mathbf{S}$ with $(\hat{\lambda}, \hat{q}, \hat{\mathbf{R}})$ plugged in for $(\lambda, q, \mathbf{R})$, and $\hat{\sigma}^2$ is the estimate of the variance. From this, one can easily construct a credible set for the regression function of the form

$$\hat{C}_p(L) = \left\{\boldsymbol{\mu} : \|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}\|_2^2 \le L\,\hat{\sigma}^2\, s_p(\hat{\lambda}, \hat{q}, \hat{\mathbf{R}})\right\}, \qquad L \ge 1, \tag{4.10}$$

which (for an appropriate, known, sequence s_p) satisfies

$$\hat{\Pi}\{\hat{C}_p(L) \mid \mathbf{Y}\} \geq 1 - \alpha, \qquad L \geq 1,$$

and is therefore a credible set – a small, high probability region of the empirical posterior. See Serra et al. (2017) for a more detailed outline of the construction. In Theorem 3 of Serra et al. (2017), it is shown that when $\mathbf{R} = \mathbf{I}_p$, the Bayesian credible set has two important frequentist properties if L is chosen appropriately large. First, the set contains the true underlying regression function with probability converging to one, uniformly over a large subset of functions. Second, with probability converging to one, the (random) radius of $\hat{C}_p(L)$ is of the order of the minimax risk corresponding to the smoothness class to which the regression function belongs.

As discussed in the previous section, the presence of correlation should have a relatively simple scaling effect on the smoothing parameter. Consequently, the theoretical results of Serra et al. (2017) can potentially be extended to cover the correlated noise case as well by selecting larger values for the multiplier L. However, this issue will be studied in more generality elsewhere. In this paper, the focus is instead on the implementation and numerical aspects of the procedure. See for instance Sniekers et al. (2015); Serra et al. (2017); Rousseau et al. (2020); Yoo et al. (2016) for more details on the use of credible sets as confidence sets, albeit for independent noise.

4.3 Numerical simulations

In this section, we investigate the small sample performance of the proposed estimation procedure and compare it to several alternatives. The simulation setup for data coming from model (4.1) is as follows. Two regression functions are considered

$$\mu_1(t) = \sum_{i=3}^{p} \psi_{3,i}(t)\{\pi(i-1)\}^{-3}\cos(2i),$$

$$\mu_2(t) = 2\sin(4\pi t),$$

where $p = 800$, $t \in [0,1]$ and $\psi_{3,i}$ is the i-th Demmler-Reinsch basis of $\mathcal{W}_3$ given explicitly in Section 4.6.1. Both functions are then scaled so that the standard deviation is one. The standard deviation of the residuals is taken to be $\sigma = 0.33$ to imply a medium signal-to-noise ratio of three. All reported results are based on the Monte

Carlo sample with size $M = 500$. Five types of the residual processes are considered: an AR(1) process with the parameters $\phi = 0.5$ and $\phi = 0.9$, an ARMA$(2, 2)$ process with $\phi = (0.7, -0.4)$ and $\theta = (-0.2, 0.2)$ and two zero-mean Gaussian process with the correlation matrices $\mathbf{R}_1$ and $\mathbf{R}_2$ with (i, j) entries given by $\cos(6.5j)\exp(-|i-j|/10)$ and $\sin(1.5j)\exp(-|i-j|/10)/(1.5j)$, respectively. The error processes are assumed to follow a Gaussian distribution according to the model (4.1). However, we also experimented with other error processes, such as gamma and uniform, and found that there were nearly no differences in performance.

The plots in the top row of Figure 6 show both regression functions. The middle and bottom rows show the first 100 elements of the first row of $\mathbf{R}_1$ (middle, left) and $\mathbf{R}_2$ (bottom, left) and the corresponding spectral densities (middle and bottom right).

Before summarizing the simulation results, we demonstrate how the proposed method works in practice. The data are simulated as described above with the regression function $\mu_1 \in \mathcal{W}_3$ and ARMA$(2, 2)$ dependence in the residuals. After Algorithm 1 has converged, one checks the estimate for q. The top left plot of Figure 7 shows the estimating equation $T_q(\hat{\lambda}, q, \hat{\mathbf{R}})$ which has a zero very close to $q = 3$, as it should for $\mu_1 \in \mathcal{W}_3$. Next, the corresponding estimate for μ is obtained; it is shown in the top right plot of Figure 7, together with the data (black dots) and 95% confidence bands (grey area) constructed as described in Section 4.2.3 ($L = \log(p)$ was used). The estimate for $R_{1,j}$, $j = 1, \ldots, 50$, is shown in the bottom left plot of the same figure in black, which nearly coincides with the true autocorrelations, which are shown in grey. This estimate was reconstructed by the discrete Fourier transform of the estimate of the spectral density, shown in black in the bottom right plot of Figure 7. The true spectral density is shown in the same plot in grey. Hence, the proposed method allows for simultaneous, fully automatic and nonparametric estimation of the mean function μ (with data driven q, λ) and the correlation matrix $\mathbf{R}$.

No other fully nonparametric method for the joint estimation of mean and autocovariance function could be found in the literature. However, several approaches that can be at least partially compared to the proposed method were considered. Among the approaches that make a parametric assumption on $\mathbf{R}$, the well-established method based on splines implemented in the statistical software R, specifically the `gamm` function in the *mgcv* package, was considered (Wood, 2017). This estimation procedure, referred to as GAM, was applied to the processes with a parametric structure of $\{\epsilon(t_i)\}_{i=1}^{p}$ only.

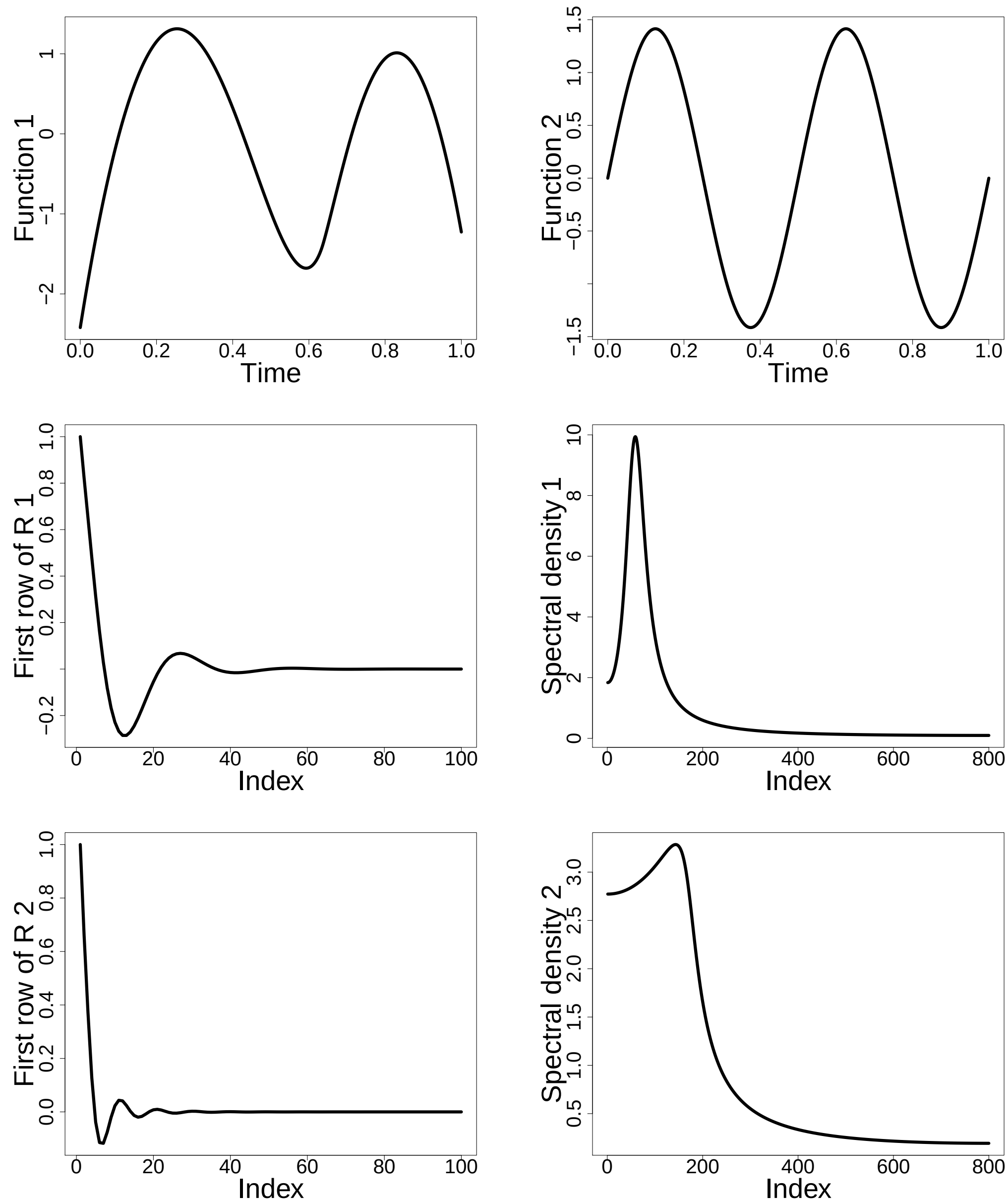

Figure 6: Regressions functions μ_1 (top left), μ_2 (top right), first 100 values of the first row of $\mathbf{R}_1$ (middle left) with the corresponding spectral density (middle right) and of the first row of $\mathbf{R}_2$ (bottom left) with the corresponding spectral density.

The correlation structure parameters in `gamm` were set to the true values, although these are not available in practice.

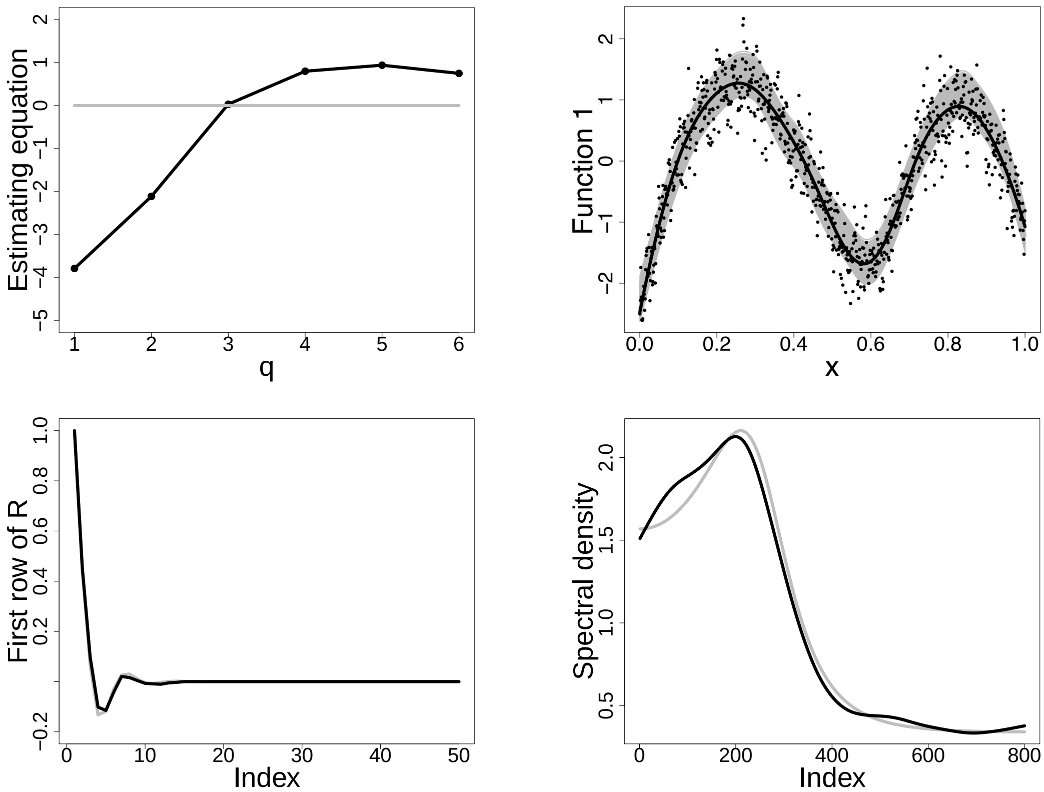

Figure 7: Estimating equation T_q in the top left and the corresponding estimator of μ_1 (black line) together with the data (black points) and point-wise 95% confidence intervals (grey area) in the top right. First 50 elements of the first row of the correlation matrix in grey with its estimator in black (bottom left) and the true spectral density in grey with its estimator in black (bottom right).

Among the approaches that do not make any parametric assumption on $\mathbf{R}$, the method proposed by Herrmann et al. (1992), hereafter referred to as HER, was considered. This kernel-based method uses sample autocorrelation estimators to improve bandwidth selection and is developed under the assumption of m-dependence in the residuals. However, the authors claim that the method works well for the estimation of μ if the residuals satisfy "some mixing conditions". Generally, this approach focuses on the estimation of μ, and the quality of the estimator for r is not discussed.

Finally, a comparison is made between the proposed estimator for the autocovariance function and the standard banded nonparametric estimator, hereafter referred to as

BAND. Since this approach requires a known mean, the true mean μ was plugged into the estimator, resulting in the calculation of

$$\hat{R}_{i,j}(b) = \hat{r}_{|i-j|}\mathbb{1}(|i-j| \leq b),\ i,j = 1,\ldots,p,$$
$$\hat{r}_k = \frac{1}{p}\sum_{i=1}^{p-k}\{Y(t_i) - \mu(t_i)\}\{Y(t_{i+k}) - \mu(t_{i+k})\},\ k = 0,1,2,\ldots,$$

where b is the banding parameter that needs to be chosen and $\mathbb{1}$ is the indicator function.

All these methods require the selection of certain parameters. In the `gamm` function, a low-rank splines with a number of knots equal to $p/4$ was used, a B-spline basis of degree 3, a penalization order $q = 2$, and the correlation structure was specified according to the true dependence structure in the residuals for the first four residual processes.

For the method of Herrmann et al. (1992), the parameter m is set according to method (i) described in their paper (p. 787). Specifically, m was chosen to be the largest integer such that $\hat{h}_m \geq 6/5\ \hat{h}_{m-1}$ and $m \leq 0.2p^{1/2}$, where $\hat{h}_m$ is a selected bandwidth with the parameter m. The parameter m is linked to the assumption of m-dependence in the residuals. It was observed in the experiments that the influence of m on the estimator of μ is not very pronounced, but it does strongly affect the estimator of the autocovariance function. There is no simple data-driven approach to choose m such that both mean and autocovariance estimators are optimal in some sense. In the proposed implementation, the function `glkerns` of the *lokern* package by Eva Herrmann was used for the estimation of μ with the second-order kernel, and for the estimation of μ'' with the fourth-order kernel.

Since there is no fully data-driven approach for selecting the banding parameter b for nonparametric covariance matrix estimation, an oracle bandwidth was used instead. The chosen bandwith minimizes the empirical version of $\mathbb{E}\|\hat{\mathbf{R}}(b) - \mathbf{R}\|_\infty$, where $\|A\|_\infty$ denotes the the maximum absolute row sum of matrix A. The calculation of the expectation is based on a Monte Carlo sample of size 500.

Even though the proposed approach is adaptive and allows for estimating $\beta = 2q - 1$ from the data, q was set to 2 in order to enable a comparison with the other methods for estimating the mean, which should have the same convergence rate. All other parameters are estimated from the data. The proposed method is referred to as BAS.

The results are summarized in Table 12. For all dependence structures and all functions, the metrics

$$A(\hat{\mu}_j) = \frac{1}{Mp}\sum_{k=1}^{p}\sum_{i=1}^{M}\{\mu_j(x_k) - \hat{\mu}_{j,i}(x_k)\}^2, \quad j = 1,2,$$
$$A(\hat{\mathbf{R}}_j) = \frac{1}{Mp}\sum_{k=1}^{p}\sum_{i=1}^{M}\{\mathbf{R}_j(k) - \hat{\mathbf{R}}_{j,i}(k)\}^2, \quad j = 1,\ldots,5,$$

were used. Here, $\hat{\mu}_{j,i}$ denotes an estimator of μ_j, $j = 1,2$, in the i-th Monte Carlo run with the residuals following one of the five processes. Similarly, $\mathbf{R}_j(k)$ denotes the k-th entry of the first row of one of the $j = 1,\ldots,5$ true residual correlation matrices and $\hat{\mathbf{R}}_{j,i}(k)$ is its estimator in the i-th Monte Carlo run.

$\boldsymbol{\mu}_1$

	$A(\hat{\mu}_1)$			$A(\hat{\mathbf{R}})$			
Correlation	BAS	GAM	HER	BAS	GAM	HER	BAND
AR1(0.5)	6.992	6.866	6.433	0.019	0.004	0.017	0.012
AR1(0.9)	107.605	96.652	300.940	0.830	0.140	2.934	0.293
ARMA(2,2)	4.163	4.012	3.797	0.015	0.011	0.031	0.013
GP1	4.161	–	3.727	0.200	–	0.789	0.124
GP2	5.495	–	4.771	0.047	–	0.087	0.048

$\boldsymbol{\mu}_2$

	$A(\hat{\mu}_2)$			$A(\hat{\mathbf{R}})$			
Correlation	BAS	GAM	HER	BAS	GAM	HER	BAND
AR1(0.5)	7.207	7.028	6.114	0.022	0.005	0.016	0.012
AR1(0.9)	121.442	113.355	308.155	0.848	0.163	2.885	0.271
ARMA(2,2)	4.417	4.236	3.482	0.016	0.012	0.033	0.014
GP1	4.449	–	3.603	0.197	–	0.786	0.117
GP2	5.769	–	4.541	0.050	–	0.090	0.046

Table 12: Simulation results. Values of $A(\hat{\mu}_j)$, $j = 1,2$, and $A(\hat{\mathbf{R}})$ were multiplied by 10^3.

The method HER performs best in the estimation of the mean, except for the case with an AR(1) error process with the parameter 0.9, where this approach fails completely. However, in contrast, HER performs the worst in the estimation of the autocovariances, which is likely due to the ad-hoc choice of the tuning parameter m. For parametric structures of the error process, the method GAM performs very similar to BAS in the estimation of the mean and is better in the estimation of the autocovariances. This is not surprising, as the proposed nonparametric method is expected to have a slower convergence rate than the parametric one. However, in the simulations, the correct parametric model specification was used in the `gamm` function, which is not available in practice. Finally, BAND's performance in nonparametric estimation of autocovariances is very similar to BAS, even though BAND uses the known mean and oracle banding parameter, both of which are not available in practice. All together, the proposed fully nonparametric and data-driven method is competitive with the considered alternatives, most of which rely on oracle choices for their parameters.

4.4 Real dataset

In this section, we consider data from the load forecasting track of the Global Energy Forecasting Competition 2012 (`http://www.drhongtao.com/gefcom/2012`). The data consists of hourly loads of a US facility at 20 zones, spanning from the 1st hour of January 1st, 2004 to the 6th hour of June 30th, 2008. The goal of the competition was to make a one-week out-of-sample forecast for each of the 20 time series, as well as backcast certain missing values within the observational period. However, in this study, our focus is not on forecasting the data but rather on understanding their structure. The mean over all 20 zones over the whole time period – all together $p = 1650$ observations – is used, justifying the use of a Gaussian model. Missing values were imputed using R package `Hmisc`. It is worth noting that omitting these missing values led to the same estimators and yielded the same conclusions.

Fitting the data using the proposed method yields a positive and increasing function in the estimating equation for q. This implies that the mean function has either one continuous derivative or is even less smooth. Consequently, a fit with $q = 1$ is considered more appropriate than a fit with $q = 2$. This agrees with the nature of the data which is prone to display peak loads. In the literature, similar data are treated, e.g., by a mixture of smoothing splines and wavelets, where the later pick up the peaks (Amato et al., 2017). The estimator of the mean with $q = 1$ is shown in the top plot of Figure 8 as a black line. The corresponding autocorrelations are shown by solid black line in the bottom left plot.

When setting $q = 1$ in GAM with an AR(2) process for the residuals, the resulting estimator of the mean is very similar to the one obtained with BAS. However, it exhibits slightly less pronounced peaks, although this distinction may not be clearly visible

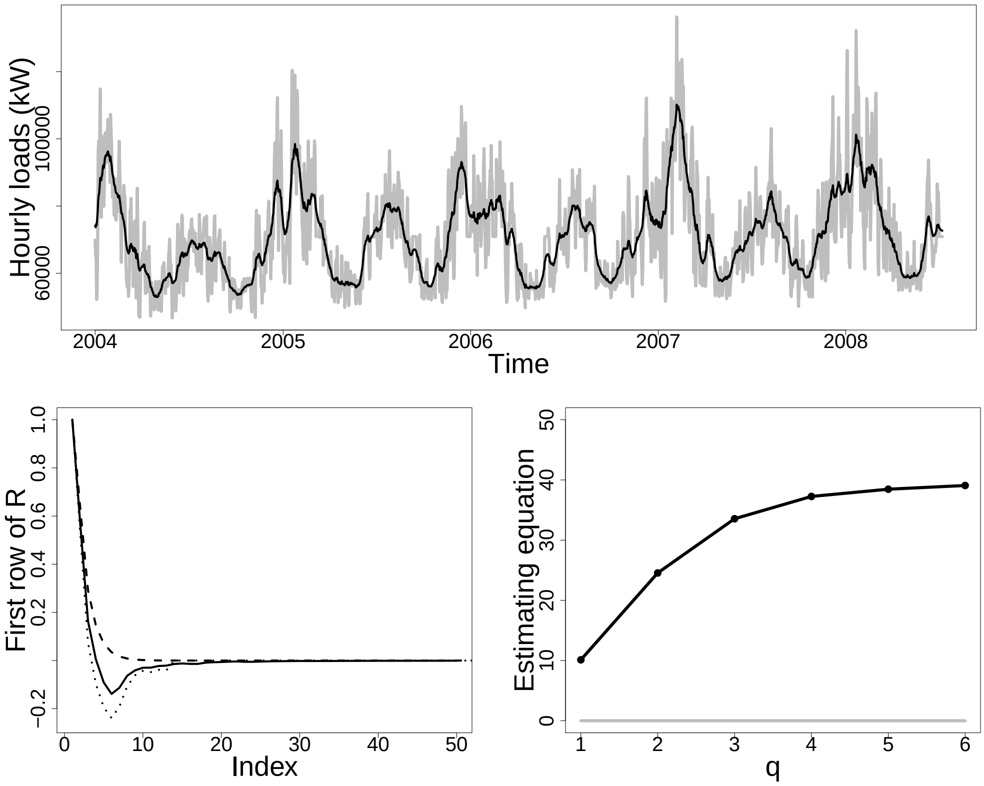

Figure 8: Estimators of the mean (top), first row of the correlation matrix (first 50 values, bottom left) and the estimating equation for the smoothness degree q (bottom right) of the hourly loads of a US facility. Data are shown in grey, BAS estimators are the black solid lines, GAM estimators are dashed lines, BAND estimator for the covariance is the dotted line.

in the plot. The corresponding autocorrelation estimator is shown as a dashed line in the bottom left plot of Figure 8. Assuming an ARMA$(2,2)$ process leads to nearly the same estimator. Since the HER method is defined only for even order kernels, the fit that would be comparable with the fit by BAS with $q = 1$ cannot be obtained. Subtracting the estimated mean from the data a banded nonparametric estimator BAND was employed for the covariance with band parameter $b = 12$. This estimator, shown as a dotted line in the bottom left plot, is closer to the BAS estimator of the covariance.

Estimation with BAS, setting $q = 2$, yields estimates of the mean that are very similar to the ones obtained by GAM, where the correlation structure is set to an AR(1) process. This is shown in the right plot of Figure 5. The corresponding autocorrelation estimators also exhibit reasonable similarity. However, the residual analysis suggests that the fit with $q = 1$ is more appropriate for this dataset.

4.5 Conclusions

Correlation is ubiquitous in various applications, especially when data is collected sequentially over time. Ignoring correlation can have severe consequences for inference. Covariance in data is typically addressed either by assuming a parametric dependence structure or by introducing tuning parameters that need to be set heuristically. To overcome these limitations, a fully automatic, nonparametric method is proposed for estimating both the mean function and the autocovariance function. The method also provides credible sets for the mean function, that quantify the uncertainty of the estimator. The choice of the splines order for the estimator is determined entirely from the data, making it fully data-driven. The approach is implemented in the R package *eBsc*, available from the authors, and delivers results in a quick and numerically stable way.

Deriving the convergence rate for the proposed estimators turned out to be a non-trivial task due to the interdependence of the estimators and will be performed in a separate work. Nevertheless, for short range dependent noise models as covered here, the results, both in terms of rates and uncertainty quantification are expected to be similar to the independent noise case. Given the lack of fully nonparametric and fully data-driven methods available in the literature, the numerical simulations compare the proposed approach with existing methods that rely on oracle knowledge of parameters. Even so, the numerical simulations suggest that the proposed method is competitive with these alternatives, demonstrating its effectiveness in practice.

4.6 Proofs and auxiliary results

This section contains a number of technical results that are used throughout this paper.

4.6.1 Demmler-Reinsch basis

Let $\{\psi_i\}_{i=1}^{\infty}$ denote the Demmler-Reinsch basis of $\mathcal{W}_\beta(M)$, such that

$$\nu_{\beta,i} \int_0^1 \psi_{\beta,i}(x)\psi_{\beta,j}(x)\,\mathrm{d}x = \nu_{\beta,i}\delta_{ij} = \int_0^1 \psi_{\beta,i}^{(\beta)}(x)\psi_{\beta,j}^{(\beta)}(x)\,\mathrm{d}x.$$

For a precise definition of the space $\mathcal{W}_\beta(M)$, $\beta \in \mathbb{N}$, see Serra et al. (2017). Rosales Marticorena (2016) found explicit expressions for $\psi_{\beta,i}$ and $\nu_{\beta,i}$ as a solution to

$$(-1)^q \psi_{\beta,i}^{(2\beta)} = \nu_{\beta,i}\psi_{\beta,i}, \tag{4.11}$$

$$\psi_{\beta,i}^{(l)}(0) = \psi_{\beta,i}^{(l)}(1) = 0, \quad l = \beta, \beta+1, \ldots, 2\beta-1. \tag{4.12}$$

In particular, $\nu_{\beta,1} = \cdots = \nu_{\beta,\beta} = 0$ and

$$\begin{aligned} \nu_{\beta,i} &= \left\{\pi\left(i - \frac{\beta+1}{2}\right)\right\}^{2\beta}, \quad i = \beta+1, \beta+2, \ldots, \\ \psi_{\beta,i}(x) &= \sqrt{2}\left[\cos\left\{\pi\left(i - \frac{\beta+1}{2}\right)x + \pi\,\frac{\beta-1}{4}\right\} + T_i(x)\right] \end{aligned} \tag{4.13}$$

where

$$\begin{aligned} T_i(x) = \sum_{a_j \in S(\beta)} r_j \Bigg[&\exp\left\{-a_j\pi\left(i - \frac{\beta+1}{2}\right)x\right\} \\ &+(-1)^{i+1}\exp\left\{-a_j\pi\left(i - \frac{\beta+1}{2}\right)(1-x)\right\}\Bigg] \end{aligned}$$

for $S(\beta) = \cup_j \left\{(-1)^{j/(2\beta)}, \overline{(-1)^{j/(2\beta)}}\right\}$, with $0 \leq j \leq \beta - 2$ taking odd values for β odd and even values for β even. Constants r_j are known and depend on β only. Note that $T_i(x)$ vanish exponentially fast away from the boundaries.

The Demmler-Reinsch basis of the natural spline space of degree $2q-1$ with knots at observations $\mathcal{S}_{2q-1}(\mathbf{x})$ is uniquely defined via

$$\eta_{q,i}\sum_{k=1}^{p}\phi_{q,i}(x_k)\phi_{q,j}(x_k) = \eta_{q,i}\delta_{ij} = \int_0^1 \phi_{q,i}^{(q)}(x)\phi_{q,j}^{(q)}(x)\mathrm{d}x,$$

and $\boldsymbol{\Phi}_q = \boldsymbol{\Phi}_q(\mathbf{x}) = [\phi_{q,1}(\mathbf{x}), \ldots, \phi_{q,p}(\mathbf{x})] = [\phi_{q,j}(x_i)]_{i,j=1}^{p}$ is the corresponding basis matrix.

Utreras Diaz (1980) used the results of Fix (1972) to show that $|p\eta_{q,i} - \nu_{q,i}| = \mathcal{O}(p^{-2})$. From Fix (1972) and Fix (1973) also follows that $\|p^{1/2}\phi_{q,i} - \psi_{q,i}\|_{\mathcal{W}_q} = \mathcal{O}(p^{-1})$, or, equivalently, that $\|\phi_{q,i} - \psi_{q,i}p^{-1/2}\|_{L_2} = \mathcal{O}(p^{-3/2})$.

To apply results by Fix (1972) and Fix (1973) the standard result Lemma 3.2 of Utreras Diaz (1980) is used.

Lemma 4.2 (Lemma 3.2 of (Utreras Diaz, 1980)) *Let $\mu \in W_\beta(M)$, $\beta \geq 2$, then*

$$\left| \frac{1}{p} \sum_{i=1}^{p} \mu(t_i) - \int_0^1 \mu(t) dt \right| \leq \frac{c}{p^2} \|\mu\|_{\mathcal{W}_\beta},$$

where $t_i = (i-1)/(p-1)$, $i = 1, \ldots, p$.

4.6.2 Matrix identities

This section contains some matrix identities. The smoother matrix $\mathbf{S}$ satisfies

$$\mathbf{S} = \mathbf{R}^{1/2} \left\{ \mathbf{I}_p + \lambda \mathbf{R}^{1/2} \boldsymbol{\Phi}_q \operatorname{diag}(p\boldsymbol{\eta}_q) \boldsymbol{\Phi}_q^t \mathbf{R}^{1/2} \right\}^{-1} \mathbf{R}^{-1/2}.$$

From this expression is it clear that $\mathbf{R}^{-1}\mathbf{S} = \mathbf{S}^t\mathbf{R}^{-1}$ so that in particular

$$\mathbf{R}^{-1}(\mathbf{I}_p - \mathbf{S}) = (\mathbf{I}_p - \mathbf{S})^t \mathbf{R}^{-1}.$$

From the definition of $\mathbf{S}$ is is also simple to check the scaling relation

$$\mathbf{R}^{-1}(\mathbf{S}^{-1} - \mathbf{I}_p) = \lambda \boldsymbol{\Phi}_q \operatorname{diag}(p\boldsymbol{\eta}_q) \boldsymbol{\Phi}_q{}^t = \mathbf{S}_{\mathbf{I}}^{-1} - \mathbf{I}_p.$$

The estimating equation for λ is obtained in a straightforward way by noting that since $\partial \mathbf{S}/\partial\lambda = -(\mathbf{I}_p - \mathbf{S})\mathbf{S}/\lambda$, then

$$\frac{\partial \mathbf{R}^{-1}(\mathbf{I}_p - \mathbf{S})}{\partial \lambda} = \frac{1}{\lambda} \mathbf{R}^{-1}(\mathbf{I}_p - \mathbf{S})\mathbf{S},$$

and the estimating equation for q is based on the fact that

$$\begin{aligned}
\frac{\partial \mathbf{R}^{-1}(\mathbf{I}_p - \mathbf{S})}{\partial q} &= \mathbf{R}^{-1}\mathbf{S} \frac{\partial \mathbf{S}^{-1}}{\partial q} \mathbf{S} = \frac{\lambda}{q} \mathbf{R}^{-1} \mathbf{S} \mathbf{R} \boldsymbol{\Phi}_q \operatorname{diag}\left\{ p\boldsymbol{\eta}_q \circ \log(p\boldsymbol{\eta}_q) \right\} \boldsymbol{\Phi}_q^t \mathbf{S} \\
&= \frac{\lambda}{q} \mathbf{S}^t \boldsymbol{\Phi}_q \operatorname{diag}\left\{ p\boldsymbol{\eta}_q \circ \log(p\boldsymbol{\eta}_q) \right\} \boldsymbol{\Phi}_q^t \mathbf{S} \\
&= \frac{1}{q} \mathbf{R}^{-1}(\mathbf{I} - \mathbf{S}) \boldsymbol{\Phi}_q \operatorname{diag}\left\{ \log(p\boldsymbol{\eta}_q) \right\} \boldsymbol{\Phi}_q^t \mathbf{S},
\end{aligned}$$

where $\circ$ denotes the Hadamard product. For the derivation of the estimating equations for the ρ_i note that

$$\frac{\partial \mathbf{S}}{\partial \rho_i} = -\mathbf{S}\frac{\partial \mathbf{S}^{-1}}{\partial \rho_i}\mathbf{S} = -\mathbf{S}\frac{\partial \mathbf{R}}{\partial \rho_i}(\mathbf{S}_{\mathbf{I}_p}^{-1} - \mathbf{I}_p)\mathbf{S} = -\mathbf{S}\frac{\partial \mathbf{R}}{\partial \rho_i}\mathbf{R}^{-1}(\mathbf{I}_p - \mathbf{S}),$$

such that, combining the above, and using the chain rule

$$\begin{aligned}\frac{\partial \mathbf{R}^{-1}(\mathbf{I}_p - \mathbf{S})}{\partial \rho_i} &= -\mathbf{R}^{-1}\frac{\partial \mathbf{R}}{\partial \rho_i}\mathbf{R}^{-1}(\mathbf{I}_p - \mathbf{S}) + \mathbf{R}^{-1}\frac{\partial \mathbf{S}}{\partial \rho_i} \\ = -\mathbf{R}^{-1}(\mathbf{I}_p - \mathbf{S})\frac{\partial \mathbf{R}}{\partial \rho_i}\mathbf{R}^{-1}(\mathbf{I}_p - \mathbf{S}) &= -(\mathbf{I}_p - \mathbf{S})^t\mathbf{R}^{-1}\frac{\partial \mathbf{R}}{\partial \rho_i}\mathbf{R}^{-1}(\mathbf{I}_p - \mathbf{S}).\end{aligned}$$

4.6.3 R and its spectral density

Let $\mathbf{R}$ be a Toeplitz matrix such that $\mathbf{R}_{i,j} = r_{i-j}$ for some sequence $(r_i)_{i\in\mathbb{Z}}$ which satisfies $\sum_{k=-\infty}^{\infty} |r_k|^2 < \infty$. Denote the Fourier spectrum of $\mathbf{R}$ as

$$\rho(x) = \sum_{k=-\infty}^{\infty} r_k \exp(\mathrm{i}kx), \quad \text{so that} \quad r_k = \frac{1}{2\pi}\int_{-\pi}^{\pi} \rho(x)\exp(-\mathrm{i}kx)\,\mathrm{d}x,$$

where i denotes the imaginary unit and $x \in [-\pi, \pi]$. If $\mathbf{R}$ is the correlation matrix of a stationary process, then additionally, $r_{i-j} = r_{j-i}$ so that

$$\begin{aligned}\rho(x) &= 1 + 2\sum_{k=1}^{\infty} r_k\cos(kx) \quad \text{and} \\ r_k &= \frac{1}{2\pi}\int_{-\pi}^{\pi}\cos(kx)\rho(x)\,\mathrm{d}x = \int_0^1 \cos(k\pi x)\rho(\pi x)\,\mathrm{d}x.\end{aligned}$$

In practice r_k is approximated via

$$r_k = \frac{1}{p}\sum_{i=1}^{p}\cos\left(k\pi\frac{i-1}{p-1}\right)\rho_i + \mathcal{O}(p^{-1}), \quad \text{where } \rho_i = \rho\left(\pi\frac{i-1}{p-1}\right).$$

4.6.4 Derivation of the posterior and marginal posterior

Consider $\mathbf{t}, \lambda, q, \mathbf{R}$ fixed throughout. Endow $(\boldsymbol{\mu}, \sigma^2)$ with a prior specified by endowing σ^2 with and inverse-gamma prior with parameters a, b, and endowing $\boldsymbol{\mu} \mid \sigma^2$ with a prior with density (w.r.t. some appropriate dominating measure) as specified in (4.4). Using the fact that $\mathbf{S}^{-1} - \mathbf{I}_p$ has rank $p-q$, the prior density is thus proportional to

$$(\sigma^2)^{-\frac{p-q}{2}-a-1} \left|\mathbf{R}^{-1}(\mathbf{S}^{-1}-\mathbf{I}_p)\right|_+^{1/2} \exp\left[-\frac{1}{2\sigma^2}\left\{\boldsymbol{\mu}^t\mathbf{R}^{-1}(\mathbf{S}^{-1}-\mathbf{I}_p)\boldsymbol{\mu}+2b\right\}\right].$$

Multiplying this by the Gaussian likelihood and completing the square shows that the posterior density of $(\boldsymbol{\mu}, \sigma^2)$ is proportional to

$$(\sigma^2)^{-\frac{2p-q}{2}-a-1} |\mathbf{SR}|^{-1/2} \exp\left(-\frac{Q}{2\sigma^2}\right), \tag{4.14}$$

where Q in the exponent is

$$Q = (\boldsymbol{\mu}-\mathbf{SY})^t(\mathbf{SR})^{-1}(\boldsymbol{\mu}-\mathbf{SY}) + 2b + \mathbf{Y}^t\mathbf{R}^{-1}(\mathbf{I}_p-\mathbf{S})\mathbf{Y}^t.$$

Integrating out $\boldsymbol{\mu}$, the marginal density of the posterior distribution on σ^2 is recognized as an inverse-gamma distribution with shape parameter $(p-q+2a)/2$ and scale parameter $\{\mathbf{Y}^t\mathbf{R}^{-1}(\mathbf{I}_p-\mathbf{S})\mathbf{Y}+2b\}/2$. Integrating out σ^2 instead leads to a marginal posterior density for $\boldsymbol{\mu}$ which is clearly a multivariate t-distribution with $p+1$ degrees of freedom, expectation $\hat{\boldsymbol{\mu}} = \mathbf{SY}$, and scale $\hat{\sigma}^2\mathbf{SR}$, where $\hat{\sigma}^2$ is given by (4.5).

As for the marginal distribution of $\mathbf{Y}$ under the prior, this is obtained by integrating out both $\boldsymbol{\mu}$ and σ^2 from (4.14). For instance, the marginal posterior for σ^2 has a density proportional to

$$(\sigma^2)^{-\frac{p-q}{2}-a-1} \exp\left[-\frac{1}{2\sigma^2}\left\{2b+\mathbf{Y}^t\mathbf{R}^{-1}(\mathbf{I}_p-\mathbf{S})\mathbf{Y}^t\right\}\right],$$

which can be easily integrated with respect to σ^2 leading to the marginal likelihood for $\mathbf{Y}$ proportional to the one specified in (4.6).

4.6.5 Proof of Lemma 4.1

(i) The spectral density function ρ corresponding to the correlation matrix $\mathbf{R}$ is a non-negative, even 2π-periodic function with Fourier series $\rho(\pi x) = 1 + 2\sum_{\tau=1}^{\infty} r_\tau \cos(\pi x \tau)$ for $x \in [-1, 1]$. By assumption, $\rho^{(\gamma)}$ is α-Hölder continuous and ρ is bounded by some constant $M > 0$. In the following, we assume that p is even. The case of odd p is discussed at the end. Let $t_i = (i-1)/(p-1)$ and $t_{j,q} = \{j - (q+1)/2\}/(p-1)$. Then, we get for $i, j = q+1, ..., p$

$$
\begin{aligned}
&\left\{\boldsymbol{\Phi}^t \mathbf{R} \boldsymbol{\Phi}\right\}_{i,j} \\
&= \sum_{l=1}^{p} \phi_{q,i}(t_l) \sum_{k=1}^{p} \phi_{q,j}(t_k) r_{|l-k|} \pm \sqrt{\frac{2}{p}} \sum_{l=1}^{p} \phi_{q,i}(t_l) \sum_{k=1}^{p} \cos\{\pi(k-1)t_{j,q} + \pi(q-1)/4\} r_{|l-k|} \\
&= \sqrt{\frac{2}{p}} \sum_{l=1}^{p} \phi_{q,i}(t_l) \sum_{k=1}^{p} \cos\{\pi(k-1)t_{j,q} + \pi(q-1)/4\} r_{|l-k|} \\
&\quad + \sum_{l=1}^{p} \phi_{q,i}(t_l) \sum_{k=1}^{p} \left[\phi_{q,j}(t_k) - \sqrt{\frac{2}{p}} \cos\{\pi(k-1)t_{j,q} + \pi(q-1)/4\}\right] r_{|l-k|}.
\end{aligned}
$$

From the identity $\cos(a+b) = \cos(a)\cos(b) - \sin(a)\sin(b)$ follows

$$
\begin{aligned}
&\sum_{k=1}^{p} \cos\{\pi(k-1)t_{j,q} + \pi(q-1)/4\} r_{|l-k|} \\
&= \cos\{\pi(l-1)t_{j,q} + \pi(q-1)/4\} \sum_{k=1}^{p} \cos\{\pi(k-l)t_{j,q}\} r_{|l-k|} \\
&\quad - \sin\{\pi(l-1)t_{j,q} + \pi(q-1)/4\} \sum_{k=1}^{p} \sin\{\pi(k-l)t_{j,q}\} r_{|l-k|}.
\end{aligned}
$$

Using that cosine is an even function and $\rho(\pi t_{j,q}) = S_{p/2}(\rho)(t_{j,q}) + \sum_{|s|>p/2}^{\infty} \cos\{\pi s t_{j,q}\} r_s$, where $S_p(\rho)(x) = 1 + 2\sum_{\tau=1}^{p} r_\tau \cos(\pi x \tau)$ is the p-th partial Fourier sum of ρ, we get

$$
\begin{aligned}
&\sum_{k=1}^{p} \cos\{\pi(k-l)t_{j,q}\} r_{|l-k|} \qquad (4.15)\\
&= \sum_{k=1}^{l-1} \cos\{\pi(k-l)t_{j,q}\} r_{|l-k|} + \sum_{k=l}^{p} \cos\{\pi(k-l)t_{j,q}\} r_{|l-k|} \\
&= \sum_{k-l=1-l}^{l-1-l} \cos\{\pi(k-l)t_{j,q}\} r_{|l-k|} + \sum_{k-l=l-l}^{p-l} \cos\{\pi(k-l)t_{j,q}\} r_{|l-k|} \\
&= \sum_{s=1}^{l-1} \cos\{-\pi s t_{j,q}\} r_s + \sum_{s=0}^{p-l} \cos\{\pi s t_{j,q}\} r_s
\end{aligned}
$$

$$= \begin{cases} \sum_{s=1}^{l-1} \cos\{\pi s t_{j,q}\} r_s + \sum_{s=0}^{p/2} \cos\{\pi s t_{j,q}\} r_s \\ + \sum_{s=p/2+1}^{p-l} \cos\{\pi s t_{j,q}\} r_s \pm \sum_{s=l}^{p/2} \cos\{\pi s t_{j,q}\} r_s, & \text{if } l \le \frac{p}{2} \\ \sum_{s=1}^{p/2} \cos\{\pi s t_{j,q}\} r_s + \sum_{s=p/2+1}^{l-1} \cos\{\pi s t_{j,q}\} r_s \\ + \sum_{s=0}^{p-l} \cos\{\pi s t_{j,q}\} r_s \pm \sum_{s=p-l+1}^{p/2} \cos\{\pi s t_{j,q}\} r_s & \text{if } l > \frac{p}{2} \end{cases}$$

$$= \begin{cases} S_{p/2}(\rho)(t_{j,q}) + \sum_{s=p/2+1}^{p-l} \cos\{\pi s t_{j,q}\} r_s - \sum_{s=l}^{p/2} \cos\{\pi s t_{j,q}\} r_s, & \text{if } l \le \frac{p}{2} \\ S_{p/2}(\rho)(t_{j,q}) + \sum_{s=p/2+1}^{l-1} \cos\{\pi s t_{j,q}\} r_s - \sum_{s=p-l+1}^{p/2} \cos\{\pi s t_{j,q}\} r_s, & \text{if } l > \frac{p}{2} \end{cases}$$

$$= \begin{cases} \rho(\pi t_{j,q}) + \sum_{s=p/2+1}^{p-l} \cos\{\pi s t_{j,q}\} r_s - \sum_{s=l}^{p/2} \cos\{\pi s t_{j,q}\} r_s \\ - \sum_{|s|>p/2}^{\infty} \cos\{\pi s t_{j,q}\} r_s, & \text{if } l \le \frac{p}{2} \\ \rho(\pi t_{j,q}) + \sum_{s=p/2+1}^{l-1} \cos\{\pi s t_{j,q}\} r_s - \sum_{s=p-l+1}^{p/2} \cos\{\pi s t_{j,q}\} r_s \\ - \sum_{|s|>p/2}^{\infty} \cos\{\pi s t_{j,q}\} r_s, & \text{if } l > \frac{p}{2}. \end{cases} \tag{4.16}$$

Note that sums for which the end index is smaller than the starting index, are treated as zero. In the same way and using that sine is an odd function with $\sin(0) = 0$, we get

$$\begin{aligned} \sum_{k=1}^{p} \sin\{\pi(k-l) t_{j,q}\} r_{|l-k|} &= \sum_{s=1}^{l-1} \sin\{-\pi s t_{j,q}\} r_s + \sum_{s=0}^{p-l} \sin\{\pi s t_{j,q}\} r_s \\ &= \begin{cases} \sum_{s=l}^{p-l} \sin\{\pi s t_{j,q}\} r_s, & \text{if } l \le \frac{p}{2} \\ - \sum_{s=p-l+1}^{l-1} \sin\{\pi s t_{j,q}\} r_s, & \text{if } l > \frac{p}{2}. \end{cases} \end{aligned}$$

Together,

$$\sqrt{\frac{2}{p}} \sum_{l=1}^{p} \phi_{q,i}(t_l) \sum_{k=1}^{p} \cos\{\pi(k-1) t_{j,q} + \pi(q-1)/4\} r_{|l-k|} \tag{4.17}$$

$$\begin{aligned} =& \sqrt{\frac{2}{p}} \sum_{l=1}^{p} \phi_{q,i}(t_l) \cos\{\pi(l-1) t_{j,q} + \pi(q-1)/4\} \rho(\pi t_{j,q}) \\ &+ \sqrt{\frac{2}{p}} \sum_{l=1}^{p/2} \phi_{q,i}(t_l) \cos\{\pi(l-1) t_{j,q} + \pi(q-1)/4\} \left[\sum_{s=p/2+1}^{p-l} \cos\{\pi s t_{j,q}\} r_s - \sum_{s=l}^{p/2} \cos\{\pi s t_{j,q}\} r_s \right] \\ &+ \sqrt{\frac{2}{p}} \sum_{l=p/2+1}^{p} \phi_{q,i}(t_l) \cos\{\pi(l-1) t_{j,q} + \pi(q-1)/4\} \left[\sum_{s=p/2+1}^{l-1} \cos\{\pi s t_{j,q}\} r_s - \sum_{s=p-l+1}^{p/2} \cos\{\pi s t_{j,q}\} r_s \right] \end{aligned} \tag{4.18}$$

$$\begin{aligned} &- \sqrt{\frac{2}{p}} \sum_{l=1}^{p} \phi_{q,i}(t_l) \cos\{\pi(l-1) t_{j,q} + \pi(q-1)/4\} \sum_{|s|>p/2}^{\infty} \cos\{\pi s t_{j,q}\} r_s \\ &- \sqrt{\frac{2}{p}} \sum_{l=1}^{p/2} \phi_{q,i}(t_l) \sin\{\pi(l-1) t_{j,q} + \pi(q-1)/4\} \sum_{s=l}^{p-l} \sin\{\pi s t_{j,q}\} r_s \end{aligned}$$

$$+\sqrt{\frac{2}{p}}\sum_{l=p/2+1}^{p}\phi_{q,i}(t_l)\sin\{\pi(l-1)t_{j,q}+\pi(q-1)/4\}\sum_{s=p-l+1}^{l-1}\sin\{\pi s t_{j,q}\}r_s. \tag{4.19}$$

Since $\phi_{q,i}(t_l)$ is an even function for i odd and an odd function for i even, it holds $\phi_{q,i}(t_l)=(-1)^{i-1}\phi_{q,i}(t_{p-l+1})$ for $l=1,\dots,p$. Furthermore, $\cos(x)=(-1)^k\cos(\pi k+x)$ and thus

$$\begin{aligned}
&\cos\left\{\pi t_{j,q}(l-1)+\pi\frac{q-1}{4}\right\}\\
&=(-1)^{j-1}\cos\left\{\pi(j-1)-\pi(j-(q+1)/2)\frac{l-1}{p-1}-\pi\frac{q-1}{4}\right\}\\
&=(-1)^{j-1}\cos\left\{\frac{2\pi(q-1)}{4}+\pi(j-(q+1)/2)-\pi(j-(q+1)/2)\frac{l-1}{p-1}-\pi\frac{q-1}{4}\right\}\\
&=(-1)^{j-1}\cos\left\{\pi(j-(q+1)/2)\frac{(p-1)-(l-1)}{p-1}+\pi\frac{q-1}{4}\right\}\\
&=(-1)^{j-1}\cos\{\pi t_{j,q}(p-l)+\pi(q-1)/4\}.
\end{aligned}$$

Then, the term (4.18) can be written as follows

$$\begin{aligned}
(4.18)&=(-1)^{i-1}(-1)^{j-1}\sqrt{\frac{2}{p}}\sum_{l=p/2+1}^{p}\phi_{q,i}(t_{p-l+1})\cos\{\pi t_{j,q}(p-l)+\pi(q-1)/4\}\\
&\quad\times\left[\sum_{s=p/2+1}^{l-1}\cos\{\pi s t_{j,q}\}r_s-\sum_{s=p-l+1}^{p/2}\cos\{\pi s t_{j,q}\}r_s\right]\\
&=(-1)^{|i-j|}\sqrt{\frac{2}{p}}\sum_{l=1}^{p/2}\phi_{q,i}(t_l)\cos\{\pi t_{j,q}(l-1)+\pi(q-1)/4\}\\
&\quad\times\left[\sum_{s=p/2+1}^{p-l}\cos\{\pi s t_{j,q}\}r_s-\sum_{s=l}^{p/2}\cos\{\pi s t_{j,q}\}r_s\right].
\end{aligned}$$

Since $\sin(x)=(-1)^k\sin(\pi k+x)$, we get by similar calculations as before

$$\begin{aligned}
&\sin\left\{\pi t_{j,q}(l-1)+\pi\frac{q-1}{4}\right\}=-(-1)^{j-1}\sin\left\{\pi(j-1)-\pi\frac{j-(q+1)/2}{p-1}(l-1)-\pi\frac{q-1}{4}\right\}\\
&=-(-1)^{j-1}\sin\{\pi t_{j,q}(p-l)+\pi(q-1)/4\}.
\end{aligned}$$

Analogously, the term (4.19) can be rewritten as follows

$$
\begin{aligned}
&-(-1)^{|i-j|}\sqrt{\frac{2}{p}}\sum_{l=p/2+1}^{p}\phi_{q,i}(t_{p-l+1})\sin\{\pi t_{j,q}(p-l)+\pi(q-1)/4\}\sum_{s=p-l+1}^{l-1}\sin\{\pi s t_{j,q}\}r_s\\
&=-(-1)^{|i-j|}\sqrt{\frac{2}{p}}\sum_{l=1}^{p/2}\phi_{q,i}(t_l)\sin\{\pi t_{j,q}(l-1)+\pi(q-1)/4\}\sum_{s=l}^{p-l}\sin\{\pi s t_{j,q}\}r_s.
\end{aligned}
$$

Thus,

$$
\begin{aligned}
&\sqrt{\frac{2}{p}}\sum_{l=1}^{p}\phi_{q,i}(t_l)\sum_{k=1}^{p}\cos\{\pi(k-1)t_{j,q}+\pi(q-1)/4\}r_{|l-k|}\\
&=\rho(\pi t_{j,q})\sqrt{\frac{2}{p}}\sum_{l=1}^{p}\phi_{q,i}(t_l)\cos\{\pi(l-1)t_{j,q}+\pi(q-1)/4\}\\
&+\{1+(-1)^{|i-j|}\}\sqrt{\frac{2}{p}}\sum_{l=1}^{p/2}\phi_{q,i}(t_l)\cos\{\pi(l-1)t_{j,q}+\pi(q-1)/4\}\left[\sum_{s=p/2+1}^{p-l}\cos(\pi s t_{j,q})r_s-\sum_{s=l}^{p/2}\cos(\pi s t_{j,q})r_s\right]\\
&-\{1+(-1)^{|i-j|}\}\sqrt{\frac{2}{p}}\sum_{l=1}^{p/2}\phi_{q,i}(t_l)\sin\{\pi(l-1)t_{j,q}+\pi(q-1)/4\}\left[\sum_{s=l}^{p/2}\sin(\pi s t_{j,q})r_s+\sum_{s=p/2+1}^{p-l}\sin(\pi s t_{j,q})r_s\right]\\
&-\sqrt{\frac{2}{p}}\sum_{l=1}^{p}\phi_{q,i}(t_l)\cos\{\pi(l-1)t_{j,q}+\pi(q-1)/4\}\sum_{|s|>p/2}^{\infty}\cos(\pi s t_{j,q})r_s\\
&=\rho(\pi t_{j,q})\sqrt{\frac{2}{p}}\sum_{l=1}^{p}\phi_{q,i}(t_l)\cos\{\pi(l-1)t_{q,j}+\pi(q-1)/4\}\\
&-\{1+(-1)^{|i-j|}\}\sqrt{\frac{2}{p}}\sum_{l=1}^{p/2}\phi_{q,i}(t_l)\cos\{\pi(l-1)t_{j,q}+\pi(q-1)/4\}\left[\sum_{s=n-l+1}^{\infty}\cos(\pi s t_{j,q})r_s+\sum_{s=l}^{\infty}\cos(\pi s t_{j,q})r_s\right]\\
&-\{1+(-1)^{|i-j|}\}\sqrt{\frac{2}{p}}\sum_{l=1}^{p/2}\phi_{q,i}(t_l)\sin\{\pi(l-1)t_{j,q}+\pi(q-1)/4\}\left[\sum_{s=l}^{\infty}\sin(\pi s t_{j,q})r_s-\sum_{s=n-l+1}^{\infty}\sin(\pi s t_{j,q})r_s\right],
\end{aligned}
$$

where we used

$$
\begin{aligned}
&-\sqrt{\frac{2}{p}}\sum_{l=1}^{p}\phi_{q,i}(t_l)\cos\{\pi(l-1)t_{j,q}+\pi(q-1)/4\}\sum_{|s|>p/2}^{\infty}\cos\{\pi s t_{j,q}\}r_s\\
&=-\{1+(-1)^{|i-j|}\}\sqrt{\frac{2}{p}}\sum_{l=1}^{p/2}\phi_{q,i}(t_l)\cos\{\pi(l-1)t_{j,q}+\pi(q-1)/4\}\sum_{|s|>p/2}^{\infty}\cos\{\pi s t_{j,q}\}r_s
\end{aligned}
$$

and added

$$
\begin{aligned}
0 &= \sqrt{\frac{2}{p}} \sum_{l=1}^{p} \phi_{q,i}(t_l) \sin\{\pi(l-1)t_{j,q} + \pi(q-1)/4\} \sum_{|s|>p/2}^{\infty} \sin\{\pi s t_{j,q}\} r_s \\
&= \{1+(-1)^{|i-j|}\} \sqrt{\frac{2}{p}} \sum_{l=1}^{p/2} \sin\{\pi(l-1)t_{j,q} + \pi(q-1)/4\} \sum_{|s|>p/2}^{\infty} \sin\{\pi s t_{j,q}\} r_s.
\end{aligned}
$$

Together,

$$
\begin{aligned}
&\{\mathbf{\Phi}^t \mathbf{R} \mathbf{\Phi}\}_{i,j} = \rho(\pi t_{j,q}) \sqrt{\frac{2}{p}} \sum_{l=1}^{p} \phi_{q,i}(t_l) \cos\{\pi(l-1)t_{q,j} + \pi(q-1)/4\} \\
&- \{1+(-1)^{|i-j|}\} \sqrt{\frac{2}{p}} \sum_{l=1}^{p/2} \phi_{q,i}(t_l) \sum_{s=p-l+1}^{\infty} \cos\{\pi(l-1+s)t_{j,q} + \pi(q-1)/4\} r_s \\
&- \{1+(-1)^{|i-j|}\} \sqrt{\frac{2}{p}} \sum_{l=1}^{p/2} \phi_{q,i}(t_l) \sum_{s=l}^{\infty} \cos\{\pi(l-1-s)t_{j,q} + \pi(q-1)/4\} r_s \\
&+ \sum_{l=1}^{p} \phi_{q,i}(t_l) \sum_{k=1}^{p} \left[\phi_{q,j}(t_k) - \sqrt{\frac{2}{p}} \cos\{\pi(k-1)t_{j,q} + \pi(q-1)/4\} \right] r_{|l-k|} \\
&= \rho(\pi t_{j,q}) \delta_{ij} - \{1+(-1)^{|i-j|}\} \sqrt{\frac{2}{p}} \sum_{l=1}^{p/2} \phi_{q,i}(t_l) \sum_{s=p-l+1}^{\infty} \cos\{\pi(l-1+s)t_{j,q} + \pi(q-1)/4\} r_s \\
&- \{1+(-1)^{|i-j|}\} \sqrt{\frac{2}{p}} \sum_{l=1}^{p/2} \phi_{q,i}(t_l) \sum_{s=l}^{\infty} \cos\{\pi(l-1-s)t_{j,q} + \pi(q-1)/4\} r_s \\
&- \rho(\pi t_{j,q}) \sum_{l=1}^{p} \phi_{q,i}(t_l) \left[\phi_{q,j}(t_l) - \sqrt{\frac{2}{p}} \cos\{\pi(l-1)t_{q,j} + \pi(q-1)/4\} \right] \\
&+ \sum_{l=1}^{p} \phi_{q,i}(t_l) \sum_{k=1}^{p} \left[\phi_{q,j}(t_k) - \sqrt{\frac{2}{p}} \cos\{\pi(k-1)t_{j,q} + \pi(q-1)/4\} \right] r_{|l-k|}.
\end{aligned} \tag{4.20}
$$

Case $|i-j|$ is odd

The four last terms in Equation (4.20) vanish for $|i-j|$ odd. For the last two terms this follows since $\phi_{q,j}(x)$ and $\cos\{\pi x(j-(q+1)/2) + \pi(q-1)/4\}$ are even/odd functions at $x = 1/2$ for j odd/even and they have the same number of sign changes. $|i-j|$ odd implies that either $\phi_{q,i}(x)$ even and $[\phi_{q,j}(x) - \sqrt{\frac{2}{p}} \cos\{\pi x(j-(q+1)/2) + \pi(q-1)/4\}]$ odd or vice-versa. In any case, their product is an odd function such that the l-th and the $(p-l+1)$-th summand cancel. Since p is even, the sum vanishes. Analogously, for the last term, the l-th and the $(p-l+1)$-th summand of the outer sum cancel each other, since

$$\begin{aligned}
&\phi_{q,i}(t_l)\sum_{k=1}^{p}\left[\phi_{q,j}(t_k)-\sqrt{\frac{2}{p}}\cos\{\pi(k-1)t_{j,q}+\pi(q-1)/4\}\right]r_{|l-k|}\\
&=(-1)^{i-1}\phi_{q,i}(x_{p-l+1})\sum_{k=1}^{p}\left[\phi_{q,j}(t_{p-k+1})-\sqrt{\frac{2}{p}}\cos\{\pi(p-k)t_{j,q}+\pi(q-1)/4\}\right]r_{|l-(p-k+1)|}\\
&=(-1)^{i-1}\phi_{q,i}(t_{p-l+1})\sum_{k=1}^{p}\left[\phi_{q,j}(t_{p-k+1})-\sqrt{\frac{2}{p}}\cos\{\pi(p-k)t_{j,q}+\pi(q-1)/4\}\right]r_{|(p-l+1)-k|}\\
&=(-1)^{|i-j|}\phi_{q,i}(t_{p-l+1})\sum_{k=1}^{p}\left[\phi_{q,j}(t_k)-\sqrt{\frac{2}{p}}\cos\{\pi(k-1)t_{j,q}+\pi(q-1)/4\}\right]r_{|(p-l+1)-k|},
\end{aligned}$$

where in the second line we mirrored $\phi_{q,i}(t_l)$ and reversed the summation from $k=1,..,p$ to $n-k+1=n,....,1$. In the fourth line, we mirrored $\phi_{q,j}(t_{p-k+1})$ and $\cos\{\pi(p-k)t_{j,q}+\pi(q-1)/4\}$ once more. Thus, $\{\mathbf{\Phi}^t\mathbf{R}\mathbf{\Phi}\}_{i,j}=\rho(\pi t_{j,q})\delta_{ij}$ for $|i-j|$ odd.

Case $|i-j|$ is even

Note that $\rho\in\mathcal{H}_\beta(M)$ implies $r_s=\mathcal{O}(s^{-\beta})$, see Grafakos (2008, Cor. 3.3.10). Together with $|\phi_{q,i}(t_l)|=\mathcal{O}(p^{-1/2})$ this implies that the second term of Equation (4.20) is of the order $\mathcal{O}(p^{-\beta+1})$. Again, by $r_s=\mathcal{O}(s^{-\beta})$, the inner sum of the third term is $\mathcal{O}(l^{-\beta+1})$ with some constant C. Then, for $p\to\infty$,

$$\sum_{l=1}^{p/2}Cl^{-\beta+1}=\begin{cases}\mathcal{O}(C), & \beta>2\\ \mathcal{O}\{\log(p)\}, & \beta=2\\ \mathcal{O}\{\log(p)p^{-\beta+2}\}, & 1<\beta<2.\end{cases}$$

Together with $|\phi_{q,i}(t_l)|=\mathcal{O}(p^{-1/2})$ this implies that the third term of Equation (4.20) is of the order $\mathcal{O}(p^{-1})$ for $\beta>2$, and of the order $\mathcal{O}\{\log(p)p^{-\beta+1}\}$ for $1<\beta\leq 2$.

Since $\mathcal{H}_\beta(M)\subset\mathcal{W}_\beta(M)$, by Lemma 4.2 we have

$$\left|\frac{1}{p}\sum_{i=1}^{p}\rho(\pi x_i)-\int_0^1\rho(\pi x)\mathrm{d}x\right|\leq\frac{c}{p^2}\|\rho\|_{\mathcal{W}^\beta}.$$

Let $\{\psi_i\}_{i=1}^{\infty}$ denote the Demmler-Reinsch basis of $\mathcal{W}_\beta(M)$, as introduced in Section 4.6.1. In particular, $\|\phi_i - \psi_i p^{-1/2}\|_{L_2} = \mathcal{O}(p^{-3/2})$ and $\|p^{1/2}\phi_i - \psi_i\|_{\mathcal{W}^\beta} = \mathcal{O}(p^{-1})$. By Equation (4.11), it follows that $\psi_{q,j}(t_l)p^{-1/2}$ equals to $(2/p)^{1/2}\cos\{\pi(l-1)t_{j,q} + \pi(q-1)/4\}$ up to a term which decreases exponentially fast. Applying the Cauchy-Schwarz inequality, one can see that for some constant $c > 0$

$$\begin{aligned}
&\left|\sum_{l=1}^{p}\phi_i(t_l)\{\phi_{q,j}(t_l) - \psi_{q,j}(t_l)p^{-1/2}\}\right| \leq \left[\sum_{l=1}^{p}\phi_i(t_l)^2\sum_{l=1}^{p}\left[(\phi_{q,j}(t_l) - \psi_{q,j}(t_l)p^{-1/2})\right]^2\right]^{1/2} \\
&= \left[\sum_{l=1}^{p}\left[\{\phi_{q,j}(t_l) - \psi_{q,j}(t_l)p^{-1/2}\}\right]^2\right]^{1/2} \\
&\leq \left[\frac{c}{p^2}\left\|(p^{1/2}\phi_{q,j} - \psi_{q,j})^2\right\|_{W^\beta} + p\cdot\left|\int_0^1\{\phi_j(x) - \psi_{q,j}(x)p^{-1/2}\}^2\,\mathrm{d}x\right|\right]^{1/2} = \mathcal{O}(p^{-1}).
\end{aligned}$$

In a similar way, for the last term of Equation (4.20) it holds

$$\begin{aligned}
&\left|\sum_{l=1}^{p}\phi_{q,i}(t_l)\sum_{k=1}^{p}\left[\phi_{q,j}(t_k) - \sqrt{\frac{2}{p}}\cos\{\pi(k-1)t_{j,q} + \pi(q-1)/4\}\right]r_{|l-k|}\right| \\
&= \left|\sum_{k=1}^{p}\left[\phi_{q,j}(t_k) - \sqrt{\frac{2}{p}}\cos\{\pi(k-1)t_{j,q} + \pi(q-1)/4\}\right]\sum_{l=1}^{p}\phi_{q,i}(t_l)r_{|l-k|}\right| \\
&\leq \left[\sum_{k=1}^{p}\left[\phi_{q,j}(t_k) - \sqrt{\frac{2}{p}}\cos\{\pi(k-1)t_{j,q} + \pi(q-1)/4\}\right]^2\right]^{1/2}\left[\sum_{k=1}^{p}\left\{\sum_{l=1}^{p}\phi_{q,i}(t_l)r_{|l-k|}\right\}^2\right]^{1/2} \\
&\leq \frac{C}{p}.
\end{aligned}$$

In the last line, we used $\left|\sum_{l=1}^{p} r_{|l-k|}\right|^2 \leq C$ for each $k = 1, ..., p$ and some constant $C > 0$, since $\beta > 1$. All in all, we have shown that in case of p even

$$(\mathbf{\Phi}_q^t\mathbf{R}\mathbf{\Phi}_q)_{i,j} = \rho_j\delta_{i,j} + \frac{1 + (-1)^{|i-j|}}{2}\cdot\begin{cases}\mathcal{O}\{\log(p)\cdot p^{-\beta+1}\}, & 1 < \beta \leq 2,\\ \mathcal{O}(p^{-1}), & 2 < \beta.\end{cases}$$

Since ρ is at least Lipschitz continuous, we have $\rho(\pi t_{j,q}) = \rho(\pi t_j) + \mathcal{O}(p^{-1})$.

Case p is odd

The three cases $l \leq (p-1)/2,\ l > (p+1)/2$ and $l = (p+1)/2$ have to be considered in Equation (4.16). The identities

$$\begin{aligned} \phi_{q,i}(t_l) &= (-1)^{i-1}\phi_{q,i}(x_{p-l+1}) \\ \cos\{\pi t_{j,q}(l-1) + \pi(q-1)/4\} &= (-1)^{j-1}\cos\{\pi t_{j,q}(p-l) + \pi(q-1)/4\} \\ \sin\{\pi t_{j,q}(p-l) + \pi(q-1)/4\} &= -(-1)^{j-1}\sin\{\pi t_{j,q}(p-l) + \pi(q-1)/4\} \end{aligned}$$

hold similarly for all $l = 1, 2, .., p$. In particular, for $l = (p+1)/2$ the expression is mapped to itself. Hence, for the case $l = (p+1)/2$ there are two additional terms in Equation (4.20), which are

$$\begin{aligned} &\sqrt{\frac{2}{p}}\phi_{q,i}(x_l)\cos\{\pi(l-1)t_{j,q} + \pi(q-1)/4\}\left[\sum_{s=p-l}^{\infty}\cos\{\pi s t_{j,q}\}r_s - \sum_{s=l}^{\infty}\cos\{\pi s t_{j,q}\}r_s\right] \\ &\sqrt{\frac{2}{p}}\phi_{q,i}(t_l)\sin\{\pi(l-1)t_{j,q} + \pi(q-1)/4\}\left[\sum_{s=l}^{\infty}\sin\{\pi s t_{j,q}\}r_s - \sum_{s=p-l}^{\infty}\sin\{\pi s t_{j,q}\}r_s\right]. \end{aligned}$$

For j even and $l = (p+1)/2$, it holds that

$$\begin{aligned} \cos\{\pi(l-1)t_{j,q} + \pi(q-1)/4\} &= \cos\left[\pi\frac{\{(p+1)/2 - 1\}\{j - (q+1)/2\}}{p-1} + \pi(q-1)/4\right] \\ &= \cos\{(j-1)/2\pi\} = 0. \end{aligned}$$

The same holds for $\phi_{j,q}$, i.e., for j even, $\phi_{j,q}(t_l) = 0$ at the reflection point $x_{(p+1)/2} = 0.5$ as $\phi_{j,q}$ is an odd function with the property $\sum_{l=1}^{p}\phi_{j,q}(t_l) = 0$. Hence, if $|i-j|$ is odd the additional terms equal zero. The fourth and fifth term vanish in the case of $|i-j|$ odd as the additional $l = (p+1)/2$-th summand equals zero. If $|i-j|$ even, the two additional terms in Equation (4.20) are of the order $\mathcal{O}(p^{-1})$ and do not matter for the overall result. For the last two terms, the asymptotic order does not change for p odd.

(ii) and (iii): The proof is identical to part (i) since only the decay of the Fourier coefficients of ρ is used and that ρ is bounded by some constant M. Note that $\rho \in \mathcal{W}_\beta$ implies $r_k = \mathcal{O}(k^{-\beta})$. □

4.6.6 Simplified representation for the likelihood

The log-likelihood of the model model is given by

$$\ell_p(\lambda, q, \mathbf{R}) = -\frac{p+1}{2}\log\{\mathbf{Y}^t\mathbf{R}^{-1}(\mathbf{I}_p - \mathbf{S})\mathbf{Y} + 1\} + \frac{1}{2}\log|\mathbf{R}^{-1}(\mathbf{I}_p - \mathbf{S})|_+, \tag{4.7}$$

where

$$\mathbf{R}^{-1}(\mathbf{I}_p - \mathbf{S}) = \mathbf{\Phi}_q\left\{\mathbf{I}_p + \operatorname{diag}(\lambda p \boldsymbol{\eta}_q)\mathbf{\Phi}_q^t\mathbf{R}\mathbf{\Phi}_q\right\}^{-1}\operatorname{diag}(\lambda p \boldsymbol{\eta}_q)\mathbf{\Phi}_q^t.$$

Suppose that Lemma 4.1 holds in such a way that

$$(\mathbf{\Phi}_q^t\mathbf{R}\mathbf{\Phi}_q)_{ij} = \rho(\pi t_j)\delta_{ij} + \frac{1+(-1)^{|i-j|}}{2}\mathcal{O}(p^{-1}),\ i,j = q+1,\dots,p,$$

which imposes some regularity constraints on the spectral density ρ. In this case,

$$\mathbf{\Phi}_q^t\mathbf{R}\mathbf{\Phi}_q = \operatorname{diag}(\boldsymbol{\rho}) + \mathbf{E},$$

where $\boldsymbol{\rho} = \{\rho(\pi t_1),\dots,\rho(\pi t_p)\}^t$ and $\mathbf{E}$ is matrix with elements $E_{ij}=\frac{1+(-1)^{|i-j|}}{2}\mathcal{O}(p^{-1})$.

Note that Lemma 4.1 gives the values of the matrix for $i,j = q+1,\dots,p$, but the values for the first $q+1$ rows and columns are at most $\mathcal{O}(1)$ (follows from the summability of the rows and columns of $\mathbf{R}$ and that $(\mathbf{\Phi}_q)_{ij} = \mathcal{O}(p^{-1/2})$).

Now,

$$\begin{aligned}\mathbf{R}^{-1}(\mathbf{I}_p - \mathbf{S}) &= \mathbf{\Phi}_q\left\{\mathbf{I}_p + \operatorname{diag}(\lambda p\boldsymbol{\eta}_q\boldsymbol{\rho}) + \operatorname{diag}(\lambda p\boldsymbol{\eta}_q)\mathbf{E}\right\}^{-1}\operatorname{diag}(\lambda p\boldsymbol{\eta}_q)\mathbf{\Phi}_q^t\\ &= \mathbf{\Phi}_q\left\{\mathbf{I}_p + \operatorname{diag}\left(\frac{\lambda p\boldsymbol{\eta}_q}{1+\lambda p\boldsymbol{\eta}_q\boldsymbol{\rho}}\right)\mathbf{E}\right\}^{-1}\operatorname{diag}\left(\frac{\lambda p\boldsymbol{\eta}_q}{1+\lambda p\boldsymbol{\eta}_q\boldsymbol{\rho}}\right)\mathbf{\Phi}_q^t.\end{aligned}$$

Note that the matrix

$$\operatorname{diag}\left(\frac{\lambda p\boldsymbol{\eta}_q}{1+\lambda p\boldsymbol{\eta}_q\boldsymbol{\rho}}\right)\mathbf{E},$$

has first q rows equal to zero and since the elements of the diagonal matrix are less than one, the order of this matrix is the same as that of $\mathbf{E}$. Denote

$$\operatorname{diag}\left(\frac{\lambda p\boldsymbol{\eta}_q}{1+\lambda p\boldsymbol{\eta}_q\boldsymbol{\rho}}\right)\mathbf{E} = \frac{1}{p}\widetilde{\mathbf{E}},$$

where elements of $\widetilde{\mathbf{E}}$ are of order $\mathcal{O}(1)$ and the first q rows are zero.
Hence,

$$\log|\mathbf{R}^{-1}(\mathbf{I}_p - \mathbf{S})|_+ = \log\left|\operatorname{diag}\left(\frac{\lambda p\boldsymbol{\eta}_q}{1+\lambda p\boldsymbol{\eta}_q\boldsymbol{\rho}}\right)\right|_+ - \log\left|\left(\mathbf{I}_p + p^{-1}\widetilde{\mathbf{E}}\right) D_q\right|_+,$$

where $D_q = \operatorname{diag}(0_q, 1_{p-q})$. It is easy to see that the first term is of order $O\{p\lambda^{-1/(2q)}\}$, while the second term is $\mathcal{O}(p^{-1})$, which follows with Hadamard's bound: $|\mathbf{A}|^2 \leq \prod_{j=1}^{p}\sum_{i=1}^{p} a_{ij}^2$. Since the ij elements of $(I_p + p^{-1}\widetilde{\mathbf{E}})D_q$ are of order $\delta_{ij} + \mathcal{O}(p^{-1})$ for $i, j = q+1, \ldots, p$ and 0 for $i, j = 1, \ldots, q$

$$\left|(\mathbf{I}_p + p^{-1}\widetilde{\mathbf{E}})D_q\right|_+^2 \leq \prod_{j=q+1}^{p}\{1 + \mathcal{O}(p^{-1})\} = 1 + \mathcal{O}(p^{-1}),$$

so that $\log\left|(\mathbf{I}_p + p^{-1}\widetilde{\mathbf{E}})D_q\right|_+^2 = \mathcal{O}(p^{-1})$.

Writing

$$(\mathbf{I}_p + p^{-1}\,\widetilde{\mathbf{E}})^{-1} = \mathbf{I}_p - (\mathbf{I}_p + p^{-1}\,\widetilde{\mathbf{E}})^{-1}p^{-1}\widetilde{\mathbf{E}}$$

leads to

$$\mathbf{Y}^t\mathbf{R}^{-1}(\mathbf{I}_p - \mathbf{S})\mathbf{Y} = \mathbf{B}^t\operatorname{diag}\left(\frac{\lambda p\boldsymbol{\eta}_q}{1+\lambda p\boldsymbol{\eta}_q\boldsymbol{\rho}}\right)\mathbf{B} - \mathbf{B}^t(\mathbf{I}_p + p^{-1}\,\widetilde{\mathbf{E}})^{-1}p^{-1}\widetilde{\mathbf{E}}\operatorname{diag}\left(\frac{\lambda p\boldsymbol{\eta}_q}{1+\lambda p\boldsymbol{\eta}_q\boldsymbol{\rho}}\right)\mathbf{B}.$$

Using the definition of the matrix inverse $A^{-1} = |A|^{-1}\operatorname{adj}(A)$ and Hadamard's bound as above, implies that the matrix between the two vectors in the second term has elements of order $\mathcal{O}(p^{-1})$ for $i, j = q+1, \ldots, p$ and zero for $i, j = 1, \ldots, q$. Hence,

$$\mathbf{Y}^t\mathbf{R}^{-1}(\mathbf{I}_p - \mathbf{S})\mathbf{Y} = \sum_{i=q+1}^{p}\frac{B_i^2\lambda p\eta_{q,i}}{1+\lambda p\eta_{q,i}}\{1 + \mathcal{O}(p^{-1})\} - \sum_{\substack{i,j=q+1\\ i\neq j}}^{p} B_iB_jE_{ij}^*,$$

for a matrix

$$\mathbf{E}^* = (\mathbf{I}_p + p^{-1}\,\widetilde{\mathbf{E}})^{-1}(p^{-1}\widetilde{\mathbf{E}})\operatorname{diag}\left(\frac{\lambda p\boldsymbol{\eta}_q}{1+\lambda p\boldsymbol{\eta}_q\boldsymbol{\rho}}\right),$$

which has elements of order $\mathcal{O}(p^{-1})$. The second term in the last display is of order $\mathcal{O}_p(1)$ by the Cauchy-Schwarz inequality. Finally,

$$\mathbf{Y}^t\mathbf{R}^{-1}(\mathbf{I}_p - \mathbf{S})\mathbf{Y} = \sum_{i=q+1}^{p} \frac{B_i^2 \lambda p \eta_{q,i}}{1+\lambda p \eta_{q,i}} \{1 + \mathcal{O}(p^{-1})\} + \mathcal{O}_p(1).$$

4.6.7 Consistency of the estimators

Suppose that $\mathbf{Y} \sim \mathcal{N}(\boldsymbol{\mu}_0, \sigma^2\mathbf{R}_0)$ represents the distribution of the data.

The estimator of the regression function

Since $\mathbf{Y} - \boldsymbol{\mu}_0 \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{R}_0)$, we have

$$\mathbb{E}\|\hat{\boldsymbol{\mu}}_{\lambda,q,\mathbf{R}} - \boldsymbol{\mu}_0\|^2 = \boldsymbol{\mu}_0^t(\mathbf{I}_p - \mathbf{S}_{\lambda,q,\mathbf{R}})^t(\mathbf{I}_p - \mathbf{S}_{\lambda,q,\mathbf{R}})\boldsymbol{\mu}_0 + \sigma^2 \mathrm{tr}\big(\mathbf{S}_{\lambda,q,\mathbf{R}}\mathbf{R}_0\mathbf{S}_{\lambda,q,\mathbf{R}}^t\big).$$

Then, if $\mathbf{A} \leq \mathbf{B}$ means that $\mathbf{B} - \mathbf{A}$ is positive semidefinite, using the fact that the eigenvalues of $\mathbf{R}$ are in the interval $[\delta, 1/\delta]$, and the identity $\mathbf{R}^{-1}(\mathbf{S}_{\lambda,q,\mathbf{R}}^{-1} - \mathbf{I}_p) = \mathbf{S}_{\lambda,q,\mathbf{I}_p}^{-1} - \mathbf{I}_p$, we get

$$\mathbf{I}_p - \mathbf{S}_{\lambda,q,\mathbf{R}} = \mathbf{S}_{\lambda,q,\mathbf{R}}\mathbf{R}\mathbf{R}^{-1}(\mathbf{S}_{\lambda,q,\mathbf{I}_p}^{-1} - \mathbf{I}_p) \leq \frac{1}{\delta}(\mathbf{S}_{\lambda,q,\mathbf{I}_p}^{-1} - \mathbf{I}_p) \leq \frac{1}{\delta^2}(\mathbf{S}_{\delta\lambda,q,\mathbf{I}_p}^{-1} - \mathbf{I}_p),$$

where the definition of the smoother is used. In a similar way, since by assumption the eigenvalues of $\mathbf{R}_0$ are on $[\delta, 1/\delta]$, using the von Neumann trace inequality,

$$\mathrm{tr}\Big(\mathbf{S}_{\lambda,q,\mathbf{R}}\mathbf{R}_0\mathbf{S}_{\lambda,q,\mathbf{R}}\Big) \leq \frac{1}{\delta^2}\mathrm{tr}\Big(\mathbf{S}_{\lambda,q,\mathbf{R}}\mathbf{R}\mathbf{S}_{\lambda,q,\mathbf{R}}\Big).$$

Then, again by the definition of the smoother it follows that

$$\begin{aligned}\mathbf{S}_{\lambda,q,\mathbf{R}}\mathbf{R}\mathbf{S}_{\lambda,q,\mathbf{R}} &= \boldsymbol{\Phi}_q\left\{\boldsymbol{\Phi}_q^t\mathbf{R}^{-1}\boldsymbol{\Phi}_q + \lambda\,\mathrm{diag}(p\boldsymbol{\eta}_q)\right\}^{-2}\boldsymbol{\Phi}_q^t \\ &\leq \frac{1}{\delta^2}\boldsymbol{\Phi}_q\left\{\mathbf{I}_p + \delta\lambda\,\mathrm{diag}(p\boldsymbol{\eta}_q)\right\}^{-2}\boldsymbol{\Phi}_q^t = \frac{1}{\delta^2}\mathbf{S}_{\delta\lambda,q,\mathbf{I}_p}^2.\end{aligned}$$

Combining all of the above, we can conclude that

$$\mathbb{E}\|\hat{\boldsymbol{\mu}}_{\lambda,q,\mathbf{R}} - \boldsymbol{\mu}_0\|^2 \leq \frac{1}{\delta^4}\left\{\boldsymbol{\mu}^t(\mathbf{I}_p - \mathbf{S}_{\delta\lambda,q,\mathbf{I}_p})^2\boldsymbol{\mu}_0 + \sigma^2\mathrm{tr}\big(\mathbf{S}_{\delta\lambda,q,\mathbf{I}_p}^2\big)\right\}.$$

The quantity in $\{\cdot\}$ is just the usual risk for a smoothing spline with smoothing parameter $\delta\lambda$ for the model with independent noise with variance σ^2. The conclusion is that, up to constants, the risk (and therefore the rate) of $\hat{\mu}_0$ is the same as in the independent noise case. It is stressed again, though, that the finite sample performance of the smoother $\mathbf{S}_{\lambda,q,\mathbf{R}}$, as illustrated by the numerical simulations, is far superior to that of the smoother that ignores the correlation structure.

The estimator of the correlation matrix

Note that since both $\hat{\mathbf{R}}$ and $\mathbf{R}_0$ are Toeplitz matrices, so is their difference. It is known that

$$\|\hat{\mathbf{R}} - \mathbf{R}_0\|_2 = \sqrt{\lambda_{\max}\left\{(\hat{\mathbf{R}} - \mathbf{R}_0)^2\right\}} \leq 2\pi\|\hat{\boldsymbol{\rho}} - \boldsymbol{\rho}_0\|_\infty,$$

see e.g., p. 115 in Cai et al. (2013). As such, consistency of the estimator $\hat{\mathbf{R}}$ follows from consistency of the spline estimator $\hat{\boldsymbol{\rho}}$.

Now, note that the $\hat{\rho}_i$ are a spline smoothed version of $\tilde{\rho}_i$ that satisfy $T_{\rho_i}(\lambda, q, \tilde{\boldsymbol{\rho}}) = 0$. As such, if $\tilde{\rho}_i$ satisfies the i-th equation, then

$$\frac{B_i^2(\lambda p\eta_{q,i}\tilde{\rho}_i)^2}{(1+\lambda p\eta_{q,i}\tilde{\rho}_i)^2} = \frac{\lambda p\eta_{q,i}\tilde{\rho}_i}{1+\lambda p\eta_{q,i}\tilde{\rho}_i}\tilde{\rho}_i \approx \tilde{\rho}_i$$

for all but the first few i. Next, note that the fraction on the left is just the square of one of the entries in $\boldsymbol{\Phi}_q^t(\mathbf{I}_p - \mathbf{S}_{\lambda,q,\tilde{\boldsymbol{\rho}}})\mathbf{Y}$ where $\tilde{\boldsymbol{\rho}}$ is identified with the resulting estimate of the correlation matrix. The quantity $(\mathbf{I}_p - \mathbf{S}_{\lambda,q,\tilde{\boldsymbol{\rho}}})\mathbf{Y}$ corresponds to the residuals of a smoothing spline fit, and as just seen, regardless of the estimate of the correlation, this estimator is consistent with a rate depending on the smoothness of the regression function, so that the expectation of the residuals is negligible. As such, for any $\boldsymbol{\rho}$,

$$\frac{B_i\lambda p\eta_{q,i}\rho_i}{1+\lambda p\eta_{q,i}\rho_i} = \left\{\boldsymbol{\Phi}_q^t(\mathbf{I}_p - \mathbf{S}_{\lambda,q,\boldsymbol{\rho}})\boldsymbol{\Phi}_q\mathbf{B}\right\}_i,$$

where $\mathbf{B} = \boldsymbol{\Phi}_q^t\mathbf{Y}$, and where approximately

$$\boldsymbol{\Phi}_q^t(\mathbf{I}_p - \mathbf{S}_{\lambda,q,\boldsymbol{\rho}})\boldsymbol{\Phi}_q\mathbf{B} \sim N\left(\mathbf{0},\ \boldsymbol{\Phi}_q^t(\mathbf{I}_p - \mathbf{S}_{\lambda,q,\boldsymbol{\rho}})\mathbf{R}_0(\mathbf{I}_p - \mathbf{S}_{\lambda,q,\boldsymbol{\rho}})^t\boldsymbol{\Phi}_q\right).$$

Then, approximately

$$\left\{\boldsymbol{\Phi}_q^t(\mathbf{I}_p - \mathbf{S}_{\lambda,q,\boldsymbol{\rho}})\mathbf{R}_0(\mathbf{I}_p - \mathbf{S}_{\lambda,q,\boldsymbol{\rho}})^t\boldsymbol{\Phi}_q\right\}_{i,j} = \left(\frac{\lambda p\eta_{q,i}\rho_{0,i}}{1+\lambda p\eta_{q,i}\rho_{0,i}}\right)^2\rho_{0,i}\delta_{i,j} \approx \rho_{0,i}\delta_{i,j}$$

for all but the first few i. It therefore makes sense to model

$$\tilde{\rho}_i = \frac{B_i^2(\lambda p \eta_{q,i} \tilde{\rho}_i)^2}{(1 + \lambda p \eta_{q,i} \tilde{\rho}_i)^2} = \rho_{0,i} + \nu_i,$$

where ν_i are independent and have expectation zero and bounded variance. In such a case (Eggermont and LaRiccia, 2009), smoothing spline smoother applied to the $\tilde{\rho}_i$ delivers an estimate of the spectral density with a convergence rate depending on the smoothness of the spectral density (or, equivalently, on the speed of decay of the autocorrelation function.) In particular, consistency of $\hat{\mathbf{R}}$ in the spectral norm follows.

The heuristic argument above can be made precise but this is out of scope in this work as the concern was with the implementation of the numerical procedure.

Bibliography

Abramowitz, M. and Stegun, I. A. (1964). *Handbok of mathematical functions with formulas, graphs, and mathematical tables*, volume 55. US Government printing office.

Altman, N. (1990). Kernel smoothing of data with correlated errors. *J. Am. Statist. Ass.*, 85:749–759.

Amato, U., Antoniadis, A., De Feis, I., and Goude, Y. (2017). Estimation and group variable selection for additive partial linear models with wavelets and splines. *South African Statist. J.*, 51(2):235–272.

Azuma, K. (1967). Weighted sums of certain dependent random variables. *Tohoku Math. J., Second Series*, 19(3):357–367.

Bartlett, M. S. (1950). Periodogram analysis and continuous spectra. *Biometrika*, 37(1/2):1–16.

Bentkus, R. (1985). Rate of uniform convergence of statistical estimators of spectral density in spaces of differentiable functions. *Lithuanian Math. J.*, 25(3):209–219.

Bickel, P. J. and Levina, E. (2008a). Covariance regularization by thresholding. *Ann. Statist.*, 36:2577–2604.

Bickel, P. J. and Levina, E. (2008b). Regularized estimation of large covariance matrices. *Ann. Statist.*, 36(1):199–227.

Brockwell, P. and Davis, R. (2009). *Time series: theory and methods*. Springer Science & Business Media.

Brockwell, P. J. and Davis, R. A. (2002). *Introduction to time series and forecasting*. Springer.

Brown, L. D., Cai, T. T., Zhou, H. H., et al. (2010). Nonparametric regression in exponential families. *Ann. Statist.*, 38(4):2005–2046.

Cai, T. T., Ren, Z., and Zhou, H. H. (2013). Optimal rates of convergence for estimating Toeplitz covariance matrices. *Probab. Theory Related Fields*, 156(1-2):101–143.

Cai, T. T., Zhang, C.-H., and Zhou, H. H. (2010). Optimal rates of convergence for covariance matrix estimation.

Cai, T. T. and Zhou, H. H. (2010). Nonparametric regression in natural exponential families. *Borrowing Strength: Theory Powering Applications - A Festschrift for Lawrence D. Brown. Inst. Math. Stat. (IMS) Collect*, 6:199–215.

Cai, T. T. and Zhou, H. H. (2012). Optimal rates of convergence for sparse covariance matrix estimation. *Ann. Statist.*, pages 2389–2420.

Chiu, S.-T. (1989). Bandwidth selection for kernel estimate with correlated noise. *Statist. Probab. Lett.*, 8:347–354.

Choudhuri, N., Ghosal, S., and Roy, A. (2004). Bayesian estimation of the spectral density of a time series. *J. Amer. Statist. Assoc.*, 99(468):1050–1059.

Chu, C.-K. and Marron, J. (1991). Comparison of two bandwidth selectors with dependent errors. *Ann. Statist.*, 19:1906–1918.

Craven, P. and Wahba, G. (1978). Smoothing noisy data with spline functions. *Numer. Math.*, 31(4):377–403.

Du, X., Aubry, A., De Maio, A., and Cui, G. (2020). Toeplitz structured covariance matrix estimation for radar applications. *IEEE Signal Process. Lett.*, 27:595–599.

Edwards, M. C., Meyer, R., and Christensen, N. (2019). Bayesian nonparametric spectral density estimation using B-spline priors. *Statist. Comput.*, 29(1):67–78.

Eggermont, P. and LaRiccia, V. (2009). *Maximum penalized likelihood estimation*, volume 2 of *Springer series in Statistics*. Springer.

Ephraim, Y. and Roberts, W. J. (2005). On second-order statistics of log-periodogram with correlated components. *IEEE Signal Process. Lett.*, 12(9):625–628.

Fan, J., Liao, Y., and Liu, H. (2016). An overview of the estimation of large covariance and precision matrices. *Econom. J.*, 19(1).

Fang, Y., Wang, B., and Feng, Y. (2016). Tuning-parameter selection in regularized estimations of large covariance matrices. *J. Stat. Comput. Simul.*, 86(3):494–509.

Fix, G. J. (1972). Effects of quadrature errors in finite element approximation of steady state, eigenvalue and parabolic problems. In Aziz, A., editor, *The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations*, pages 525 – 556. Academic Press.

Fix, G. J. (1973). Eigenvalue approximation by the finite element method. *Adv. Math.*, 10(2):300 – 316.

Franaszczuk, P. J., Blinowska, K. J., and Kowalczyk, M. (1985). The application of parametric multichannel spectral estimates in the study of electrical brain activity. *Biol. Cybern.*, 51(4):239–247.

Gradshteyn, I. S. and Ryzhik, I. M. (2014). *Table of integrals, series, and products*. Academic press.

Grafakos, L. (2008). *Classical Fourier analysis*, volume 2. Springer.

Gray, R. M. et al. (2006). Toeplitz and circulant matrices: A review. *Foundations and Trends® in Communications and Information Theory*, 2(3):155–239.

Grenander, U. and Szegö, G. (1958). *Toeplitz forms and their applications*. Univ of California Press.

Gunter, A. I. (2001). Bias and variance of averaged and smoothed periodogram-based log-amplitude spectra. In *ISPA 2001. Proceedings of the 2nd International Symposium on Image and Signal Processing and Analysis. In conjunction with 23rd International Conference on Information Technology Interfaces (IEEE Cat.*, pages 452–457. IEEE.

Hall, P. and Keilegom, I. V. (2003). Using difference-based methods for inference in nonparametric regression with time series errors. *J. R. Statist. Soc. B*, 65(2):443–456.

Hall, P., Lahiri, S., and Polzehl, J. (1995). On bandwidth choice in nonparametric regression with both short- and long-range dependent errors. *Ann. Statist.*, 23:1921–1936.

Hart, J. (1994). Automated kernel smoothing of dependent data by using time series cross-validation. *J. R. Statist. Soc. B*, 56:529–542.

Hastie, T. and Tibshirani, R. (1990). *Generalized additive models*. Wiley Online Library.

Herrmann, E., Gasser, T., and Kneip, A. (1992). Choice of bandwidth for kernel regression when residuals are correlated. *Biometrika*, 79:783–795.

Hurvich, C. and Zeger, S. (1990). A frequency domain selection criterion for regression with autocorrelated errors. *J. Am. Statist. Ass.*, 85:705–714.

Karagiannis, T., Molle, M., and Faloutsos, M. (2006). Understanding the limitations of estimation methods for long-range dependence. *University of California*.

Katznelson, Y. (2004). *An introduction to harmonic analysis*. Cambridge University Press.

Klockmann, K. and Krivobokova, T. (2023a). Efficient nonparametric estimation of Toeplitz covariance matrices. *arXiv preprint arXiv:2303.10018*.

Klockmann, K. and Krivobokova, T. (2023b). On second-order statistics of the log-average periodogram.

Klockmann, K. and Krivobokova, T. (2023c). *vstdct: "Nonparametric Estimation of Toeplitz Covariance Matrices"*. R package version 0.1.

Kohn, R., Ansley, C., and Wong, C. (1992). Nonparametric spline regression with autoregressive moving average errors. *Biometrika*, 79:335–346.

Komlós, J., Major, P., and Tusnády, G. (1975). An approximation of partial sums of independent rv'-s, and the sample df. i. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 32:111–131.

Kooperberg, C., Stone, C. J., and Truong, Y. K. (1995). Rate of convergence for logspline spectral density estimation. *J. Time Ser. Anal.*, 16(4):389–401.

Kotz, S. and Nadarajah, S. (2004). *Multivariate t-distributions and their applications*. Cambridge University Press.

Krivobokova, T. (2013). Smoothing parameter selection in two frameworks for penalized splines. *J. R. Statist. Soc. B*, 75(4):725–741.

Krivobokova, T., Briones, R., Hub, J. S., Munk, A., and de Groot, B. L. (2012). Partial least-squares functional mode analysis: application to the membrane proteins aqp1, aqy1, and clc-ec1. *Biophys. J.*, 103(4):786–796.

Krivobokova, T., Serra, P., Rosales, F., and Klockmann, K. (2022). Joint non-parametric estimation of mean and auto-covariances for gaussian processes. *Comput. Statist. Data Anal.*, 173:107519.

Lebedev, N. N., Silverman, R. A., and Livhtenberg, D. (1965). Special functions and their applications. *Phys. Today*, 18(12):70.

Lee, Y. K., Mammen, E., and Park, B. U. (2010). Bandwidth selection for kernel regression with correlated errors. *Statistics*, 44(4):327–340.

Mason, D. M. and Zhou, H. H. (2012). Quantile coupling inequalities and their applications.

Maturana-Russel, P. and Meyer, R. (2021). Bayesian spectral density estimation using p-splines with quantile-based knot placement. *Comput. Statist.*, 36(3):2055–2077.

McMurry, T. L. and Politis, D. N. (2010). Banded and tapered estimates for autocovariance matrices and the linear process bootstrap. *J. Time Ser. Anal.*, 31(6):471–482.

Opsomer, J., Wang, Y., and Yang, Y. (2001). Nonparametric regression with correlated errors. *Statist. Sci.*, 16(134–153).

Patel, J. K. and Read, C. B. (1996). *Handbook of the normal distribution*, volume 150. CRC Press.

Pawitan, Y. and O'Sullivan, F. (1994). Nonparametric spectral density estimation using penalized Whittle likelihood. *J. Amer. Statist. Assoc.*, 89(426):600–610.

Pourahmadi, M. (2013). *High-dimensional covariance estimation: with high-dimensional data*, volume 882. John Wiley & Sons.

Purahmadi, M. (2011). Covariance estimation: the GLM and regularized perspectives. *Statist. Sci.*, 26:369–387.

Quah, D. (2000). Internet cluster emergence. *European economic review*, 44(4-6):1032–1044.

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

Robinson, P. M. (1989). Nonparametric estimation of time-varying parameters. In Hackl, P., editor, *Statistical Analysis and Forecasting of Economic Structural Change*, pages 253 – 264. Springer, Berlin.

Robinson, P. M. (1995). Log-periodogram regression of time series with long range dependence. *Ann. Statist.*, pages 1048–1072.

Rosales, F., Krivobokova, T., and Serra, P. (2022). *eBsc: "Empirical Bayes Smoothing Splines with Correlated Errors"*. R package version 4.15.

Rosales Marticorena, L. F. (2016). *Empirical Bayesian Smoothing Splines for Signals with Correlated Errors: Methods and Applications*. PhD thesis, Georg-August-Universität Göttingen, http://hdl.handle.net/11858/00-1735-0000-0028-87F9-6.

Rosenblatt, M. (2012). *Stationary sequences and random fields*. Springer Science & Business Media.

Rousseau, J., Szabo, B., et al. (2020). Asymptotic frequentist coverage properties of bayesian credible sets for sieve priors. *Ann. Statist.*, 48(4):2155–2179.

Rujivan, S., Sutchada, A., Chumpong, K., and Rujeerapaiboon, N. (2023). Analytically computing the moments of a conic combination of independent noncentral chi-square random variables and its application for the extended cox–ingersoll–ross process with time-varying dimension. *Mathematics*, 11(5):1276.

Sandberg, J. and Hansson-Sandsten, M. (2012). Optimal cepstrum smoothing. *Signal Process.*, 92(5):1290–1301.

Schoenberg, I. (1973). *Cardinal Spline Interpolation*. SIAM, Philadelphia.

Schwarz, K. (2012). *A unified framework for spline estimators*. PhD thesis, University of Göttingen.

Schwarz, K. and Krivobokova, T. (2016). A unified framework for spline estimators. *Biometrika*, 103(1):121–131.

Serra, P., Krivobokova, T., et al. (2017). Adaptive empirical Bayesian smoothing splines. *Bayesian Analysis*, 12(1):219–238.

Shao, X. and Wu, W. B. (2007). Asymptotic spectral theory for nonlinear time series. *Ann. Statist.*, 35(4):1773–1801.

Singer, M., Krivobokova, T., Munk, A., and De Groot, B. (2016). Partial least squares for dependent data. *Biometrika*, 103(2):351–362.

Sniekers, S., van der Vaart, A., et al. (2015). Adaptive bayesian credible sets in regression with a gaussian process prior. *Electron. J. Statist.*, 9(2):2475–2527.

Speckman, P. and Sun, D. (2003). Fully Bayesian spline smoothing and intrinsic autoregressive priors. *Biometrika*, 90(2):289–302.

Stuart, A. and Ord, K. (2010). *Kendall's advanced theory of statistics, Vol. 2A*, volume 1. John Wiley & Sons.

Taqqu, M. S. and Teverovsky, V. (1997). Robustness of Whittle-type estimators for time series with long-range dependence. *Communications in statistics. Stochastic models*, 13(4):723–757.

Thomson, D. J. (1982). Spectrum estimation and harmonic analysis. *Proceedings of the IEEE*, 70(9):1055–1096.

Tsybakov, A. B. (2009). *Introduction to nonparametric estimation*. Springer series in statistics. Springer.

Utreras Diaz, F. (1980). Pur ie choix du paramétre d'ajustement dans le lissage par fonctions spline. *Numer. Math.*, 34:15–28.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press.

Wahba, G. (1980). Automatic smoothing of the log periodogram. *J. Amer. Statist. Assoc.*, 75(369):122–132.

Walker, A. (1964). Asymptotic properties of least-squares estimates of parameters of the spectrum of a stationary non-deterministic time-series. *J. Austral. Math.*, 4(3):363–384.

Wang, Y. (1998). Smoothing spline models with correlated random errors. *J. Am. Statist. Ass.*, 93:341–348.

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(2):70–73.

Whittle, P. (1957). Curve and periodogram smoothing. *J. R. Stat. Soc. Ser. B Methodol.*, 19(1):38–47.

Wood, S. (2017). *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC.

Wu, W. B. and Pourahmadi, M. (2009). Banding sample autocovariance matrices of stationary processes. *Statist. Sin.*, pages 1755–1768.

Wu, W. B. and Xiao, H. (2012). Covariance matrix estimation in time series. In *Handb. Statist.*, volume 30, pages 187–209. Elsevier.

Xiao, H. and Wu, W. (2012). Covariance matrix estimation for stationary time series. *Ann. Statist.*, 40(1):466–493.

Yang, Y. et al. (2001). Nonparametric regression with dependent errors. *Bernoulli*, 7(4):633–655.

Yoo, W. W., Ghosal, S., et al. (2016). Supremum norm posterior contraction and credible sets for nonparametric multivariate regression. *Ann. Statist.*, 44(3):1069–1102.

Zygmund, A. (2002). *Trigonometric series*, volume 1. Cambridge university press.