Computational Statistics and Data Analysis 173 (2022) 107519

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

www.elsevier.com/locate/csda

Joint non-parametric estimation of mean and auto-covariances for Gaussian processes

Tatyana Krivobokova [a], Paulo Serra [b,*], Francisco Rosales [c], Karolina Klockmann [a]

[a] Department of Statistics and Operations Research, University of Vienna, Austria
[b] Department of Mathematics, Vrije Universiteit Amsterdam, the Netherlands
[c] Graduate school of Business, Universidad ESAN, Peru

ARTICLE INFO

Article history:
Received 13 July 2021
Received in revised form 21 April 2022
Accepted 22 April 2022
Available online 5 May 2022

Keywords:
Demmler-Reinsch basis
Empirical Bayes
Spectral density
Stationary process

ABSTRACT

Gaussian processes that can be decomposed into a smooth mean function and a stationary autocorrelated noise process are considered and a fully automatic nonparametric method to simultaneous estimation of mean and auto-covariance 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. Performance is demonstrated in simulations and real data analysis. The method is implemented in the R package *eBsc*.[1]

1. Introduction

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

$$\begin{aligned} Y_i &= f(t_i) + \sigma\epsilon_i,\ \mathbb{E}(\epsilon_i) = 0,\ \sigma > 0,\ \ i = 1, \dots, n, \\ \operatorname{corr}(\epsilon_i, \epsilon_j) &= r(i-j) = r_{|i-j|} \in [-1, 1],\ \ i, j = 1, \dots, n, \end{aligned} \tag{1}$$

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 represent time points. The functions f and r are unknown, but it is assumed that the noise terms $\{\epsilon_i\}_{i=1}^n$ are sampled from a stationary noise process and that f 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 $\{Y_i\}_{i=1}^n$ might be measures of some experimental quantity observed with a time dependent measurement error; in this case estimation of f is of interest, while r is considered as a nuisance parameter. It might also be that $\{Y_i\}_{i=1}^n$ are measurements from a stochastic process indexed by time or space with a seasonal or other deterministic effect, described by f; in this case the focus is rather on estimation of the autocorrelation r.

Non-parametric 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 – e.g., for prediction – as well as in multivariate analysis, in problems such as clustering, principal component analysis (PCA), linear and quadratic discriminant analysis and regression

* Corresponding author. De Boelelaan 1111, 1081 HV Amsterdam, the Netherlands.
E-mail address: p.j.de.andradeserra@vu.nl (P. Serra).

[1] The *eBsc* package is available as a supplementary material and on *CRAN*.

https://doi.org/10.1016/j.csda.2022.107519

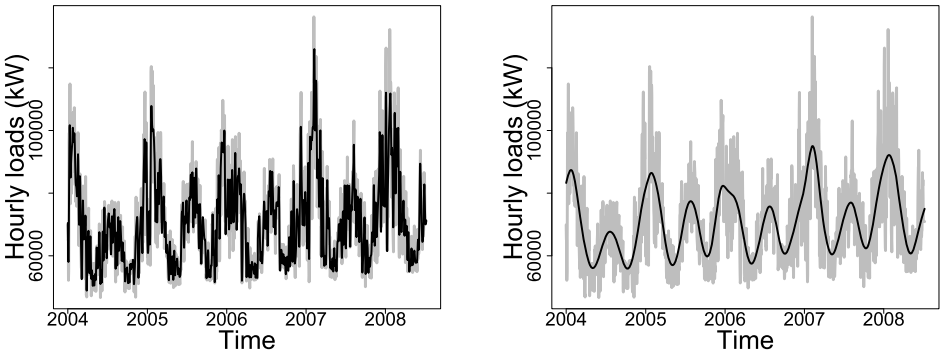

Fig. 1. Hourly loads (kW) for a US utility estimated ignoring dependence in the data (left) and assuming that errors follow an autoregressive process of order one (right). Grey line shows the data and black lines show estimators. Estimators are obtained by functions `gam` (left) and `gamm` (right) of the R package *mgcv*.

analysis. Two frameworks are usually considered when the mean is known: either there are n independent observations of a p-dimensional vector with correlated components, cf. Bickel and Levina (2008a,b); or there is one observation of an n-dimensional vector sampled from a stationary process, cf. Xiao and Wu (2012). The second framework is the one relevant for the present setting. Irrespectively of the framework, natural (moment or sample auto-covariances) estimators for the covariance matrix of the observed vectors are well known to be inconsistent (in, e.g., operator-norm) and some form of regularisation is needed (e.g., banding, tapering, thresholding) to ensure positive definiteness and consistency of the estimator. There is no simple data-driven approach to select the regularisation parameter in this context, and thus fully taking the correlation structure into account; see, e.g., Purahmadi (2011). Minimax rates of estimators of the correlation structure depend on the decay of the autocorrelation function as the lag increases (or, alternatively, smoothness of the corresponding spectral density); cf. Yang et al. (2001), Cai et al. (2010), Purahmadi (2011), Xiao and Wu (2012), see also Fan et al. (2016). As such, non-parametric estimation of the covariance structure of the process is of great importance.

If the mean function is unknown, then a natural approach to covariance estimation is to obtain a consistent estimator for the mean function first and then apply methods for (nonparametric) covariance estimation with a known mean. Unfortunately, ignoring dependence on the error process hinders the estimation of f.

To exemplify the problem, consider a time series of hourly loads (kW) for a US utility (grey line in Fig. 1), described in detail in Section 4. This process has clear seasonal effects over years and possibly over weeks and days. These deterministic effects can be modelled by a function f. If no parametric assumptions on f is made and errors are treated as i.i.d., then all standard nonparametric estimators of f are heavily affected by the ignored dependence in the errors. In particular, automatic selectors of the smoothing parameter (e.g., by cross-validation) choose a biased smoothing parameter leading to a nearly interpolating estimator in case of positively correlated $\{\epsilon_i\}_{i=1}^n$; see the black line on the left plot in Fig. 1. This problem has been known for several decades; for an overview see Opsomer et al. (2001).

Hence, there is a need for an approach that allows for correct smoothing of the mean over a large family of possible auto-covariance functions. Available methods rely on an explicit, parametric assumption on the correlation structure, e.g., assuming that $\{\epsilon_i\}_{i=1}^n$ follows an ARMA(p,q) process. Once r is parametrised, the usual smoothing parameter selection criteria are adjusted to incorporate r and both f and r are estimated simultaneously. For example, Hart (1994) introduces a time series cross-validation for f estimated by kernel estimators assuming errors to follow an AR(p) process, see also Altman (1990) and Hall and Keilegom (2003). Kohn et al. (1992) use spline smoothing for estimation of f and a general ARMA(p,q) model for the errors, estimating all parameters from either generalised cross validation or maximum likelihood. Similar ideas are employed in Wang (1998). This approach does allow for simultaneous estimation of f and r, but is limited to parametric models for r. Of course, the true correlation structure might be much more complex than any parametric one and if the parametric model for the error process is misspecified, then the estimator of f is strongly affected. More importantly, in practice it is difficult to select and verify a parametric model for r, if the mean function f 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 Fig. 1, assuming that the errors follow an AR(1) process. However, it is difficult to verify if this model is appropriate for the data at hand. For more detailed analysis of this dataset see Section 4.

Another group of methods for estimating f is based on trying to eliminate the influence correlation has on the smoothing parameter thus treating r as a nuisance parameter. For instance, Chu and Marron (1991) and Hall et al. (1995) study the modified cross-validation criterion obtained by leaving out 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. A different route is taken by Herrmann et al. (1992) and Lee et al. (2010): in the smoothing parameter selection criteria they incorporate some sample estimators of auto-covariances that depend on unknown parameters linked to assumptions on the error process. These methods require careful selection of tuning parameters, for which no data-driven methods are available.

In this paper, a likelihood based method that provides an estimator for a regression function f observed with correlated, additive noise is presented. Estimators of the noise level σ, and of the autocorrelation function r are developed as well. Furthermore, the proposed empirical Bayesian framework provides a computationally attractive way of constructing confidence sets for the regression function that take the correlation structure into account. There are quite a few novel points in this work. Contrary to other approaches in the literature, the proposed method is completely automatic so that no tuning parameters need to be set by the user, and it is also fully nonparametric. The proposed estimate of the autocorrelations is also novel – spline smoothers are used to estimate the spectral density of the noise process. The autocorrelations are then reconstructed from the estimate of the spectral density, rather than by tapering or thresholding some empirical estimate. The resulting covariance matrix is positive definite by construction, and consistent in operator norm. Further properties of this estimator require new tools to be developed and will be investigated in a separate work.

This paper is structured as follows. In Section 2 the proposed estimators are introduced, Section 3 contains simulation results, Section 4 presents a real data example, and Section 5 closes the paper with some conclusions. Technical results are in the Appendix to this paper.

2. Construction of the estimators

Assume that in model (1) the regression function f belongs to a Sobolev space $\mathcal{W}_\beta$, $\beta > 1/2$; see Appendix A.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 penalisation framework is followed (cf. (2)) so that the estimators are expected to perform well under other data distributions. To estimate f, so called infill asymptotics are used, cf. Robinson (1989) for more details. Define $\boldsymbol{Y} = (Y_1, \dots, Y_n)^T$, $\boldsymbol{f} = f(\boldsymbol{t}) = \{f(t_1), \dots, f(t_n)\}^T$, $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_n)^T$, where $t_i = (i-1)/(n-1)$, $i = 1, \dots, n$. The observations in (1) are modelled as

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

The symmetric $n \times n$ matrix $\boldsymbol{R}$ satisfies the regularity assumptions described in Lemma 1. These do imply that the first row of $\boldsymbol{R}$ is absolutely summable which means that the noise process is modelled as having short-term dependence. However, the assumptions are otherwise quite mild; in particular, a large (non-parametric) collection of covariance structures for the noise terms covering 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 $\boldsymbol{R}$ lie on the interval $[\delta, 1/\delta]$.

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

$$\min_{f \in \mathcal{W}_q} \left[\frac{1}{n} \{\boldsymbol{Y} - f(\boldsymbol{t})\}^T \boldsymbol{R}^{-1} \{\boldsymbol{Y} - f(\boldsymbol{t})\} + \lambda \int_0^1 \{f^{(q)}(t)\}^2 dt \right], \tag{2}$$

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

$$\boldsymbol{S} = \boldsymbol{S}_{\lambda,q,\boldsymbol{R}} = \boldsymbol{\Phi}_q \left\{ \boldsymbol{\Phi}_q^T \boldsymbol{R}^{-1} \boldsymbol{\Phi}_q + \lambda \operatorname{diag}(n\boldsymbol{\eta}_q) \right\}^{-1} \boldsymbol{\Phi}_q^T \boldsymbol{R}^{-1}, \tag{3}$$

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

2.1. Estimation of the remaining parameters

Although a frequentist paradigm is followed, the estimator $\hat{f}$ 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; cf. Section 2.3.

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

$$\left| \frac{\boldsymbol{R}^{-1}(\boldsymbol{S}^{-1} - \boldsymbol{I}_n)}{2\pi\sigma^2} \right|_+^{1/2} \exp\left\{ -\frac{\boldsymbol{f}^T \boldsymbol{R}^{-1}(\boldsymbol{S}^{-1} - \boldsymbol{I}_n)\boldsymbol{f}}{2\sigma^2} \right\}, \tag{4}$$

where $|\cdot|_+$ denotes the product of non-zero eigenvalues ($\boldsymbol{S}$ has exactly q eigenvalues equal to 1). This prior corresponds to a non-informative part on the null-space of $\boldsymbol{R}^{-1}(\boldsymbol{S}^{-1} - \boldsymbol{I}_n)$ and a proper Gaussian prior on the remaining space. (This prior

distribution happens to be independent of $\boldsymbol{R}$. This follows from the identity $\boldsymbol{R}^{-1}(\boldsymbol{S}^{-1}-\boldsymbol{I}_n)=\boldsymbol{S}_I^{-1}-\boldsymbol{I}_n$, where $\boldsymbol{S}_I$ denotes the smoother matrix with $\boldsymbol{R}=\boldsymbol{I}_n$; cf. Appendix A.2 for the derivation of the identity.) The mean of the corresponding posterior distribution for $\boldsymbol{f}|(\boldsymbol{t},\lambda,\sigma^2,\boldsymbol{R})$ is the smoothing spline estimator $\hat{\boldsymbol{f}}=\boldsymbol{S}\boldsymbol{Y}$; cf. Speckman and Sun (2003). The variance σ^2 given $(\boldsymbol{t},\lambda,q,\boldsymbol{R})$ is endowed with an inverse-gamma prior $\mathrm{IG}(a,b)$, $a,b>0$.

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

$$\hat{\sigma}^2=\frac{\boldsymbol{Y}^T\boldsymbol{R}^{-1}(\boldsymbol{I}_n-\boldsymbol{S})\boldsymbol{Y}+1}{n+1}. \tag{5}$$

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

$$\frac{\Gamma\{a+\frac{n-q}{2}\}\left|\boldsymbol{R}^{-1}(\boldsymbol{I}_n-\boldsymbol{S})\frac{2a-q}{2b}\right|_+^{1/2}}{\{\pi(2a-q)\}^{n/2}\Gamma(a-q/2)}\left\{1+\frac{\boldsymbol{Y}^T\boldsymbol{R}^{-1}(\boldsymbol{I}_n-\boldsymbol{S})\boldsymbol{Y}}{2b}\right\}^{-(n+2a-q)/2}. \tag{6}$$

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 $o(n)$ and lead to a proper prior and marginal distributions), so to simplify the log-likelihood, these are set $b=1/2$ and $a=(q+1)/2$ to obtain

$$\ell_n(\lambda,q,\boldsymbol{R})=-\frac{n+1}{2}\log\left\{\boldsymbol{Y}^T\boldsymbol{R}^{-1}(\boldsymbol{I}_n-\boldsymbol{S})\boldsymbol{Y}+1\right\}+\frac{1}{2}\log\left|\boldsymbol{R}^{-1}(\boldsymbol{I}_n-\boldsymbol{S})\right|_+, \tag{7}$$

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 (5) which is the proposed estimator for the variance of the noise.

2.2. Estimating equations and algorithm

The log-likelihood (7) depends on the unknown parameters via the matrix $\boldsymbol{R}^{-1}(\boldsymbol{I}_n-\boldsymbol{S})$ only and can be represented in a more convenient way as

$$\boldsymbol{R}^{-1}(\boldsymbol{I}_n-\boldsymbol{S})=\boldsymbol{\Phi}_q\left\{\boldsymbol{I}_n+\mathrm{diag}(\lambda n\boldsymbol{\eta}_q)\boldsymbol{\Phi}_q^T\boldsymbol{R}\boldsymbol{\Phi}_q\right\}^{-1}\mathrm{diag}(\lambda n\boldsymbol{\eta}_q)\boldsymbol{\Phi}_q^T,$$

using the Demmler-Reinsch basis; cf. Appendix A.1. The key result that allows us to handle efficiently the simultaneous estimation of f and r (or, equivalently, $\boldsymbol{R}$) is given in the following lemma whose proof can be found in Appendix A.5,

Lemma 1. *Let $\boldsymbol{R}$ be a $n\times n$ covariance matrix of a stationary process with spectral density ρ. Set $\rho_j=\rho(\pi t_j)$.*

(i) If ρ belongs to the Hölder space $\mathcal{C}^{\gamma,\alpha}$, $\gamma\in\mathbb{N}$, $0<\alpha\le 1$, i.e. $\rho^{(\gamma)}$ is α-Hölder continuous, then, for $i,j=q+1,\dots,n$,

$$\{\boldsymbol{\Phi}_q^T\boldsymbol{R}\boldsymbol{\Phi}_q\}_{i,j}=\rho_j\delta_{i,j}+\mathbb{I}\{|i-j|\ \textit{is even}\}\cdot\begin{cases}O(\log(n)\cdot n^{-\gamma-\alpha+1}), & 1<\gamma+\alpha\le 2,\\ O(n^{-1}), & 2<\gamma+\alpha.\end{cases}$$

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

$$\{\boldsymbol{\Phi}_q^T\boldsymbol{R}\boldsymbol{\Phi}_q\}_{i,j}=\rho_j\delta_{i,j}+\mathbb{I}\{|i-j|\ \textit{is even}\}\cdot\begin{cases}O(\log(n)\cdot n^{-1}), & \beta=2,\\ O(n^{-1}), & \beta=3,4,\dots.\end{cases}$$

(Note that the $O(n^{-1})$ term is uniform over i,j.)

That is, the Demmler-Reinsch basis $\boldsymbol{\Phi}_q$ asymptotically diagonalises $\boldsymbol{R}$ for any q and the log-likelihood (7) can thus be represented as

$$\begin{aligned}\ell_n(\lambda,q,\boldsymbol{R})=&-\frac{n+1}{2}\log\left[\sum_{i=q+1}^{n}\frac{B_i^2\lambda n\eta_{q,i}}{1+\lambda n\eta_{q,i}\rho_i}\{1+o(1)\}+O_P(1)\right]+\\&+\frac{1}{2}\sum_{i=q+1}^{n}\log\left(\frac{\lambda n\eta_{q,i}}{1+\lambda n\eta_{q,i}\rho_i}\right)\{1+o(1)\},\end{aligned} \tag{8}$$

where $B_i = \{\boldsymbol{\Phi}_q^T \boldsymbol{Y}\}_i$, see Appendix A.6. The representation in (8) should hold under wider generality but Lemma 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 Appendix A.1 for more details), it is straightforward to differentiate the approximative log-likelihood (8) w.r.t. λ, q and ρ_i to obtain after appropriate scaling the estimating equations

$$T_\lambda(\lambda, q, \boldsymbol{\rho}) = \left\{ \sum_{i=q+1}^{n} \frac{B_i^2 \lambda n \eta_{q,i}}{\left(1+\lambda n \eta_{q,i}\rho_i\right)^2} - \hat{\sigma}^2 \sum_{i=q+1}^{n} \frac{1}{1+\lambda n \eta_{q,i}\rho_i} \right\}\{1+o(1)\},$$

$$T_q(\lambda, q, \boldsymbol{\rho}) = \left\{ \sum_{i=q+1}^{n} \frac{B_i^2 \lambda n \eta_{q,i} \log(n\eta_{q,i})}{\left(1+\lambda n \eta_{q,i}\rho_i\right)^2} - \hat{\sigma}^2 \sum_{i=q+1}^{n} \frac{\log\left(n\eta_{q,i}\right)}{1+\lambda n \eta_{q,i}\rho_i} \right\}\{1+o(1)\},$$

$$T_{\rho_i}(\lambda, q, \boldsymbol{\rho}) = \left(\frac{B_i^2 \lambda n \eta_{q,i}\rho_i}{1+\lambda n \eta_{q,i}\rho_i} - \rho_i \right)\{1+o(1)\}, \quad i = 1, \dots, n,$$

where

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

Joint estimates for all unknown model parameters are obtained by solving these equations simultaneously over $\lambda > 0$, $\rho_i > 0$, $i = 1, \dots, n$, and $q \in \mathbb{N}$. However, this cannot be done directly due to interdependences and non-linearities in the equations. In practice, the equations are solved iteratively. For each fixed q, start with an initial guess $\hat{\boldsymbol{R}}^{(0)}$ (typically just an identity matrix), to obtain a preliminary estimate $\hat{\lambda}^{(0)}$ and iterate to get $\hat{\lambda}_q$ and $\hat{\boldsymbol{R}}_q$. Finally, q is chosen to solve $T_q(\hat{\lambda}_q, q, \hat{\boldsymbol{R}}_q) = 0$.

Since $\rho_i = \rho(\pi t_i)$ are values of a smooth function ρ at given points, estimation of ρ should be carried out over a space of smooth functions. In the Bayesian framework this can, conceivably, be accomplished by introducing a suitable prior on $\boldsymbol{R}$, which acts as a penalty term. For simplicity, a post-processing procedure is performed instead. First, $\tilde{\rho}_i$ are obtained as solutions of the corresponding estimating equations. Second, $\tilde{\rho}_i$ are smoothed, i.e., $\hat{\boldsymbol{\rho}} = \boldsymbol{S}_{\xi,p,\boldsymbol{I}}\tilde{\boldsymbol{\rho}}$, where $\boldsymbol{S}_{\xi,p,\boldsymbol{I}}$ is a smoother matrix (3) with parameters ξ, p and $\boldsymbol{I}_n$. (Note that this is inevitable: the estimates $\tilde{\rho}_i$ are inconsistent.)

All together, 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, $\hat{\boldsymbol{R}}$ is recovered from $\hat{\boldsymbol{\rho}}$ by the discrete Fourier transform; for the details see Appendix A.3. The summary of the estimation procedure is given in Algorithm 1:

for q *in* Q_n **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_n;
set $\hat{\lambda}$ to $\hat{\lambda}_{\hat{q}}$;
set $\hat{\boldsymbol{\rho}} = \hat{\boldsymbol{\rho}}_{\hat{q}}$;
set $(\hat{\boldsymbol{R}})_{i,j} = \hat{r}_{|i-j|}$, with $\hat{r}_k = n^{-1}\sum_{l=1}^{n} \cos\{k\pi\{l-1\}/\{n-1\})\,\hat{\rho}_l$;

Algorithm 1: Estimation procedure.

Here, Q_n denotes the collection of values for q under consideration. The stopping rule is standard: after each iteration, the change in the value of the estimate of λ and the norm of the change in the estimate of $\boldsymbol{\rho}$ are checked; if these fall below a threshold, the algorithm terminates. The algorithm is fast and typically converges after just a few iterations.

In summary, the procedure outlined in Algorithm 1 provides an approximation for the optimisers of the likelihood (8) with the extra property that the estimate of the spectral density is smooth (which is not accomplished by simply maximising the likelihood.)

The estimators for $\hat{f}$ and $\hat{\boldsymbol{R}}$ are consistent; cf. Appendix A.7. However, deriving convergence rate for $\hat{f}$ and $\hat{\boldsymbol{R}}$ is not a trivial task due to the interdependence of the two estimators and requires a separate treatment. However, as far as $\hat{f}$ is

concerned, in light of Lemma 1 it is clear from (3) that $\boldsymbol{R}$ has a scaling effect on the smoother: as n grows $\boldsymbol{S}_{\boldsymbol{I},\lambda/\delta} \leq \boldsymbol{S}_{\boldsymbol{R},\lambda} \leq \boldsymbol{S}_{\boldsymbol{I},\delta\lambda}$ for any $\boldsymbol{R}$ with eigenvalues on $[\delta, 1/\delta]$, where $\boldsymbol{A} \leq \boldsymbol{B}$ means that $\boldsymbol{B} - \boldsymbol{A}$ is positive semi-definite. The conclusion is that the effect of the smoother $S_{\boldsymbol{R},\lambda}$ is comparable to that of $S_{\boldsymbol{I},\lambda}$ for a difference choice of λ. It is stressed, however, that the small sample behaviour of the smoother $S_{\boldsymbol{R},\lambda}$ is far superior in the presence of correlation; the smoother $S_{\boldsymbol{I},\lambda}$ is well known to underperform particularly for positively correlated data since, contrary to the smoother $S_{\boldsymbol{R},\lambda}$, it does not take the correlation structure of the noise into consideration. From the likelihood in (8) it is clear, however, that the presence of correlation simply results in the presence of the (bounded) ρ_i in the denominators. This means that it should be possible to study the asymptotics of the proposed estimator $\hat{f}$ in a similar way as when there is no correlation in the data as in Serra et al. (2017).

Also the asymptotic behaviour of the estimator $\hat{q}$ should also be similar to when there is not correlation in the data; cf. Serra et al. (2017). Either way, the estimator should perform well even if q is set by the user and not selected from the data; $q = 2$ is a popular choice. 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 lead to the same large sample behaviour for estimators. However, the small sample behaviour of estimators that ignore correlation can be significantly improved by use of the proposed approach.

2.3. Confidence sets

Once estimates $\hat{\lambda}$, $\hat{\boldsymbol{R}}$, $\hat{q}$, and $\hat{\sigma}^2$ for respectively λ, $\boldsymbol{R}$, β, and σ^2 are available, these can be plugged into the marginal posterior for $\boldsymbol{f}$ to obtain the so called empirical, marginal posterior for $\boldsymbol{f}$:

$$\begin{aligned}\hat{\Pi}^{\boldsymbol{f}}\left(\cdot \mid \boldsymbol{Y}\right) &= \Pi^{\boldsymbol{f}}\left(\cdot \mid \boldsymbol{Y}, \hat{\sigma}^2, \hat{\lambda}, \hat{q}, \hat{\boldsymbol{R}}\right) \\ &= \Pi^{\boldsymbol{f}}\left(\cdot \mid \boldsymbol{Y}, \sigma^2, \lambda, q, \boldsymbol{R}\right)\big|_{(\lambda,q,\boldsymbol{R},\sigma^2)=(\hat{\lambda},\hat{q},\hat{\boldsymbol{R}},\hat{\sigma}^2)}.\end{aligned} \tag{9}$$

Given $\boldsymbol{Y}$, this is just a multivariate $t_{n+1}\left(\hat{\boldsymbol{f}}, \hat{\sigma}^2\hat{\boldsymbol{S}}\hat{\boldsymbol{R}}\right)$ distribution. Here, $\hat{\boldsymbol{f}}$ is the spline estimate, $\hat{\boldsymbol{S}}$ is the smoother matrix $\boldsymbol{S}$ with $(\hat{\lambda}, \hat{q}, \hat{\boldsymbol{R}})$ plugged in for $(\lambda, q, \boldsymbol{R})$, 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

$$\hat{C}_n(L) = \left\{\boldsymbol{f} : \|\boldsymbol{f} - \hat{\boldsymbol{f}}\|_2^2 \leq L\,\hat{\sigma}^2\, s_n(\hat{\lambda}, \hat{q}, \hat{\boldsymbol{R}})\right\}, \qquad L \geq 1, \tag{10}$$

which (for an appropriate, known, sequence s_n) satisfy

$$\hat{\Pi}\left\{\hat{C}_n(L) \mid \boldsymbol{Y}\right\} \geq 1 - \alpha, \qquad L \geq 1,$$

and are therefore a credible set – a small, high probability region of the empirical posterior; cf. Serra et al. (2017) for a more detailed outline of the construction. In (Serra et al., 2017, Theorem 3), it is shown that in the case where $\boldsymbol{R} = \boldsymbol{I}_n$ this Bayesian credible set has two important frequentist properties if L is taken appropriately large: a) the set contains the true underlying regression function with probability converging to 1, uniformly over a large subset of functions, and b) with probability converging to 1, the (random) radius of $\hat{C}_n(L)$ is of the order of the minimax risk corresponding to the smoothness class to which the regression function belongs.

As argued in the previous section, the presence of correlation should have a relatively simple scaling effect on the smoothing parameter. As such, the theoretical results of Serra et al. (2017) can in principle be extended to cover the correlated noise case as well, by (eventually) picking larger values of the multiplier L, but this problem 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.

3. Numerical simulations

In this section, the small sample performance of the proposed estimation procedure is investigated, and compared to several alternatives. The simulation setup for data coming from model (1) is as follows. Two regression functions are considered

$$\begin{aligned}f_1(t) &= \sum_{i=3}^{n} \psi_{3,i}(t)\{\pi(i-1)\}^{-3}\cos(2i), \\ f_2(t) &= 2\sin(4\pi t),\end{aligned}$$

where $\psi_{3,i}$ is the i-th Demmler-Reinsch basis of $\mathcal{W}_3$ given explicitly in Appendix A.1, $n = 800$ and $t \in [0, 1]$. Both functions are subsequently scaled to have standard deviation 1. The standard deviation of the residuals is taken to be $\sigma = 0.33$ to imply a medium signal-to-noise ratio of 3. All reported results are based on the Monte Carlo sample $M = 500$. 5 types of the

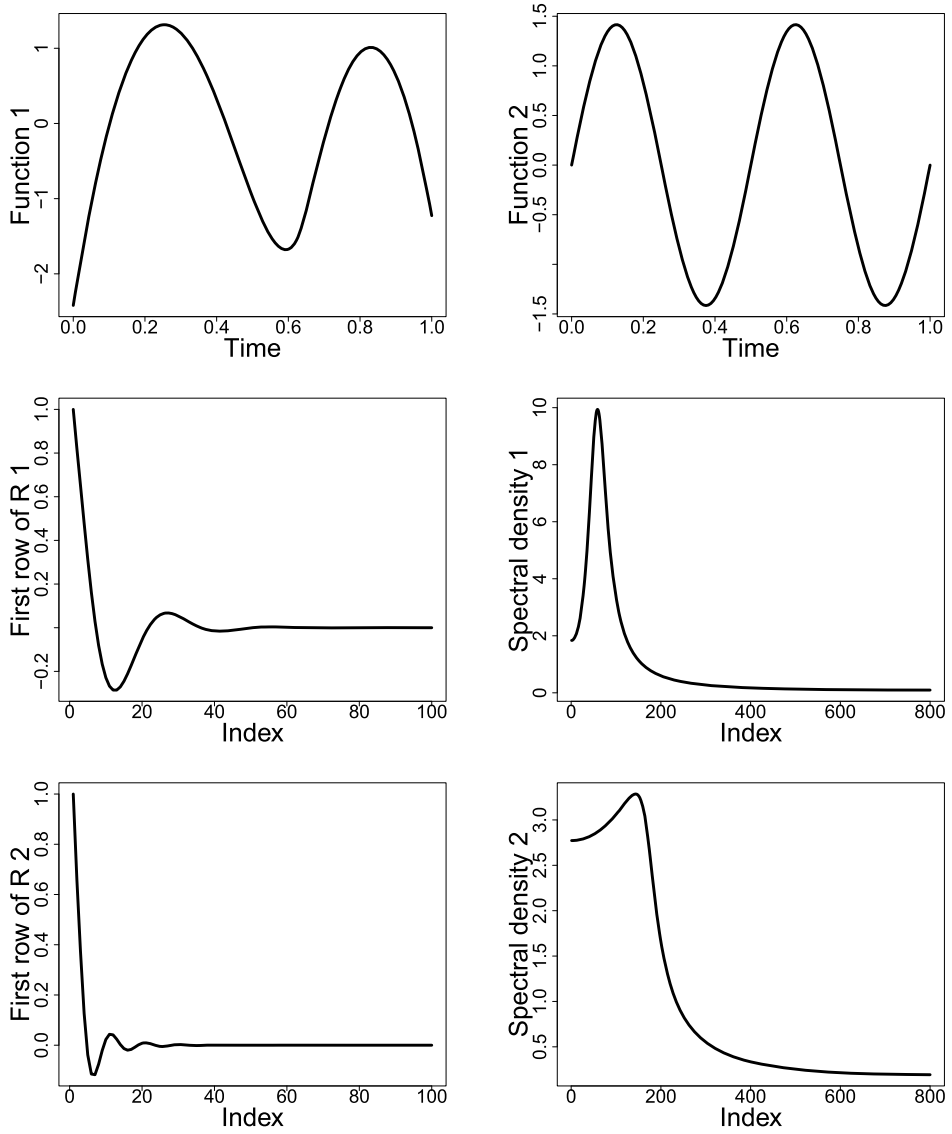

Fig. 2. Regression functions f_1 (top left), f_2 (top right), first 100 values of the first row of $\boldsymbol{R}_1$ (middle left) with the corresponding spectral density (middle right) and of the first row of $\boldsymbol{R}_2$ (bottom left) with the corresponding spectral density.

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 $\boldsymbol{R}_1$ and $\boldsymbol{R}_2$ 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 distribution of the error processes is taken to be Gaussian according to the model (1). However, other error processes were also experimented with, such as Gamma and uniform and found nearly no differences in performance.

The plots in the top row of Fig. 2 show both regression functions. The middle and bottom rows show the first 100 elements of the first row of $\boldsymbol{R}_1$ (middle, left) and $\boldsymbol{R}_2$ (bottom, left) and the corresponding spectral densities (middle and bottom right).

Before the summary of the simulation results comes a demonstration of how the proposed method works in practice. The data are simulated as described above with the regression function $f_1 \in \mathcal{W}_3$ and ARMA$(2, 2)$ dependence in the residuals. After Algorithm 1 has converged, one checks the estimate for q. The top left plot of Fig. 3 shows the estimating equation $T_q(\hat{\lambda}, q, \hat{\boldsymbol{R}})$ which has a zero very close to $q = 3$, as it should for $f_1 \in \mathcal{W}_3$. Next, the corresponding estimate for f is obtained; it is shown in the top right plot of Fig. 3, together with the data (black dots) and 95% confidence bands (grey area) constructed as described in Section 2.3 ($L = \log n$ 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

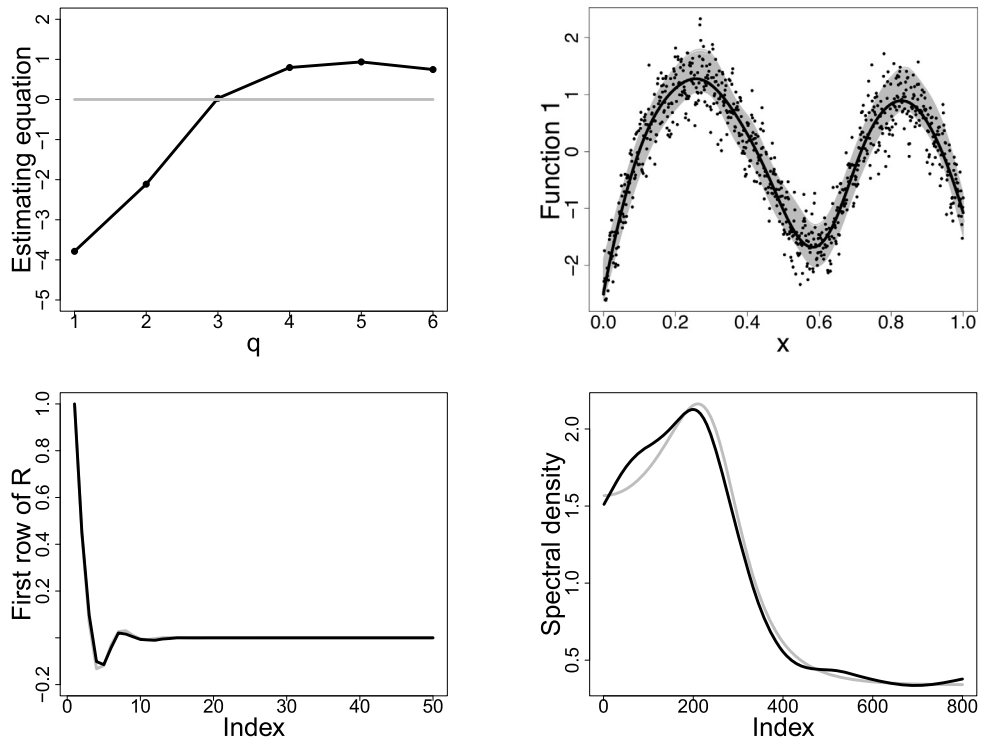

Fig. 3. Estimating equation T_q in the top left and the corresponding estimator of f_1 (black line) together with the data (black points) and point-wise 95% confidence intervals (grey area) in the top right. First 50 elements of the first row of the correlation matrix in grey with its estimator in black (bottom left) and the true spectral density in grey with its estimator in black (bottom right).

the bottom right plot of Fig. 3. 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 f (with data driven q, λ) and $\boldsymbol{R}$.

There is no other fully nonparametric method for joint estimation of mean and auto-covariance available that could be found in the literature. Nonetheless, several approaches that can be at least partly compared to the proposed method were considered. Among approaches that make a parametric assumption on $\boldsymbol{R}$, the well-established method based on splines that is implemented in the statistical software R in the function `gamm` of package *mgcv* was considered, see Wood (2017). This estimation procedure (further denoted by GAM) was applied to the processes with parametric structure of $\{\epsilon(t_i)\}_{i=1}^n$ only and set the correlation structure parameters in `gamm` to the truth, which is not available in practice.

Among approaches that make no parametric assumption on $\boldsymbol{R}$, the method of Herrmann et al. (1992), subsequently referred to as HER, was considered. This kernel based method uses sample autocorrelation estimators to improve bandwidth selection and is developed under assumption of m-dependence in the residuals. However, the authors claim that the method works well for estimation of f if the residuals satisfy "some mixing conditions". In general, this approach focuses on estimation of f and the quality of the estimator for r is not discussed.

Finally, a comparison with the proposed estimator for the auto-covariance with the standard banded nonparametric estimator (further BAND) is also presented. Since this approach requires a known mean, the true mean f into the estimator was plugged in, calculating

$$\hat{R}_{i,j}(b) = \hat{r}_{|i-j|}\mathbb{I}(|i-j| \le b),\ i, j = 1, \dots, n,$$
$$\hat{r}_k = \frac{1}{n}\sum_{i=1}^{n-k}\{Y(t_i) - f(t_i)\}\{Y(t_{i+k}) - f(t_{i+k})\},\ k = 0, 1, 2, \dots,$$

where b is the banding parameter that needs to be chosen.

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

For the method of Herrmann et al. (1992) the parameter m is set according to method (i) described in that paper (p. 787). Namely, m was chosen to be the largest integer such that $\hat{h}_m \ge 6/5\,\hat{h}_{m-1}$ and $m \le 0.2\sqrt{n}$, where $\hat{h}_m$ is a selected bandwidth with the parameter m. This parameter m is linked to the assumption of m-dependence in the residuals. In the experiments it was noticed that the influence of m on the estimator of f is not very pronounced, but it does affect

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

f₁							
	$A(\hat{f}_1)$			$A(\hat{R}_1)$			
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
f₂							
	$A(\hat{f}_2)$			$A(\hat{R}_2)$			
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

the estimator of the auto-covariance quite strongly. There is no simple data-driven approach to choose m such that both mean and auto-covariance estimators are optimal in some sense. In the proposed implementation, the function `glkerns` of package *lokern* by Eva Herrmann was used for the estimation of f with the second order kernel and for the estimation of f'' with the fourth order kernel.

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

Even though the proposed approach is adaptive and β can be estimated from the data, q was set to 2 to enable comparing the proposed method with all other procedures for estimating the mean, which should have the same convergence rate. All other parameters are estimated from the data. The proposed method as is referred to as BAS.

The results are summarised in Table 1. For all dependence structures and all functions, the metrics

$$A(\hat{f}_j) = \frac{1}{Mn}\sum_{k=1}^{n}\sum_{i=1}^{M}\{f_j(x_k) - \hat{f}_{j,i}(x_k)\}^2,$$

$$A(\hat{R}_j) = \frac{1}{Mn}\sum_{k=1}^{n}\sum_{i=1}^{M}\{R_j(k) - \hat{R}_{j,i}(k)\}^2, \quad j = 1, 2,$$

were used. Here $\hat{f}_{j,i}$ denotes an estimator of f_j in i-th Monte Carlo run with the residuals following one of the six processes. Similarly, $R_j(k)$ denotes the k-th entry of the first row of one of six true residual correlation matrices added to the jth regression function f_j and $\hat{R}_{j,i}(k)$ is its estimator in the ith Monte Carlo run.

Method HER performs best in estimation of the mean, except for the case with AR(1) error process with the parameter 0.9, where this approach fails completely. For estimation of auto-covariances HER performs, in contrast, worst, which is most likely due to the ad-hoc choice of the tuning parameter m. For parametric structures of the error process method GAM performs very similar to BAS in estimation of the mean and is better in estimation of the auto-covariances. This is not surprising, since the proposed nonparametric method should have a slower convergence rate than the parametric one. However, in simulations the correct parametric model specification in the function `gamm` was used, which is not available in practice. Finally, BAND's performance in nonparametric estimation of auto-covariances 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 to the considered alternatives, most of which rely on oracle choices for their parameters.

4. Real dataset

In this section, data from the load forecasting track of the Global Energy Forecasting Competition 2012 (http://www.drhongtao.com/gefcom/2012) is considered. These are data on hourly loads of a US facility at 20 zones 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. The interest does not lie in forecasting the data, but rather understanding their structure. The mean over all 20 zones over the

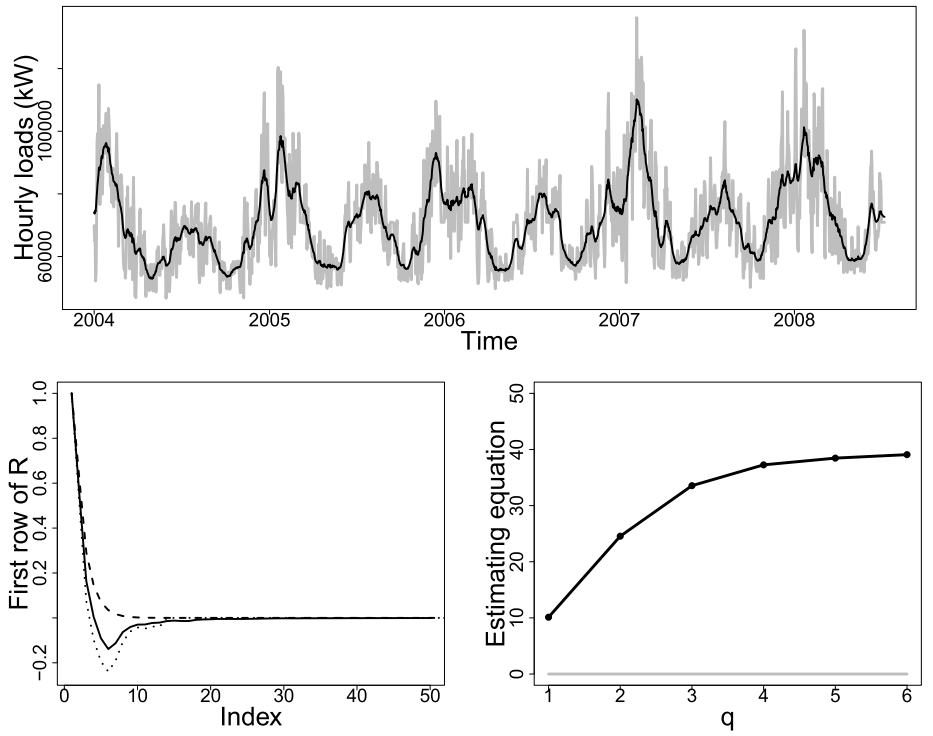

Fig. 4. Estimators of the mean (top), first row of the correlation matrix (first 50 values, bottom left) and the estimating equation for the smoothness degree q (bottom right) of the hourly loads of a US facility. Data are shown in grey, BAS estimators are the black solid lines, GAM estimators are dashed lines, BAND estimator for the covariance is the dotted line.

whole time period – all together $n = 1650$ observations – is used, justifying the use of a Gaussian model. Missing values were imputed using R package `Hmisc`; omitting these missing values lead to the same estimators and same conclusions.

Fitting the data with the proposed approach delivers a positive and increasing function in the estimating equation for q, which implies that the mean function has either 1 continuous derivative or is even less smooth. This suggests that a fit with $q = 1$ is 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, see Amato et al. (2017). The estimator of the mean with $q = 1$ is shown in the top plot of Fig. 4 as a black line. The corresponding autocorrelations are shown as solid black line in the bottom left plot.

Setting $q = 1$ in GAM with an AR(2) process for the residuals results in the estimator of the mean that is very close to the one obtained with BAS, but having somewhat less pronounced peaks (not clearly visible in the plot). The corresponding autocorrelation estimator is shown as a dashed line in the bottom left plot of Fig. 4. 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 $b = 12$, which is shown as a dotted line in the bottom left plot and is closer to the BAS estimator of the covariance.

Estimation with BAS setting $q = 2$ gives estimates of the mean that are very close to the one obtained by GAM, setting the correlation structure to an AR(1) process; this is shown in Fig. 1, right plot. The corresponding autocorrelation estimators are also reasonably close. The residual analysis suggests, however, that the fit with $q = 1$ is more appropriate for this data.

5. Conclusions

Correlation is ubiquitous in applications, particularly when data is collected sequentially over time, and ignoring this can have severe consequences for inference. Covariance in data is typically taken into account by either making parametric assumptions on the dependence structure, or by relying on introducing tuning parameters that then have be set heuristically. A fully automatic, nonparametric method to estimate both mean function and auto-covariance function is proposed. It further supplies credible sets for the mean function that quantify the uncertainty of the estimator. The order of the splines to be used in the estimator is also estimated from the data, 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.

Derivation of the convergence rate turned out to be a non-trivial task due to the interdependence of the estimators and will be performed in a separate work. Nonetheless, 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. Since there are no competing fully non-parametric, fully data-driven, methods available in the literature, in the numerical simulations a comparison is made between the proposed approach and methods from the literature that require oracle knowledge of parameters. Even so, the numerical simulations suggest that the proposed method is competitive to these alternatives.

Acknowledgements

The authors are greatly indebted to Dr. Christoph Lehrenfeld for helpful discussions, as was as the co-editor and two anonymous reviewers for useful feedback during the review process.

T.K. would like to thank the Isaac Newton Institute for Mathematical Sciences, Cambridge, for support and hospitality during the programme "Statistical Scalability", where work on this paper was undertaken and supported by EPSRC grant numbers EP/K032208/1 and EP/R014604/1.

P.S. carried out part of this research while he was a postdoctoral researcher at the Institute of Mathematical Stochastics, Georg-August-Universität Göttingen and would like to acknowledge together with T.K. the support of the German Research Foundation (Deutsche Forschungsgemeinschaft) as part of the Institutional Strategy of the University of Göttingen.

Appendix A. Auxiliary results

In this appendix, contains a number of technical results that are used throughout this paper.

A.1. Demmler-Reinsch basis

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

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

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

$$\begin{aligned}&(-1)^q \psi_{\beta,i}^{(2\beta)} = \nu_{\beta,i}\psi_{\beta,i},\\ &\psi_{\beta,i}^{(l)}(0) = \psi_{\beta,i}^{(l)}(1) = 0,\ \ l = \beta, \beta+1, \ldots, 2\beta-1.\end{aligned}$$

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

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

where

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

for $S(\beta) = \cup_j \left\{(-1)^{j/(2\beta)}, \overline{(-1)^{j/(2\beta)}}\right\}$, with $0 \le j \le \beta-2$ 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 $\mathcal{S}_{2q-1}(\boldsymbol{x})$ is uniquely defined via

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

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

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

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

Lemma 2. *Let $f \in W_\beta(M)$, $\beta \geq 2$, then*

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

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

A.2. Matrix identities

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

$$\boldsymbol{S} = \boldsymbol{R}^{1/2} \left\{ \boldsymbol{I}_n + \lambda \boldsymbol{R}^{1/2} \boldsymbol{\Phi}_q \operatorname{diag}(n\boldsymbol{\eta}_q) \boldsymbol{\Phi}_q^T \boldsymbol{R}^{1/2} \right\}^{-1} \boldsymbol{R}^{-1/2}.$$

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

$$\boldsymbol{R}^{-1}(\boldsymbol{I}_n - \boldsymbol{S}) = (\boldsymbol{I}_n - \boldsymbol{S})^T \boldsymbol{R}^{-1}.$$

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

$$\boldsymbol{R}^{-1}(\boldsymbol{S}^{-1} - \boldsymbol{I}_n) = \lambda \boldsymbol{\Phi}_q \operatorname{diag}(n\boldsymbol{\eta}_q) \boldsymbol{\Phi}_q{}^T = \boldsymbol{S}_{\boldsymbol{I}}^{-1} - \boldsymbol{I}_n.$$

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

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

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

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

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

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

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

$$\begin{aligned}
\frac{\partial \boldsymbol{R}^{-1}(\boldsymbol{I}_n - \boldsymbol{S})}{\partial \rho_i} &= -\boldsymbol{R}^{-1}\frac{\partial \boldsymbol{R}}{\partial \rho_i}\boldsymbol{R}^{-1}(\boldsymbol{I}_n - \boldsymbol{S}) + \boldsymbol{R}^{-1}\frac{\partial \boldsymbol{S}}{\partial \rho_i} \\
&= -\boldsymbol{R}^{-1}(\boldsymbol{I}_n - \boldsymbol{S})\frac{\partial \boldsymbol{R}}{\partial \rho_i}\boldsymbol{R}^{-1}(\boldsymbol{I}_n - \boldsymbol{S}) = -(\boldsymbol{I}_n - \boldsymbol{S})^T\boldsymbol{R}^{-1}\frac{\partial \boldsymbol{R}}{\partial \rho_i}\boldsymbol{R}^{-1}(\boldsymbol{I}_n - \boldsymbol{S}).
\end{aligned}$$

A.3. $\boldsymbol{R}$ and its spectral density

Let $\boldsymbol{R}$ be a Toeplitz matrix in that $(\boldsymbol{R})_{i,j} = R_{i,j} = r_{i-j}$ for some sequence $\{r_i\}_{i\in\mathbb{Z}}$. Let us further assume that $R_{i,j} = 0$, if $|i-j| > m$, $m \in \mathbb{N}$. Denote the Fourier spectrum of $\boldsymbol{R}$ as

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

where $\boldsymbol{i}$ denotes the imaginary unit. If $\boldsymbol{R}$ is the correlation matrix of a stationary process, then additionally, $r_{i-j} = r_{j-i}$ so that

$$\rho(x) = 1 + 2\sum_{k=1}^{m} r_k \cos(kx), \quad \text{and,}$$

$$r_k = \frac{1}{2\pi}\int_{-\pi}^{\pi} \cos(kx)\rho(x)\,dx = \int_0^1 \cos(k\pi x)\rho(\pi x)\,dx.$$

In practice r_k is approximated via

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

A.4. Derivation of the posterior and marginal posterior

Consider $\boldsymbol{t}, \lambda, q, \boldsymbol{R}$ fixed throughout. Endow $(\boldsymbol{f}, \sigma^2)$ with a prior specified by endowing σ^2 with and inverse gamma prior with parameters a, b, and endowing $\boldsymbol{f} \mid \sigma^2$ with a prior with density (w.r.t. some appropriate dominating measure) as specified in (4). Using the fact that $\boldsymbol{S}^{-1} - \boldsymbol{I}_n$ has rank $n - q$, the prior density is thus proportional to

$$\left(\sigma^2\right)^{-\frac{n-q}{2}-a-1} \times \left|\boldsymbol{R}^{-1}(\boldsymbol{S}^{-1} - \boldsymbol{I}_n)\right|_+^{1/2} \times \exp\left\{-\frac{1}{2\sigma^2}\left[\boldsymbol{f}^T\boldsymbol{R}^{-1}(\boldsymbol{S}^{-1} - \boldsymbol{I}_n)\boldsymbol{f} + 2b\right]\right\}.$$

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

$$\left(\sigma^2\right)^{-\frac{2n-q}{2}-a-1} \times |\boldsymbol{S}\boldsymbol{R}|^{-1/2} \times \exp\left\{-\frac{Q}{2\sigma^2}\right\}, \tag{A.2}$$

where Q in the exponent is

$$Q = \left(\boldsymbol{f} - \boldsymbol{S}\boldsymbol{Y}\right)^T(\boldsymbol{S}\boldsymbol{R})^{-1}\left(\boldsymbol{f} - \boldsymbol{S}\boldsymbol{Y}\right) + 2b + \boldsymbol{Y}^T\boldsymbol{R}^{-1}\left(\boldsymbol{I}_n - \boldsymbol{S}\right)\boldsymbol{Y}^T.$$

Integrating out $\boldsymbol{f}$, the marginal density of the posterior distribution on σ^2 is recognised as an inverse gamma distribution with shape parameter $(n - q + 2a)/2$ and scale parameter $\{\boldsymbol{Y}^T\boldsymbol{R}^{-1}(\boldsymbol{I}_n - \boldsymbol{S})\boldsymbol{Y} + 2b\}/2$. Integrating out σ^2 instead leads to a marginal posterior density for $\boldsymbol{f}$ which is clearly a multivariate t-distribution with $n + 1$ degrees of freedom, expectation $\hat{\boldsymbol{f}} = \boldsymbol{S}\boldsymbol{Y}$, and scale $\hat{\sigma}^2\boldsymbol{S}\boldsymbol{R}$, where $\hat{\sigma}^2$ is given by (5).

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

$$\left(\sigma^2\right)^{-\frac{n-q}{2}-a-1} \times \exp\left\{-\frac{1}{2\sigma^2}\left[2b + \boldsymbol{Y}^T\boldsymbol{R}^{-1}\left(\boldsymbol{I}_n - \boldsymbol{S}\right)\boldsymbol{Y}^T\right]\right\},$$

which can be easily integrated with respect to σ^2 leading to the marginal likelihood for $\boldsymbol{Y}$ proportional to the one specified in (6).

A.5. Proof of Lemma 1

(*i*) W.l.o.g. assume that n is even. Noting that $t_i = (i-1)/(n-1)$ and denoting $t_{j,q} = \{j - (q+1)/2\}/(n-1)$, for $i, j = q+1, \ldots, n$,

$$\begin{aligned}
\{\boldsymbol{\Phi}_q^T\boldsymbol{R}\boldsymbol{\Phi}_q\}_{i,j} &= \sum_{l=1}^{n}\phi_{q,i}(t_l)\sum_{k=1}^{n}\sqrt{\frac{2}{n}}\cos\{\pi(k-1)t_{j,q} + \pi(q-1)/4\}r_{|l-k|} \\
&\quad + \sum_{l=1}^{n}\phi_{q,i}(t_l)\sum_{k=1}^{n}\left[\phi_{q,j}(t_k) - \sqrt{\frac{2}{n}}\cos\{\pi(k-1)t_{j,q} + \pi(q-1)/4\}\right]r_{|l-k|} \\
&= \rho(\pi t_{j,q})\delta_{i,j} \\
&\quad - \{1 + (-1)^{|i-j|}\}\sqrt{\frac{2}{n}}\sum_{l=1}^{n/2}\phi_{q,i}(t_l)\sum_{s=n-l+1}^{\infty}\cos\{\pi(l-1+s)t_{j,q} + \pi(q-1)/4\}r_s \\
&\quad - \{1 + (-1)^{|i-j|}\}\sqrt{\frac{2}{n}}\sum_{l=1}^{n/2}\phi_{q,i}(t_l)\sum_{s=l}^{\infty}\cos\{\pi(l-1-s)t_{j,q} + \pi(q-1)/4\}r_s
\end{aligned} \tag{A.3}$$

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

It was used that $\sum_{|s|\le n/2}\cos(\pi s t_{j,q})r_s=\rho(\pi t_{j,q})+\sum_{|s|>n/2}^{\infty}\cos(\pi s t_{j,q})r_s$, the cosine addition identity, and that cosine and sine are even and odd functions, respectively. Four last terms in (A.3) vanish for $|i-j|$ odd; for the last two terms this follows since $\phi_{q,j}(t_k)$ has the same number of sign changes as $\cos\{\pi(k-1)t_{j,q}+\pi(q-1)/4\}$ (see Appendix A.1) and $\phi_{q,i}(t_k)$ is an even function for i odd and odd function for i even. If $|i-j|$ is even, then the second term in (A.3) is of order $O(n^{-\gamma-\alpha+1})$ since $\phi_{q,i}(x)=O(n^{-1/2})$ and the Fourier coefficients r_s of $\rho\in C^{\gamma,\alpha}$ are $O(s^{-\gamma-\alpha})$. Again, by $r_s=O(s^{-\gamma-\alpha})$, the inner sum of the third term is $O(l^{-\gamma-\alpha+1})$. Thus, for $|i-j|$ even, the third term is of order $O(n^{-1})$, if $2<\gamma+\alpha$, and of order $O(\log(n)\cdot n^{-\gamma-\alpha+1})$, if $1<\gamma+\alpha\le 2$. To see that the last two terms are of order $O(n^{-1})$ for $|i-j|$ even, use the Cauchy-Schwarz inequality, Lemma 2 and the result from Appendix A.1 that $\|\phi_{q,i}-\psi_{q,i}/\sqrt{n}\|_{L_2}=O(n^{-3/2})$, where $\psi_{q,i}$ is the Demmler-Reinsch basis of the Sobolev space $\mathcal{W}_\beta(M)$ given in (A.1), with the tail parts $T_i(x)$ vanishing exponentially away from the boundaries. Additionally, for the last term it is used that $\left|\left(\sum_{l=1}^{n}r_{|l-k|}\right)^2\right|=O(C)$ for each $k=1,\ldots,n$ since $\gamma+\alpha>1$. Finally, since ρ is Lipschitz continuous, it follows $\rho(\pi t_{j,q})=\rho(\pi t_j)+O(n^{-1})=\rho_j+O(n^{-1})$.

(ii) The proof is identical to (i) since only the decay properties of the Fourier coefficients of ρ are used. Note that $r_s=O(s^{-\beta})$ for $\rho\in\mathcal{W}_\beta(M)$.

A.6. Simplified representation for the likelihood

The log-likelihood of the model is given by

$$\ell_n(\lambda,q,\mathbf{R})=-\frac{n+1}{2}\log\{\mathbf{Y}^T\mathbf{R}^{-1}(\mathbf{I}_n-\mathbf{S})\mathbf{Y}+1\}+\frac{1}{2}\log|\mathbf{R}^{-1}(\mathbf{I}_n-\mathbf{S})|_+, \tag{7}$$

where

$$\mathbf{R}^{-1}(\mathbf{I}_n-\mathbf{S})=\mathbf{\Phi}_q\left\{\mathbf{I}_n+\mathrm{diag}(\lambda n\boldsymbol{\eta}_q)\mathbf{\Phi}_q^T\mathbf{R}\mathbf{\Phi}_q\right\}^{-1}\mathrm{diag}(\lambda n\boldsymbol{\eta}_q)\mathbf{\Phi}_q^T.$$

Suppose that Lemma 1 holds in such a way that

$$\{\mathbf{\Phi}_q^T\mathbf{R}\mathbf{\Phi}_q\}_{ij}=\rho(\pi t_j)\delta_{ij}+\mathbb{I}\{|i-j|\text{ is even}\}O(n^{-1}),\ i,j=q+1,\ldots,n,$$

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

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

where $\boldsymbol{\rho}=\{\rho(\pi t_1),\ldots,\rho(\pi t_n)\}^T$ and $\mathbf{E}$ is matrix with elements $E_{ij}=\mathbb{I}\{|i-j|\text{ is even}\}O(n^{-1})$. Note that the Lemmas give the values of the matrix for $i,j=q+1,\ldots,n$, but the values for the first $q+1$ rows and columns are at most $O(1)$ (follows from the summability of the rows and columns of $\mathbf{R}$ and that $\{\mathbf{\Phi}_q\}_{ij}=O(n^{-1/2})$).

Now,

$$\mathbf{R}^{-1}(\mathbf{I}_n-\mathbf{S})=\mathbf{\Phi}_q\left\{\mathbf{I}_n+\mathrm{diag}(\lambda n\boldsymbol{\eta}_q\boldsymbol{\rho})+\mathrm{diag}(\lambda n\boldsymbol{\eta}_q)\mathbf{E}\right\}^{-1}\mathrm{diag}(\lambda n\boldsymbol{\eta}_q)\mathbf{\Phi}_q^T$$
$$=\mathbf{\Phi}_q\left\{\mathbf{I}_n+\mathrm{diag}\left(\frac{\lambda n\boldsymbol{\eta}_q}{1+\lambda n\boldsymbol{\eta}_q\boldsymbol{\rho}}\right)\mathbf{E}\right\}^{-1}\mathrm{diag}\left(\frac{\lambda n\boldsymbol{\eta}_q}{1+\lambda n\boldsymbol{\eta}_q\boldsymbol{\rho}}\right)\mathbf{\Phi}_q^T.$$

Note that the matrix

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

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

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

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

Hence,

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

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

$$\left|(\mathbf{I}_n+n^{-1}\widetilde{\mathbf{E}})D_q\right|_+^2\le\prod_{j=q+1}^n\{1+O(n^{-1})\}=1+O(n^{-1}),$$

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

Writing

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

leads to

$$\begin{aligned}&\mathbf{Y}^T\mathbf{R}^{-1}(\mathbf{I}_n-\mathbf{S})\mathbf{Y}=\\&=\mathbf{B}^T\operatorname{diag}\left(\frac{\lambda n\boldsymbol{\eta}_q}{1+\lambda n\boldsymbol{\eta}_q\boldsymbol{\rho}}\right)\mathbf{B}-\mathbf{B}^T(\mathbf{I}_n+n^{-1}\widetilde{\mathbf{E}})^{-1}n^{-1}\widetilde{\mathbf{E}}\operatorname{diag}\left(\frac{\lambda n\boldsymbol{\eta}_q}{1+\lambda n\boldsymbol{\eta}_q\boldsymbol{\rho}}\right)\mathbf{B}.\end{aligned}$$

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 $O(n^{-1})$ for $i,j=q+1,\dots,n$ and zero for $i,j=1,\dots,q$. Hence,

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

for a matrix

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

which has elements of order $O(1/n)$. The second term in the last display is of order $O_p(1)$ by the Cauchy-Schwarz inequality. Finally,

$$\mathbf{Y}^T\mathbf{R}^{-1}(\mathbf{I}_n-\mathbf{S})\mathbf{Y}=\sum_{i=q+1}^n\frac{B_i^2\lambda n\eta_{q,i}}{1+\lambda n\eta_{q,i}}\{1+O(n^{-1})\}+O_p(1).$$

A.7. Consistency of the estimators

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

The estimator of the regression function. Note that since $\boldsymbol{Y}-\boldsymbol{f}_0\sim N(\mathbf{0},\sigma^2\boldsymbol{R}_0)$,

$$\mathbb{E}\|\hat{\boldsymbol{f}}_{\lambda,q,\boldsymbol{R}}-\boldsymbol{f}_0\|^2=\boldsymbol{f}_0^T(\boldsymbol{I}_n-\boldsymbol{S}_{\lambda,q,\boldsymbol{R}})^T(\boldsymbol{I}_n-\boldsymbol{S}_{\lambda,q,\boldsymbol{R}})\boldsymbol{f}_0+\sigma^2\operatorname{tr}\left(\boldsymbol{S}_{\lambda,q,\boldsymbol{R}}\boldsymbol{R}_0\boldsymbol{S}_{\lambda,q,\boldsymbol{R}}^T\right).$$

Then, if $\boldsymbol{A}\le\boldsymbol{B}$ means that $\boldsymbol{B}-\boldsymbol{A}$ is positive semi-definite, using the fact that the eigenvalues of $\boldsymbol{R}$ are on $[\delta,1/\delta]$, and the identity $\boldsymbol{R}^{-1}(\boldsymbol{S}_{\lambda,q,\boldsymbol{R}}^{-1}-\boldsymbol{I}_n)=\boldsymbol{S}_{\lambda,q,\boldsymbol{I}_n}^{-1}-\boldsymbol{I}_n$, leading to

$$\boldsymbol{I}_n-\boldsymbol{S}_{\lambda,q,\boldsymbol{R}}=\boldsymbol{S}_{\lambda,q,\boldsymbol{R}}\boldsymbol{R}\boldsymbol{R}^{-1}(\boldsymbol{S}_{\lambda,q,\boldsymbol{I}_n}^{-1}-\boldsymbol{I}_n)\le\frac{1}{\delta}(\boldsymbol{S}_{\lambda,q,\boldsymbol{I}_n}^{-1}-\boldsymbol{I}_n)\le\frac{1}{\delta^2}(\boldsymbol{S}_{\delta\lambda,q,\boldsymbol{I}_n}^{-1}-\boldsymbol{I}_n),$$

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

$$\operatorname{tr}\left(\boldsymbol{S}_{\lambda,q,\boldsymbol{R}}\boldsymbol{R}_0\boldsymbol{S}_{\lambda,q,\boldsymbol{R}}\right)\le\frac{1}{\delta^2}\operatorname{tr}\left(\boldsymbol{S}_{\lambda,q,\boldsymbol{R}}\boldsymbol{R}\boldsymbol{S}_{\lambda,q,\boldsymbol{R}}\right).$$

Then, again by the definition of the smoother that

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

Combining all of the above, conclude that

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

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 $\hat{\boldsymbol{f}}_0$ is the same as in the independent noise case. It is stressed again, though, that the finite sample performance of the smoother $\boldsymbol{S}_{\lambda,q,\boldsymbol{R}}$, as illustrated by the numerical simulations, is far superior to that of the smoother that ignores the correlation structure.

The estimator of the correlation. Note that since both $\hat{\boldsymbol{R}}$ and $\boldsymbol{R}_0$ are Toeplitz matrices, so is their difference. It is known that

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

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

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

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 $\boldsymbol{\Phi}_q^T(\boldsymbol{I}_n - \boldsymbol{S}_{\lambda,q,\tilde{\boldsymbol{\rho}}})\boldsymbol{Y}$ where $\tilde{\boldsymbol{\rho}}$ is identified with the resulting estimate of the correlation matrix. The quantity $(\boldsymbol{I}_n - \boldsymbol{S}_{\lambda,q,\tilde{\boldsymbol{\rho}}})\boldsymbol{Y}$ 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 $\boldsymbol{\rho}$,

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

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

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

Then, approximately

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

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

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

where ν_i are independent and have expectation 0 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 auto-correlations.) In particular, consistency of $\hat{\boldsymbol{R}}$ in 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.

References

Bickel, P., Levina, E., 2008a. Covariance regularization by thresholding. Ann. Stat. 36, 2577–2604.
Bickel, P., Levina, E., 2008b. Regularized estimation of large covariance matrices. Ann. Stat. 36, 199–227.
Xiao, H., Wu, W., 2012. Covariance matrix estimation for stationary time series. Ann. Stat. 40 (1), 466–493.
Purahmadi, M., 2011. Covariance estimation: the GLM and regularized perspectives. Stat. Sci. 26, 369–387.
Yang, Y., et al., 2001. Nonparametric regression with dependent errors. Bernoulli 7 (4), 633–655.
Cai, T., Zhang, C.-H., Zhou, H., 2010. Optimal rates of convergence for covariance matrix estimation. Ann. Stat. 38, 2118–2144.
Fan, J., Liao, Y., Liu, H., 2016. An overview of the estimation of large covariance and precision matrices. Econom. J. 19 (1).
Opsomer, J., Wang, Y., Yang, Y., 2001. Nonparametric regression with correlated errors. Stat. Sci. 16 (134–153).
Hart, J., 1994. Automated kernel smoothing of dependent data by using time series cross-validation. J. R. Stat. Soc. B 56, 529–542.
Altman, N., 1990. Kernel smoothing of data with correlated errors. J. Am. Stat. Assoc. 85, 749–759.
Hall, P., Keilegom, I.V., 2003. Using difference-based methods for inference in nonparametric regression with time series errors. J. R. Stat. Soc. B 65 (2), 443–456.
Kohn, R., Ansley, C., Wong, C., 1992. Nonparametric spline regression with autoregressive moving average errors. Biometrika 79, 335–346.
Wang, Y., 1998. Smoothing spline models with correlated random errors. J. Am. Stat. Assoc. 93, 341–348.
Chu, C.-K., Marron, J., 1991. Comparison of two bandwidth selectors with dependent errors. Ann. Stat. 19, 1906–1918.
Hall, P., Lahiri, S., Polzehl, J., 1995. On bandwidth choice in nonparametric regression with both short- and long-range dependent errors. Ann. Stat. 23, 1921–1936.
Chiu, S.-T., 1989. Bandwidth selection for kernel estimate with correlated noise. Stat. Probab. Lett. 8, 347–354.
Hurvich, C., Zeger, S., 1990. A frequency domain selection criterion for regression with autocorrelated errors. J. Am. Stat. Assoc. 85, 705–714.
Herrmann, E., Gasser, T., Kneip, A., 1992. Choice of bandwidth for kernel regression when residuals are correlated. Biometrika 79, 783–795.
Lee, Y.K., Mammen, E., Park, B.U., 2010. Bandwidth selection for kernel regression with correlated errors. Statistics 44 (4), 327–340.
Robinson, P.M., 1989. Nonparametric estimation of time-varying parameters. In: Hackl, P. (Ed.), Statistical Analysis and Forecasting of Economic Structural Change. Springer, Berlin, pp. 253–264.
Speckman, P., Sun, D., 2003. Fully Bayesian spline smoothing and intrinsic autoregressive priors. Biometrika 90 (2), 289–302.
Kotz, S., Nadarajah, S., 2004. Multivariate t-Distributions and Their Applications. Cambridge University Press.
Serra, P., Krivobokova, T., et al., 2017. Adaptive empirical bayesian smoothing splines. Bayesian Anal. 12 (1), 219–238.
Sniekers, S., van der Vaart, A., et al., 2015. Adaptive bayesian credible sets in regression with a gaussian process prior. Electron. J. Stat. 9 (2), 2475–2527.
Rousseau, J., Szabo, B., et al., 2020. Asymptotic frequentist coverage properties of bayesian credible sets for sieve priors. Ann. Stat. 48 (4), 2155–2179.
Yoo, W.W., Ghosal, S., et al., 2016. Supremum norm posterior contraction and credible sets for nonparametric multivariate regression. Ann. Stat. 44 (3), 1069–1102.
Wood, S., 2017. Generalized Additive Models: An Introduction with R. Chapman and Hall/CRC.
Amato, U., Antoniadis, A., De Feis, I., Goude, Y., 2017. Estimation and group variable selection for additive partial linear models with wavelets and splines. S. Afr. Stat. J. 51 (2), 235–272.
Rosales Marticorena, L.F., 2016. Empirical bayesian smoothing splines for signals with correlated errors: methods and applications. Ph.D. thesis. Georg-August-Universität Göttingen. http://hdl.handle.net/11858/00-1735-0000-0028-87F9-6.
Utreras Diaz, F., 1980. Pur ie choix du paramétre d'ajustement dans le lissage par fonctions spline. Numer. Math. 34, 15–28.
Fix, G.J., 1972. Effects of quadrature errors in finite element approximation of steady state, eigenvalue and parabolic problems. In: Aziz, A. (Ed.), The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations. Academic Press, pp. 525–556.
Fix, G.J., 1973. Eigenvalue approximation by the finite element method. Adv. Math. 10 (2), 300–316.
Cai, T.T., Ren, Z., Zhou, H.H., 2013. Optimal rates of convergence for estimating Toeplitz covariance matrices. Probab. Theory Relat. Fields 156 (1), 101–143.
Eggermont, P., LaRiccia, V., 2009. Maximum Penalized Likelihood Estimation. Springer Series in Statistics, vol. 2. Springer.