Adaptive Non-parametric Estimation of Mean and Autocovariance in Regression with Dependent Errors

Paulo Serra [1]   Tatyana Krivobokova [2]   Francisco Rosales [3]

18th December 2018

Abstract

We develop a fully automatic non-parametric approach to simultaneous estimation of mean and autocovariance functions in regression with dependent errors. Our empirical Bayesian approach is adaptive, numerically efficient and allows for the construction of confidence sets for the regression function. Consistency of the estimators is shown and small sample performance is demonstrated in simulations and real data analysis. The method is implemented in the R package *eBsc* that accompanies the paper.

Key words and phrases. Demmler-Reinsch basis, empirical Bayes, spectral density, stationary process.

[1]Department of Mathematics and Computer Science, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, the Netherlands; p.j.serra@tue.nl.

[2]Institute for Mathematical Stochastics, Georg-August-Universität-Göttingen, Germany.

[3]Department of Finance, Universidad del Pacífico, Peru.

1　Introduction

Consider a fixed design non-parametric regression model

$$Y_i = f(t_i) + \sigma\epsilon_i, \ \mathbb{E}(\epsilon_i) = 0, \ \sigma > 0, \ \ i = 1, \ldots, n$$

$$\text{cov}(\epsilon_i, \epsilon_j) = r(i - j) = r_{|i-j|} \in [-1, 1], \ \ i, j = 1, \ldots, n,$$

(1)

where $r(\cdot)$ denotes the autocorrelation function of the noise process, and $r_{|i|}$ denotes autocorrelation at lag i. Typically, design points $t_i \in \mathbb{R}$ are equidistant and represent time points. Functions f and r are unknown, but stationarity of $\{\epsilon_i\}_{i=1}^n$ and smoothness of f assumed. 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. Another instance is $\{Y_i\}_{i=1}^n$ being some time or space indexed stochastic process with a seasonal or other deterministic effect, described by f. In this case the focus is rather on estimation of the autocorrelation r.

Obviously, having a consistent estimator of f would deliver a reasonable estimator of r. Unfortunately, it is impossible to get a consistent estimator of f from a single replication without accounting for r as shown in (Hart and Wehrley, 1986). Consider a time series of hourly loads (kW) for a US utility (grey line in Figure 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 smooth function f. If no parametric assumptions on f is made and errors are treated as i.i.d., then all standard non-parametric estimators of f are heavily affected by the ignored dependence in the errors. In particular, automatic selectors of the smoothing parameter (e.g., 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 left plot in Figure 1.

This problem has been known for several decades; for an overview see (Opsomer et al.,
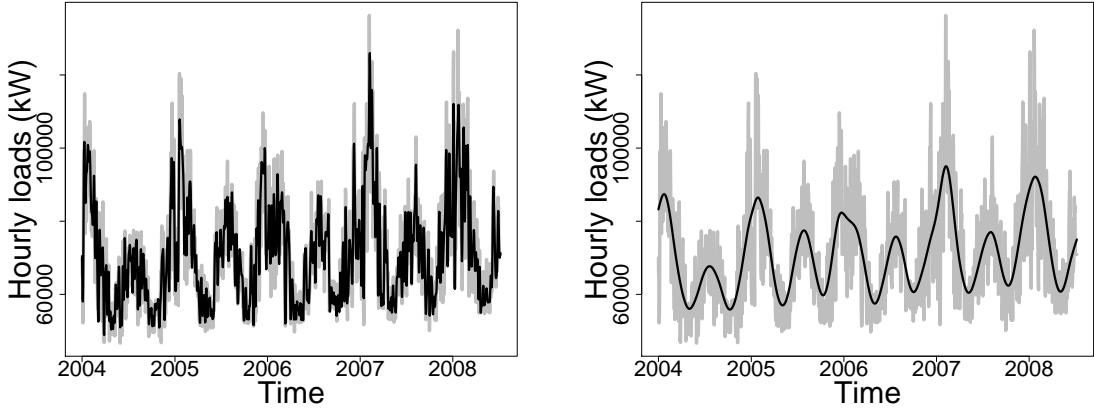
Figure 1: Hourly loads (kW) for a US utility estimated ignoring dependence in the data (left) and assuming that errors follow an autoregressive process of order one (right). Grey line shows the data and black lines show estimators. Estimators are obtained by functions `gam` (left) and `gamm` (right) of the R package *mgcv*.

2001). Since then, many approaches have been suggested for regression with dependent errors. The basic idea is to take smoothing parameter selection criteria that are geared towards independent data, and modify them to take into account the dependence in the errors.

One group of methods consists of making an explicit, parametric assumption on the correlation structure, e.g., assuming that errors follow 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 and 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 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), (Durban and Currie, 2003) and (Krivobokova and Kauermann, 2007). 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

3

Figure 1, assuming that the errors follow an $AR(1)$ process. This approach makes for computationally attractive algorithms, and the impact of the correlation structure on the estimation of the regression function is clearer. However, there are several drawbacks. First, the true correlation structure might be much more complex than any parametric one, which is especially problematic if the focus is on estimation of r. Moreover, if the correlation structure is indeed misspecified, then estimators of f are typically strongly affected and perhaps even inconsistent. Maximum likelihood methods do seem to be more robust to misspecification, see (Krivobokova and Kauermann, 2007). More importantly, in practice there is no information on r is available and some strategies for the selection of a parametric model for r (its type and order) are needed.

Another group of methods makes no parametric assumption on r and tries to eliminate the influence of the error dependence on the smoothing parameter in different ways. For example, (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 autocovariances, that depend on unknown parameters linked to assumptions on the error process. Using the method of (Herrmann et al., 1992) for estimating the hourly loads gives an estimator of f very similar to the one obtained with function `gamm` in Figure 1.

All together, both groups of methods are rather focused on obtaining a reasonable estimator for f, while r is treated as a nuisance parameter. Moreover, all these methods depend on unknown parameters that require some knowledge of r. Hence, there is a need for a fully automatic (independent on any unknown parameters) method for non-parametric estimation of both f and r simultaneously.

Note that estimation of the correlation structure of a data vector with a known mean is already a challenging problem. It is important in multivariate analysis in problems such as 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 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 relevant one for our problem. Irrespectively of the framework, natural (moment) estimators for correlation matrix of the observed vectors are not consistent (in, e.g., operator-norm) and regularisation is needed (e.g., banding, tapering, thresholding) to ensure positive definiteness and consistency of the estimate. The minimax rates of convergence of these estimators can be shown to depend on the decay of the autocorrelation function as the lag increases; cf. (Yang et al., 2001), (Cai et al., 2010), (Purahmadi, 2011), (Xiao and Wu, 2012), see also (Fan et al., 2016).

In this paper we develop a likelihood based method that provides an adaptive estimator of the regression function f, as well as estimators of the noise level σ and autocorrelations r. Starting from an arbitrary guess for the autocorrelations we iteratively update all other parameters until convergence. Furthermore, our empirical Bayesian framework provides a computationally attractive way of construction confidence sets for the regression function that take the correlation structure into account.

There are quite a few novel points in our work. Contrary to other approaches in the literature, our method is completely automatic so that no tuning parameters need to be set by the user, and it is also fully non-parametric. To the best of our knowledge, our estimate of the autocorrelations is also novel – we use spline smoothers to estimate both the regression function and the spectral density of the noise process adaptively. The autocorrelations are then reconstructed from the estimate of the spectral density, rather

than by tapering or thresholding some empirical estimate.

This paper is structured as follows. In Section 2 we introduce our estimator, Section 3 contains simulation results, Section 4 presents two real data examples, and Section 5 closes the paper with some conclusions. All of the asymptotics and respective proofs, as well as auxiliary results are gathered in the Appendix.

2   Construction of the estimators

Assume that in model (1) the design points are given by $t_i = (i-1)/(n-1)$, regression function f is a function from a Sobolev space $\mathcal{W}_\beta$, $\beta > 1/2$ and the noise terms ϵ_i are sampled from a stationary, Gaussian noise process with zero mean and variance $\sigma^2 > 0$. Setting $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$, $\boldsymbol{f} = f(\boldsymbol{t}) = \{f(t_1), \ldots, f(t_n)\}^T$, $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T$ we can write this model as

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

The entries of $\boldsymbol{\mathcal{R}}$ are $\mathcal{R}_{i,j} = r_{|i-j|}$, so that $\boldsymbol{\mathcal{R}}$ is a Toeplitz matrix, such that $r_{|i-j|} = 0$ for $|i-j| > m$, for some fixed $m \in \mathbb{N}$ independent on n. We further assume that for some $0 < \delta < 1$, $\boldsymbol{\mathcal{R}} \in \mathcal{M}_{n,\delta}$, where $\mathcal{M}_{n,\delta}$ denotes the space of $n \times n$ matrices whose eigenvalues all lie on the interval $[\delta, 1/\delta]$. Moreover, the spectral density of $\{\epsilon_i\}_{i=1}^n$ is at least Lipschitz continuous.

We estimate f using a smoothing spline, i.e., we find $\hat{f}$ 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 correlation matrix $\boldsymbol{R} \in \mathcal{M}_{n,\delta}$. Note that β and $\boldsymbol{\mathcal{R}}$ are unknown in practice and are replaced by some "working" values in (2). Subsequently, β, λ and $\boldsymbol{\mathcal{R}}$ 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 $\boldsymbol{S}$ we choose the so-called Demmler-Reinsch basis of the natural spline space of degree $2q - 1$, which is defined in Section B.1 in the Appendix,

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

where $\boldsymbol{\Phi}_q$ is the $n \times n$ 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. Subsequently, to keep the notation simple we omit the dependence on these parameters, except if these are set to a particular value. Note that estimation of β from the data makes our estimator adaptive to the unknown smoothness of the signal; see (Serra and Krivobokova, 2017).

2.1 Bayesian interpretation

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. We start by endowing $\boldsymbol{f}$ 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 is often used for Bayesian smoothing splines and 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. Note that 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 B.2 for the derivation of the identity. It is well-known that the corresponding posterior distribution for $\boldsymbol{f}|(\boldsymbol{t}, \lambda, \sigma^2, \boldsymbol{R})$ has a mean equal to the smoothing spline estimator $\hat{\boldsymbol{f}} = \boldsymbol{S}\boldsymbol{Y}$ and variance $\sigma^2 \boldsymbol{S}\boldsymbol{R}$; 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 is a multivariate t-distribution with $n + 1$ degrees of freedom, mean $\hat{\boldsymbol{f}} = \boldsymbol{S}\boldsymbol{Y}$, and scale $\hat{\sigma}^2 \boldsymbol{S}\boldsymbol{R}$.

It remains to estimate unknown λ, q and $\boldsymbol{R}$. To do so, we employ the empirical Bayes approach and estimate these parameter from the marginal distribution of $\boldsymbol{Y}$, given $(\boldsymbol{t}, \lambda, q, \boldsymbol{R})$, which is a multivariate t-distribution whose density is

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

see (Kotz and Nadarajah, 2004). It remains to set the parameters a and b. From our analysis we concluded that asymptotically the choice of parameters a and b is irrelevant (as long as a and b are $o(n)$ and lead to proper a prior and marginal distributions). Hereon after, to simplify the log-likelihood, we set $b = 1/2$ and $a = (q + 1)/2$ and obtain

$$\ell_n(\lambda, q, \boldsymbol{R}) = -\frac{n+1}{2} \log \left\{ \boldsymbol{Y}^T \boldsymbol{R}^{-1}(\boldsymbol{I}_n - \boldsymbol{S})\boldsymbol{Y} + 1 \right\} + \frac{1}{2} \log \left| \boldsymbol{R}^{-1}(\boldsymbol{I}_n - \boldsymbol{S}) \right|_+, \qquad (5)$$

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

$$\hat{\sigma}^2 = \frac{\boldsymbol{Y}^T \boldsymbol{R}^{-1}(\boldsymbol{I}_n - \boldsymbol{S})\boldsymbol{Y} + 1}{n + 1}.$$ (6)

2.2   Estimating equations and algorithm

By differentiating the log-likelihood (5) (see identities for the derivatives in Appendix B.2) with respect to the smoothing parameter λ, values of the spectral density $\rho_i = \rho(\pi t_i)$, $i = 1, \ldots, n$ that correspond to $\boldsymbol{R}$ and penalty order q, we obtain the following estimating equations for these parameters.

$$T_\lambda(\lambda, q, \boldsymbol{R}) = \frac{1}{n_{\lambda,q}} \left[\boldsymbol{Y}^T \boldsymbol{R}^{-1}(\boldsymbol{I}_n - \boldsymbol{S})\boldsymbol{S}\boldsymbol{Y} - \hat{\sigma}^2 \left\{ \mathrm{tr}\left(\boldsymbol{S}\right) - q \right\} \right],$$

$$T_q(\lambda, q, \boldsymbol{R}) = \frac{1}{n'_{\lambda,q}} \left[\boldsymbol{Y}^T \boldsymbol{R}^{-1}(\boldsymbol{I}_n - \boldsymbol{S})\boldsymbol{\Phi}_q \mathrm{diag}\left[n\boldsymbol{\eta}_q \circ \log(n\boldsymbol{\eta}_q) \right] \boldsymbol{\Phi}_q^T \boldsymbol{Y} - \hat{\sigma}^2 \mathrm{tr}\left\{ \boldsymbol{n}\boldsymbol{\eta}_q \circ \log(n\boldsymbol{\eta}_q) \right\} \right],$$

$$T_{\rho_i}(\lambda, q, \boldsymbol{R}) = \mathrm{tr}\left[\boldsymbol{R}^{-1} \frac{\partial \boldsymbol{R}}{\partial \rho_i} \boldsymbol{R}^{-1} \left\{ (\boldsymbol{I}_n - \boldsymbol{S})\boldsymbol{Y}\boldsymbol{Y}^T(\boldsymbol{I}_n - \boldsymbol{S})^T - \hat{\sigma}^2(\boldsymbol{I}_n - \boldsymbol{S})\boldsymbol{R} \right\} \right], \quad i = 1, \ldots, n,$$

where $\circ$ denotes the Hadamard product and $\hat{\sigma}^2$ as given in (6). The scaling factors in the first two equations are given by $n_{\lambda,q} = \lambda^{-1/(2q)} + n\lambda^{1/(2q)}$ and $n'_{\lambda,q} = n_{\lambda,q}(\log \lambda)^2$, $\lambda > 0$, $q \in \mathbb{N}$. These equations might be difficult to evaluate and in practice we use approximate expressions, which are given in Appendix B.5.

These estimating equations need to be solved simultaneously. In practice we proceed as follows. For each fixed q we start with an initial guess $\hat{\boldsymbol{R}}^{(0)}$ (typically just an identity matrix), 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 be accomplished by introducing a suitable prior on $\boldsymbol{R}$, which then acts as a penalty term

in the posterior. For simplicity, we perform a two-step procedure 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 $\boldsymbol{I}_n$. 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 $\hat{\rho}$, 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 B.4. The summary of the estimation procedure is given in Algorithm 1.

---

for q *in Q_n* **do**

    set $j = 1$ and $\hat{\boldsymbol{\rho}}^{(0)}$

    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 $\tilde{\boldsymbol{\rho}}^{(j)}$ to a solution of $T_{\boldsymbol{\rho}}(\lambda^{(j)}, q, \hat{\boldsymbol{\rho}}^{(j-1)}) = 0$;

        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 $\hat{r}_k = n^{-1}\sum_{l=1}^{n} \cos\left(k\pi\{l - 1\}/\{n - 1\}\right)\hat{\rho}_l$;

Algorithm 1: Recursive estimation procedure.

Here Q_n denotes collections of values for q. The stopping rule is standard: after each iteration we compare the change in the value of the estimate of λ and the norm of the change in the estimate of $\boldsymbol{\rho}$; if these fall bellow a threshold, then we stop iterating. The consistency of estimators $\hat{\boldsymbol{R}}$, $\hat{\lambda}$ and $\hat{q}$ is addressed in Appendix A.

2.3 Confidence sets

Once estimates $\hat{\lambda}$, $\hat{\boldsymbol{R}}$, $\hat{q}$, and $\hat{\sigma}^2$ for respectively λ, $\mathcal{R}$, q, 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 $\boldsymbol{f}$:

$$\hat{\Pi}^{\boldsymbol{f}}\big(\,\cdot\mid\boldsymbol{Y}\big)=\Pi^{\boldsymbol{f}}\big(\,\cdot\mid\boldsymbol{Y},\hat{\sigma}^2,\hat{\lambda},\hat{q},\hat{\boldsymbol{R}}\big)=\Pi^{\boldsymbol{f}}\big(\,\cdot\mid\boldsymbol{Y},\sigma^2,\lambda,q,\boldsymbol{R}\big)\big|_{\left(\lambda,q,\boldsymbol{R},\sigma^2\right)=\left(\hat{\lambda},\hat{q},\hat{\boldsymbol{R}},\hat{\sigma}^2\right)}. \quad (7)$$

Given $\boldsymbol{Y}$, this is just a $t_{n+1}\big(\hat{\boldsymbol{f}}\,,\hat{\sigma}^2\hat{\boldsymbol{S}}\,\hat{\boldsymbol{R}}\big)$ distribution, which is centred at the spline estimate $\hat{\boldsymbol{f}}$, and whose covariance matrix depends on the (random) smoother $\hat{\boldsymbol{S}}$ – which is just the smoother $\boldsymbol{S}$ with $(\hat{\lambda},\hat{q},\hat{\boldsymbol{R}})$ plugged in for $(\lambda,q,\boldsymbol{R})$ – and the estimates $\hat{\sigma}^2$ and $\hat{\boldsymbol{R}}$. From this we can easily construct a credible set for the regression function.

If $\boldsymbol{f}$ is distributed according to $\Pi^{\boldsymbol{f}}(\,\cdot\mid\boldsymbol{Y},\lambda,q,\boldsymbol{R},\sigma^2)$, then $\boldsymbol{f}-\hat{\boldsymbol{f}}$ is distributed according to $t_{n+1}(\boldsymbol{0},\sigma^2\boldsymbol{S}\boldsymbol{R})$. This means that $\|\boldsymbol{f}-\hat{\boldsymbol{f}}\|_2^2$ has the same distribution as the random variable $\sigma^2\boldsymbol{Z}^T\boldsymbol{S}\boldsymbol{R}\boldsymbol{Z}/U$, where $\boldsymbol{Z}\sim N(\boldsymbol{0},\boldsymbol{I}_n)$, and $U\sim\mathcal{X}_{n+1}^2$, with $\boldsymbol{Z}$ and $\boldsymbol{U}$ independent. From this it follows that there exits a (known) sequence of quantiles $s_n(\lambda,q,\boldsymbol{R})$ such that for every $n,\lambda,q,\boldsymbol{R}$, and each $\alpha\in(0,1)$,

$$\Pi\Big\{\boldsymbol{f}:\|\boldsymbol{f}-\hat{\boldsymbol{f}}\|_2^2\leq\sigma^2\,s_n(\lambda,q,\boldsymbol{R})\mid\boldsymbol{Y},\lambda,q,\boldsymbol{R},\sigma^2\Big\}=1-\alpha.$$

We can then define the sets

$$\hat{C}_n(L)=\Big\{\boldsymbol{f}:\|\boldsymbol{f}-\hat{\boldsymbol{f}}\|_2^2\leq L\,\hat{\sigma}^2\,s_n(\hat{\lambda},\hat{q},\hat{\boldsymbol{R}})\Big\},\qquad L\geq 1, \quad (8)$$

which (simply by definition of the quantile sequence s_n) satisfy

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

and are therefore a credible set – a small, high probability region of the empirical posterior. These sets can be easily simulated by sampling a large number of draws from the posterior and keeping the $(1 - \alpha)$-fraction of them that are the closest (in ℓ_2-norm) to $\hat{\boldsymbol{f}}$. These will give a good visual representation of the uncertainty in the estimate of the regression function. One can of course construct confidence sets for f (or functionals of f) in terms of other norms (such as ℓ_∞- norm) by simply adjusting the quantile sequence s_n. However, the coverage properties of such sets is still not so well understood outside the setting of regression with Gaussian i.i.d. noise, and of the Gaussian white noise model.

The construction above can be found in (Szabó et al., 2015), and is also used in (Serra and Krivobokova, 2017), where the behaviour of this set as a confidence set is studied, under a frequentist assumption on the distribution of the data. In the latter paper, for the case where $\boldsymbol{\mathcal{R}} = \boldsymbol{I}_n$, it is shown (cf. Serra and Krivobokova, 2017, Theorem 3) that this set has two important properties if L is taken appropriately large. Firstly, the set contains the true underlying regression function with probability converging to 1, uniformly over a large subset of functions; secondly, 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. In other words, uniformly over a large set of regression function, this credible set covers the true regression function (honest coverage), and has a size (radius) that adapts to the smoothness of the underlying regression function (adaptive coverage).

The asymptotic results from Appendix A show that the presence of correlation has a relatively simple scaling effect on the smoothing parameter. So, in principle, the theoretical results of (Serra and Krivobokova, 2017) can be extended to cover the correlated noise case as well, by simply picking larger values of the multiplier L, but this problem will be studied in more generality elsewhere. In this paper, we focus instead on the implementation, and numerical aspects of the procedure.

3 Numerical simulations

In this section we investigate the small sample performance of our estimation procedure and compare it to two alternatives. Among approaches that make a parametric assumption on $\mathcal{R}$ we consider the well-established method based on splines that is implemented in the statistical software R in the function `gamm` of package *mgcv*, see (Wood, 2017). Among approaches that make no parametric assumption on $\mathcal{R}$ we consider the method by (Herrmann et al., 1992). This kernel based method uses sample autocorrelation estimators to improve bandwidth selection and is developed under assumptions of m-dependence in the residuals. However, the authors state that the method still works if the residuals satisfy "some mixing conditions".

Our simulation set up is as follows. We consider three regression functions

$$
\begin{aligned}
f_1(x) &= \sum_{i=3}^{n} \psi_{3,i}(x)\{\pi(i-1)\}^{-3}\cos(2i) \\
f_2(x) &= \sum_{i=4}^{n} \psi_{5,i}(x)\{\pi(i-1)\}^{-5}\cos(2i) \\
f_3(x) &= 2\sin(4\pi x),
\end{aligned}
$$

where $\psi_{j,i}$, $j \in \{3,5\}$ is the i-th Demmler-Renisch basis of $\mathcal{W}_j$ given explicitly in Appendix B.1, $n = 500$ and x is a sequence of n equally spaced data on $[0,1]$. All three 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 $M = 500$. We consider 9 types of the residual processes: independent and identically distributed, an $AR(1)$ process with the parameters $\phi \in \{-0.9, -0.5, 0.5, 0.9\}$, a $MA(1)$ process with the parameters $\theta \in \{-0.5, 0.5\}$, an ARMA(2,2) process with $\phi = (0.7, -0.4)$ and $\theta = (-0.2, 0.2)$ and a zero-mean Gaussian process with the correlation matrix with (i,j) entries $\cos(6.5j)\exp(-|i-j|/20)$.
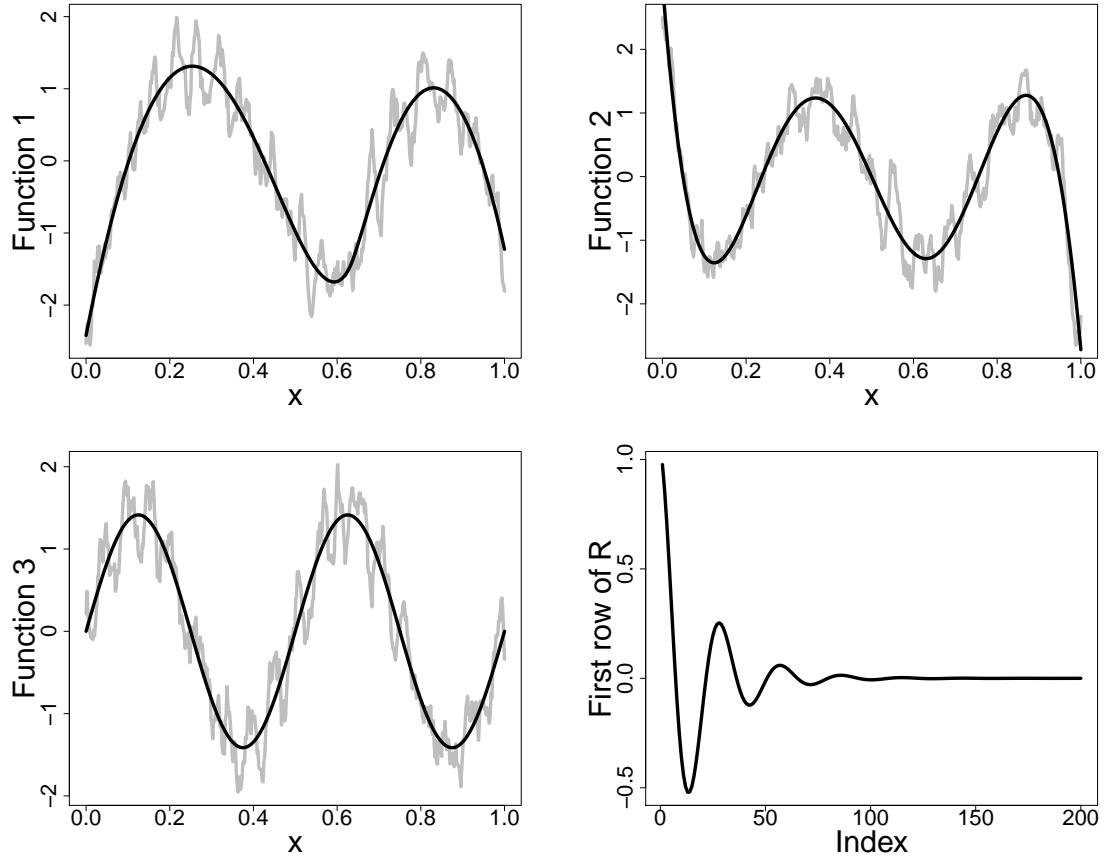
13

Figure 2: Regression functions (black lines) f_1 (top left), f_2 (top right) and f_3 (bottom left) with the data (grey lines). The data are simulated adding a zero-mean Gaussian process noise with the correlation matrix (first row) shown in the bottom right plot.

The top row and bottom left plot in Figure 2 show all three regression functions (black lines). The data (grey lines) are simulated using a zero-mean Gaussian noise process with the correlation matrix $\mathcal{R}_{i,j} = \cos(6.5 \cdot j) \exp(-|i - j|/20)$. Elements $\mathcal{R}_{1,j}$, $j = 1, \ldots, 200$ are shown in the bottom right plot of Figure 2. All three functions are to be estimated with the method of (Herrmann et al., 1992) (further denoted by HER), with function gamm (denoted by GAM) and with our approach (denoted by BAS).

Before we summarise the simulation results, we demonstrate how our 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 Figure 3 shows estimating equation
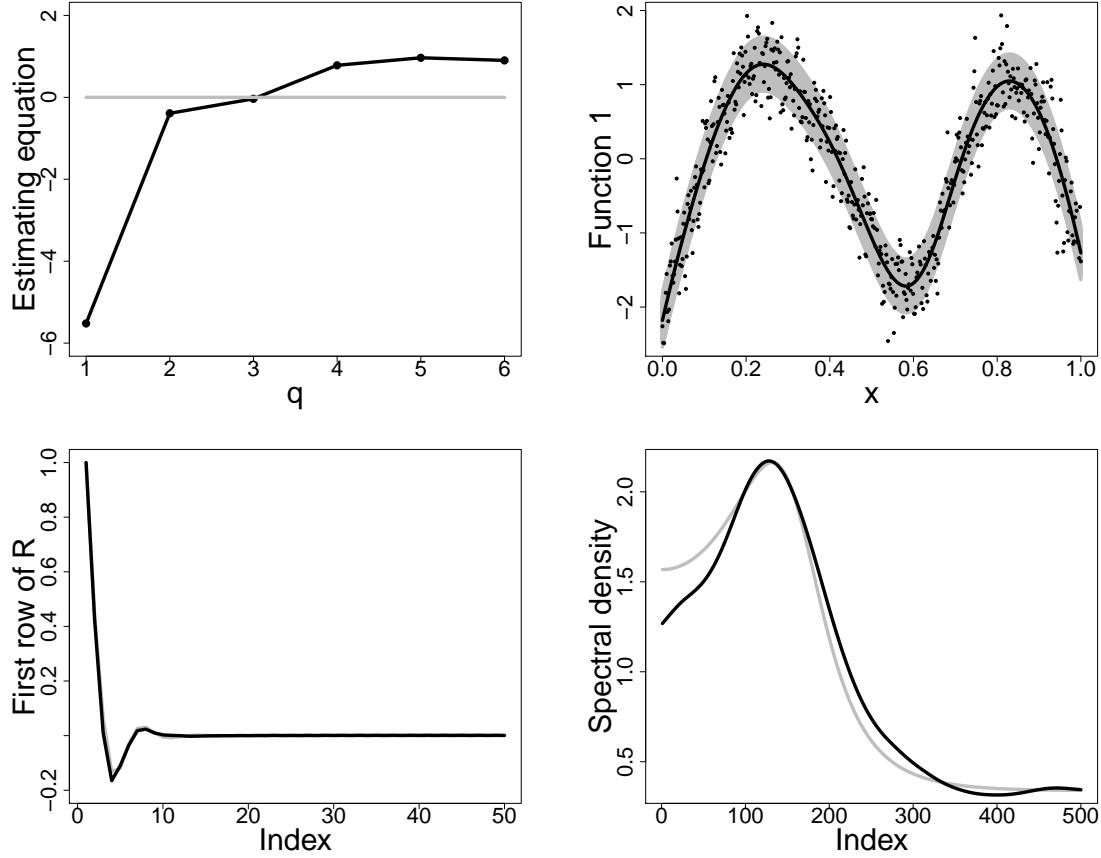
14

Figure 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 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).

$T_q(\hat{\lambda}, q, \hat{\boldsymbol{R}})$, which is zero at q which is very close to 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 Figure 3, together with the data (black dots) and confidence bands (grey area) constructed as described in Section 2.3. The estimate for $\mathcal{R}_{1,j}$, $j = 1, \ldots, 50$ is shown in the bottom left plot of the same figure in black, which nearly coincides with the truth 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 3. The true spectral density is shown in the same plot in grey. Hence, our method allows for simultaneous, fully automatic and non-parametric estimation of q, λ and $\mathcal{R}$. However, in the simulation

15

below we fixed $q = 2$ to ensure comparability of all three estimators, which all should have the same rate of convergence $n^{-2/5}$.

For the method of (Herrmann et al., 1992) the parameter m is set according to method (i) described in that paper (p. 787). Namely, we choose m to be the largest integer such that $\hat{h}_m \geq 6/5\, \hat{h}_{m-1}$ and $m \leq 0.2\sqrt{n} \approx 4.5$ for $n = 500$, 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 our experiments we noticed that the influence of m on the estimator of f is not very pronounced, but it does affect the estimator of the autocovariance quite strongly. There is no simple data-driven approach to choose m such that both mean and autocovariance estimators are optimal in some sense. Hence, we expect that HER performs comparably to BAS in estimation of f, but should be inferior in estimation of R. In our implementation we used function `glkerns` of package *lokern* by Eva Herrmann for the estimation of f with the second order kernel and for the estimation of f'' with the fourth order kernel.

In the function `gamm` we used low-rank splines with number of knots $n/4 = 125$, B-spline basis of degree 3, penalisation order $q = 2$ and specified the correlation structure according to the true dependence structure in the residuals for the first 8 residual processes. Since the parametric estimators of $\mathcal{R}$ have faster convergence rates, we expect that GAM will outperform our non-parametric estimator for $\mathcal{R}$ with BAS. Note that in practice no information on the true correlation structure is given and some model validation and selection procedures need to be employed. Since the 9th process is a non-parametric one, we set the correlation structure in the call of `gamm` to be an $AR(1)$ process, mimicking the real situation in practice where no information on the residual process is available.

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

16

we calculate

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

$$A(\hat{R}_j) \;=\; \frac{1}{Mn}\sum_{k=1}^{n}\sum_{i=1}^{M}\{\mathcal{R}_j(k) - \hat{R}_{j,i}(k)\}^2, \;\; j=1,2,3,$$

Here $\hat{f}_{j,i}$ denotes an estimator of f_j in ith Monte Carlo run with the residuals following one of the nine processes. Similarly, $\mathcal{R}_j(k)$ denotes the k-th entry of the first row of one of nine 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.

Consider first a parametric dependence structures of the residuals (first eight types). Non-parametric methods HER and BAS perform similarly for the estimation of the regression function, with HER performing better for MA-processes and BAS being better for AR-processes. As expected, performance of HER for the autocovariance estimation is uniformly worse (except for an $AR1(0.5)$ process). The GAM approach performs very much like BAS in estimating the regression function, but, as expected, outperforms in covariance matrix estimation. Of course, this comes at the cost of using the true specification for the dependence process, which is unavailable in practice. In case of a non-parametric dependence in the residuals (the ninth, Gaussian process), BAS clearly outperforms both GAM and HER. Overall, BAS shows very good small sample properties, while being fully automatic and non-parametric.

The last column of Table 1 shows the proportion of correctly estimated q out of M. In general, q is estimated more reliably for smaller qs (for the analytic function f_3 we set $q = 6$ if $T_q(\hat{\lambda}, q, \hat{\boldsymbol{R}})$ remains negative as it should). Also, high autoregressive dependence in the data and high smoothness of the signal make it more difficult to identify q. Note also that $n = 500$ is rather moderate and performance of $T_q(\hat{\lambda}, q, \hat{\boldsymbol{R}})$ becomes better with the growing sample size.

17

f_1							
	$A(\hat{f}_1)$			$A(\hat{R}_1)$			$\hat{q}=3$
Correlation	BAS	GAM	HER	BAS	GAM	HER	
$i.i.d.$	3.187	**3.182**	3.840	0.000	0.000	0.000	0.870
$AR1(-0.9)$	1.170	**1.130**	3.889	0.874	**0.177**	5.885	0.890
$AR1(-0.5)$	1.732	**1.674**	2.924	0.033	**0.010**	0.081	0.940
$AR1(0.5)$	10.737	10.262	**9.786**	0.063	**0.012**	0.037	0.658
$AR1(0.9)$	210.067	**150.041**	328.900	3.709	**0.426**	5.088	0.010
$MA1(-0.5)$	1.464	**1.126**	2.057	0.015	**0.001**	0.064	0.572
$MA1(0.5)$	8.836	6.251	**6.059**	0.040	**0.001**	0.037	0.638
$AR2MA2$	6.130	5.897	**5.598**	0.038	**0.032**	0.047	0.802
GP	**2.192**	5.656	16.396	**1.420**	15.714	5.565	0.892

f_2							
	$A(\hat{f}_2)$			$A(\hat{R}_2)$			$\hat{q}=5$
Correlation	BAS	GAM	HER	BAS	GAM	HER	
$i.i.d.$	4.559	**4.506**	6.143	0.000	0.000	0.000	0.332
$AR1(-0.9)$	2.185	**1.585**	2.380	0.973	**0.183**	5.877	0.456
$AR1(-0.5)$	2.428	2.353	**1.661**	0.033	**0.009**	0.077	0.422
$AR1(0.5)$	15.007	14.083	**9.033**	0.075	**0.013**	0.032	0.122
$AR1(0.9)$	267.605	**191.323**	338.288	4.260	**0.572**	5.067	0.000
$MA1(-0.5)$	1.882	1.695	**1.252**	0.015	**0.004**	0.065	0.166
$MA1(0.5)$	12.844	8.849	**5.356**	0.041	**0.001**	0.033	0.228
$AR2MA2$	8.477	8.138	**4.682**	0.043	**0.034**	0.048	0.226
GP	**3.907**	7.028	14.673	**1.548**	17.964	5.663	0.330

f_3							
	$A(\hat{f}_3)$			$A(\hat{R}_3)$			$\hat{q}=6$
Correlation	BAS	GAM	HER	BAS	GAM	HER	
$i.i.d.$	3.270	**3.264**	4.058	0.000	0.000	0.000	0.422
$AR1(-0.9)$	**1.093**	1.186	2.735	0.902	**0.189**	5.874	0.800
$AR1(-0.5)$	1.832	**1.753**	1.981	0.032	**0.009**	0.072	0.500
$AR1(0.5)$	11.051	10.330	**9.217**	0.080	**0.013**	0.033	0.024
$AR1(0.9)$	**215.062**	224.555	325.371	3.746	**0.803**	5.064	0.000
$MA1(-0.5)$	1.502	**1.142**	1.487	0.016	**0.001**	0.061	0.234
$MA1(0.5)$	8.353	6.688	**5.821**	0.029	**0.001**	0.036	0.118
$AR2MA2$	6.535	6.224	**5.360**	0.036	**0.034**	0.046	0.108
GP	**2.495**	6.008	15.125	**1.461**	16.095	5.662	0.312

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

4 Examples

4.1 Exchange rates of Russian Ruble

The first example is on the official exchange rates of Russian Ruble to Euro and US Dollars. The data are freely available from the Central bank of Russian Federation under `https://www.cbr.ru/eng/currency_base/dynamics/`. We obtained the data on prices of Euro and US Dollar in Rubles for the time period between 03 June 2008 and 01 June 2018 and calculated the dual currency basket weighing US Dollar prices by 0.55 and Euro prices by 0.45, i.e., $B = 0.55 \, USD + 0.45 \, EUR$.
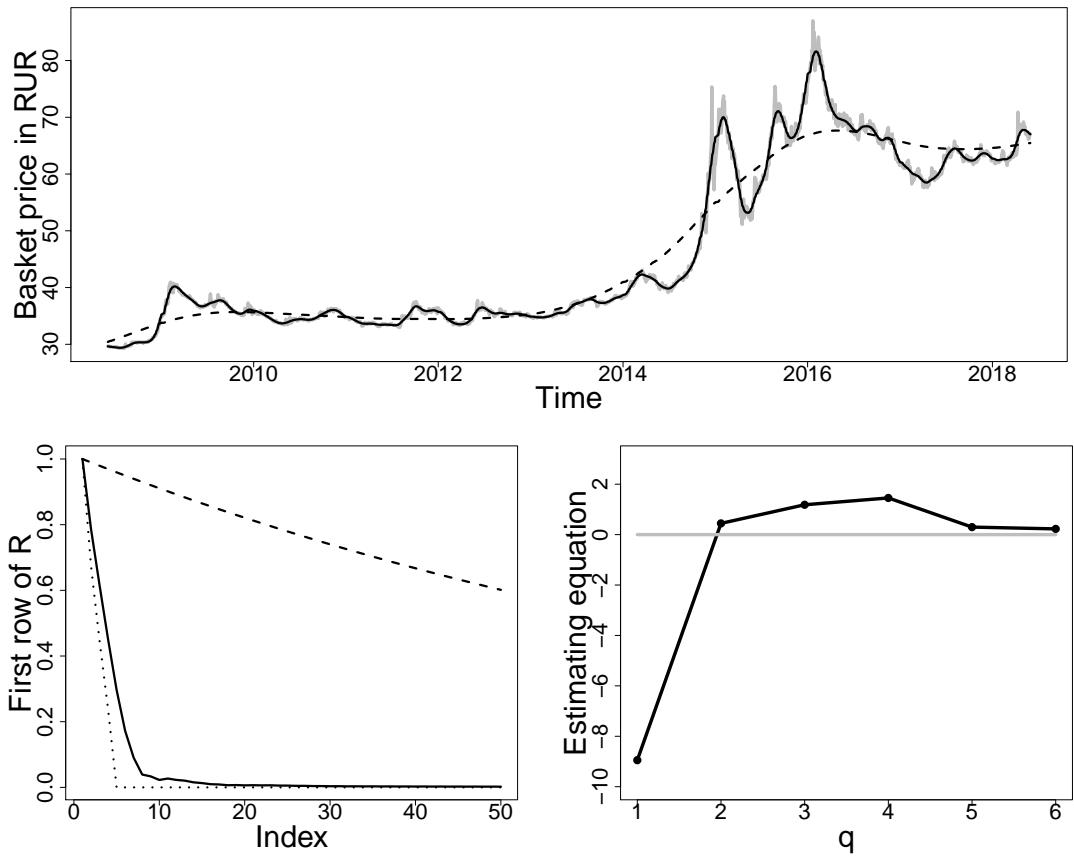

Figure 4: Estimators of the mean (top), first row of the correlation matrix (first 50 values, bottom left) and the estimating equation for the smoothness degree q (bottom right) of the dual currency basket. Data are shown in grey, BAS estimators are the black solid lines, HER estimators are doted lines and GAM estimators are dashed lines.

The top plot in Figure 4 shows the data in grey. The signal-to-noise ratio of the data is very high and the number of observations is large: $n = 2477$. Fitting the data with our algorithm delivers $\hat{q} = 2$ (bottom right plot of Figure 4). The corresponding mean estimator is shown as a black bold line on the top plot of Figure 4 and the estimator of the first row of the correlation matrix (first 50 values) is shown on the bottom left plot of Figure 4. Since q is estimated to 2 (bottom right plot of Figure 4) one can expect that both competing methods HER and GAM should deliver similar results. Method HER indeed delivers the mean estimator visually indistinguishable from the one obtained with BAS (with m set as described in the previous section). The dotted line on the bottom left plot shows the corresponding sample autocorrelation. We observed that the shape of the estimated autocorrelation appears is more influenced by m than the mean estimator. Recall that the choice of m is not data driven. Method GAM has been applied assuming an $AR(1)$ process for the residuals. The estimate of the mean seems to be way too smooth with the most variation moved to the residuals; the corresponding estimated autocorrelation decays extremely slowly. The model validation via residual analysis suggests that the estimators HER and BAS are more appropriate.

4.2   Hourly loads at a US facility

In the second example we consider data from the load forecasting track of the Global Energy Forecasting Competition 2012 (`http://www.drhongtao.com/gefcom/2012`). 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 consider each time series of the 20 zones, as well as their mean, in order to make a one week out-of-sample forecast, as well as backcast certain values set to be missed within the observational period. Here we are not interested in forecasting the data, but rather understanding their structure. We consider the mean over all 20 zones over the whole time period – all together

$n = 1650$ observations. Missing values were imputed using R package `Hmisc`; omitting these missing values lead to the same estimators and same conclusions.
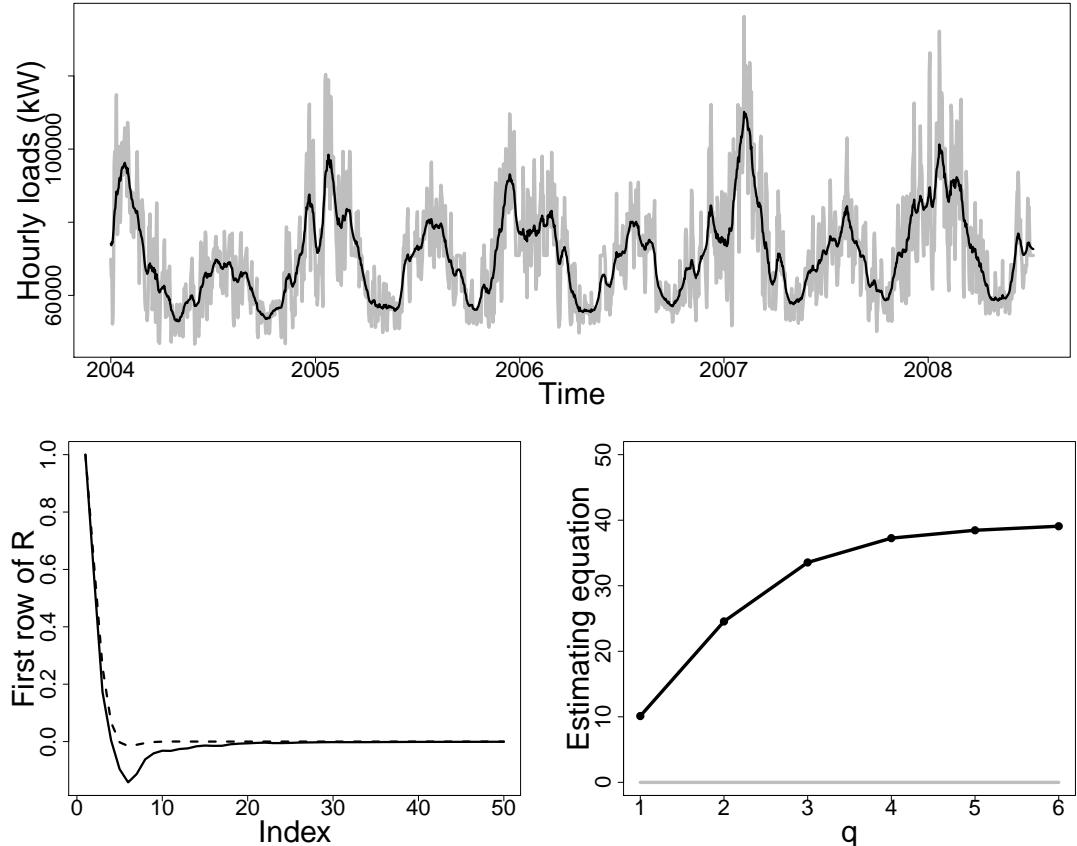

Figure 5: Estimators of the mean (top), first row of the correlation matrix (first 50 values, bottom left) and the estimating equation for the smoothness degree q (bottom right) of the hourly loads of a US facility. Data are shown in grey, BAS estimators are the black solid lines, GAM estimators are dashed lines.

Estimating with BAS delivered a positive and increasing function in the estimating equation for q. This suggests that the mean function has either 1 continuous derivative or is even less smooth. In any case, this suggests a fit with $q = 1$ to be more appropriate than a fit with $q = 2$. This agrees with the nature of the data, where many peak loads might happen; 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 Figure 5 as black line. The

corresponding autocorrelations are shown as solid black line in the bottom left top. Since the HER method is defined only for even order kernels, we can not obtain the fit that would be comparable with the fit by BAS with $q = 1$. Setting $q = 1$ in GAM with an $AR(2)$ process for the residuals results in an estimator of the mean that is very close to the one obtained with BAS, having only slightly less pronounced peaks (not visible in the plot). The corresponding autocorrelation estimator is shown as a dashed line in the bottom left plot of Figure 5.

Estimation using BAS with $q = 2$ and by HER method 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 Figure 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.

5   Conclusions

Correlation is ubiquitous in applications, particularly when data collected sequentially over time, and ignoring this can have severe consequences for inference. Covariance is typically taken into account by either making parametric assumptions on the dependence structure, or by relying on introducing hyperparameters that then have be set heuristically. We propose a fully automatic, non-parametric, adaptive method to estimate both mean function and autocovariance function, and further supply confidence sets for the mean function that quantify the uncertainty of the estimator. The approach is implemented in the R package *eBsc* (available from the authors), and delivers results in a quick and numerically stable way.

Our method is iterative. We start with an arbitrary guess of the correlation structure (e.g., independent) and recursively update the our data-driven choice for the smoothing

parameter and correlation structure. We can show that, under appropriate, mild assumptions, the initial choice of the smoothing parameter is consistent for an oracle, which is of the correct magnitude, but not optimal. After further iterations, the choice of the smoothing parameter is consistent for the oracle that has access to the true covariance structure of the data. Also our estimate of the covariance structure is consistent. The order of the splines to be used in the estimator is also estimated from the data.

Our numerical simulations suggest that when the covariance structure is parametric our (non-parametric) method is typically outperformed at estimating the autocovariance function by parametric methods, but still delivers comparable results for the mean. Under the same situation, compared to other non-parametric methods we seem to perform better at estimating the covariance structure. When the covariance structure is non-parametric we strongly outperform competing methods.

Appendix

A   Asymptotics

In this section we describe the asymptotic behaviour of the estimators from Section 2. The proofs for the results below are in Appendix C. Henceforth let $\mathbb{P}$, (resp. $\mathbb{E}$, $\mathbb{V}$) represent probability (resp. expectation, variance) with respect to $N(\boldsymbol{f}, \sigma^2 \boldsymbol{\mathcal{R}})$, where $f \in L_2$, $\sigma^2 > 0$, $\boldsymbol{\mathcal{R}}$ are the true values of the parameters of interest which determine the distribution of the data. The matrix $\boldsymbol{\mathcal{R}}$ denotes the true underlying correlation matrix of the data, while $\boldsymbol{R}$ denotes the "working" correlation matrix used in the smoother $\boldsymbol{S}$. We assume thought that $\boldsymbol{\mathcal{R}} \in \mathcal{M}_{n,\delta}$, so that the eigenvalues of $\boldsymbol{\mathcal{R}}$ are bounded away from zero and from infinity. The numerical procedure from Section 2.2 ensures that this is also the case for our estimates of the spectral density of the noise.

Our first result describes the uniform convergence (over λ and $\boldsymbol{R}$) of the criterium T_λ to it its expectation.

Proposition 1 (Uniform convergence of T_λ) *Assume that Λ_n (a collection such that $\lambda \in \Lambda_n$ satisfies $0 < \lambda < 1$, $\lambda \to 0$, and $n\lambda \to \infty$ as $n \to \infty$), is such that $|\Lambda_n| \exp\{-\inf_{\lambda \in \Lambda_n} \lambda^{-1/(2q)}\} = o(1)$. Then,*

$$\sup_{\boldsymbol{R} \in \mathcal{M}_{n,\delta}} \mathbb{P}\left\{ \sup_{\lambda \in \Lambda_n} \left| T_\lambda(\lambda, q, \boldsymbol{R}) - \mathbb{E}T_\lambda(\lambda, q, \boldsymbol{R}) \right| > \epsilon \right\} \to 0, \quad \epsilon > 0, \quad n \to \infty. \quad (9)$$

A consequence of this result is that, for any working correlation $\boldsymbol{R} \in \mathcal{M}_{n,\delta}$, the solution to the estimating equation $T_\lambda(\lambda, q, \boldsymbol{R}) = 0$ converges to the solution of $\mathbb{E}T_\lambda(\lambda, q, \boldsymbol{R}) = 0$. In particular we have the following proposition.

Proposition 2 (Consistency of the preliminary estimate of λ) *Let $f \in \mathcal{W}_\beta(M)$, $\beta > 1/2$, and assume that $\|f^{(\beta)}\|^2 > 0$. Assume that $\sigma^2 > 0$. Assume that the first row of*

$\mathcal{R} \in \mathcal{M}_{n,\delta}$ is absolutely summable, and denote by ϱ the underlying spectral density. Let Λ_n be as in Proposition 1. Assume $q > 1/2$ is such that $\kappa_q(\boldsymbol{I}_n)$, as defined in (24), is stricly positive. (Note that this constant depends on ϱ.) Denote by $\lambda_{q,\boldsymbol{I}}$ the solution to $\mathbb{E}T_\lambda(\lambda, q, \boldsymbol{I}_n) = 0$, $\lambda > 0$, and by $\hat{\lambda}_{q,\boldsymbol{I}}$ any solution to $T_\lambda(\lambda, q, \boldsymbol{I}_n) = o(1)$, $\lambda \in \Lambda_n$. Under these assumptions, $\hat{\lambda}_{q,\boldsymbol{I}}$ is consistent for the oracle $\lambda_{q,\boldsymbol{I}}$, in that $\hat{\lambda}_{q,\boldsymbol{I}}/\lambda_{q,\boldsymbol{I}} \to 1$, in $\mathbb{P}$-probability, as $n \to \infty$.

As for the oracle, if $q \leq \max\{\beta > 1/2 : f \in \mathcal{W}_\beta(M)\}$, then

$$\lambda_{q,\boldsymbol{I}} = \left[\frac{n\|f^{(q)}\|^2}{\sigma^2 \, \kappa_q(\boldsymbol{I}_n)} \{1 + o(1)\} \right]^{-\frac{2q}{2q+1}}, \tag{10}$$

and, if $f \in \mathcal{W}_\beta$, $\beta > 1/2$, and $q > \beta$, then

$$\lambda_{q,\boldsymbol{I}} \geq \left[\frac{n\|f^{(\beta)}\|^2}{\sigma^2 \, \kappa_q(\boldsymbol{I}_n)} \{1 + o(1)\} \right]^{-\frac{2q}{2\beta+1}}. \tag{11}$$

If the noise is white, then $\varrho = 1$ in which case the constant in the denominator becomes $\kappa_q(0, 2)$ and one just recuperates the oracles for the i.i.d. noise setting; cf. (Serra and Krivobokova, 2017).

The constant $\kappa_q(\boldsymbol{I}_n)$ will not match the constant $\kappa_q(\mathcal{R})$ from the oracle that we would get if $\mathcal{R}$ were known (this is the oracle from Theorem 1). Comparing the two constants one can characterise which types of noise processes lead to under-smoothing and which lead to over-smoothing if the correlation is ignored. Note also that the assumption that $\kappa_q(\boldsymbol{I}_n)$ has to be positive means that if the correlation structure is ignored, then the type of correlation in the noise introduces constraints on the values that q can take. To the best of our knowledge this has never been reported in the literature and justifies the need to pick q in a data driven way if the correlation structure is unknown. (The constant $\kappa_q(\mathcal{R})$ is always positive.)

Although the working correlation is not the true correlation of the data, the oracle is of the correct order and will lead to a risk for the resulting spline estimator of the correct (minimax) order (compare the oracle from Proposition 2 and the one from Theorem 1). Finite dimensional performance, though, leaves much to be desired, so the next result describes the behaviour of the estimator $\hat{\boldsymbol{R}}$ when λ is set to $\hat{\lambda}_{\boldsymbol{I}}$ and $\boldsymbol{R}$ is set to $\boldsymbol{I}_n$.

Proposition 3 (Consistency of the preliminary estimator of $\mathcal{R}$) *Let the coordinates of $\tilde{\boldsymbol{\rho}}$ solve $T_{\rho_i}(\hat{\lambda}_{q,\boldsymbol{I}}, q, 1) = 0$, and let $\bar{\boldsymbol{S}} = \boldsymbol{S}_{\xi, p, \boldsymbol{I}}$. Let $\hat{\boldsymbol{\rho}}$ be $\bar{\boldsymbol{S}} \tilde{\boldsymbol{\rho}}$ with entries first truncated to $[\delta, 1/\delta]$ and then scaled to add up to n. Further, assume that the entries of $\mathcal{R}$ satisfy $\mathcal{R}_{1,i} = O(|i|^{-2\alpha})$, $\alpha > 1/2$, so that in particular $\sum_{i=1}^{n} |\mathcal{R}_{1,i}|$ is finite. Then,*

$$\mathbb{E}\|\hat{\sigma}^2 \hat{\boldsymbol{\rho}} - \sigma^2 \boldsymbol{\varrho}\|_\infty = o(1).$$

We do now establish how subsequent iterates of the estimate of $\boldsymbol{\rho}$ behave, but replacing the constant function 1 with a consistent estimate of the spectral sensity should further improve the estimator.

This result ensures, in particular, that for any sequences $\xi, \lambda \in \Lambda_n$, the mean squared error of the estimate of the spectral density of the noise process is consistent in operator norm. The smoothing parameter ξ is picked using the estimating equation $T_\xi(\xi, p, 1)$.

The following theorem specifies the behaviour of the estimator for λ when $\boldsymbol{R}$ is set to any estimator $\hat{\boldsymbol{R}}$ of $\mathcal{R}$ that is consistent (e.g., the estimator from Proposition 3).

Theorem 1 (Consistency of estimates of λ) *Let $f \in \mathcal{W}_\beta(M)$, $\beta > 1/2$, and assume that $\|f^{(\beta)}\|^2 > 0$. Assume that $\sigma^2 > 0$. Assume that the first row of $\mathcal{R}$ is absolutely summable, and denote by ϱ the underlying spectral density. Let $\hat{\boldsymbol{R}}$ be any a.s. positive definite estimator for $\mathcal{R}$ that is consistent in operator norm.*

Consider the constant $\kappa_q(\mathcal{R})$ defined in (27). Assume also that Λ_n, as specified before is such that $|\Lambda_n| \exp\{-\inf_{\lambda \in \Lambda_n} \lambda^{-1/(2q)}\} = o(1)$. For any $q > 1/2$ denote by $\hat{\lambda}_q$ the solution

to $T_\lambda(\lambda, q, \hat{\boldsymbol{R}}) = o(1)$, $\lambda \in \Lambda_n$, and by $\lambda_{q,\boldsymbol{R}}$ the solution to $\mathbb{E}T_\lambda(\lambda, q, \boldsymbol{R}) = 0$, $\lambda > 0$.

Under these assumptions, $\hat{\lambda}_q$ is consistent for the oracle $\lambda_{q,\boldsymbol{R}}$, in that $\hat{\lambda}_q / \lambda_{q,\boldsymbol{R}} \to 1$, in $\mathbb{P}$-probability, as $n \to \infty$.

As for the value of the oracle, if $q \le \max\{\beta > 1/2 : f \in \mathcal{W}_\beta(M)\}$, then

$$\lambda_{q,\boldsymbol{R}} = \left[\frac{n\|f^{(q)}\|^2}{\sigma^2\,\kappa_q(\boldsymbol{R})}\{1 + o(1)\} \right]^{-\frac{2q}{2q+1}}, \tag{12}$$

and, if $f \in \mathcal{W}_\beta$, $\beta > 1/2$, and $q > \beta$, then

$$\lambda_{q,\boldsymbol{R}} \ge \left[\frac{n\|f^{(\beta)}\|^2}{\sigma^2\,\kappa_q(\boldsymbol{R})}\{1 + o(1)\} \right]^{-\frac{2q}{2\beta+1}}. \tag{13}$$

Note that these oracles match the oracles for λ when $\boldsymbol{R}$ is known.

The oracle provided by this result is the same as if $\boldsymbol{R}$ were known. This means that the resulting spline estimator, with data driven choice for λ, will attain the same rate as if $\boldsymbol{R}$ were known. In turn, since the noise is short range dependent, this corresponds to the same rate as one would obtain in the i.i.d. setting for estimating a β-smooth signal; cf. Section 3.3 of (Yang et al., 2001).

Our final theoretical result addresses the behaviour of the estimator for q.

Theorem 2 (Consistency of estimate of q) *Assume that the conditions of Theorem 1 are met. Assume in addition that the set Q_n from the definition of $\hat{q}$ is such that*

$$|Q_n| \exp\left\{ -\inf_{q \in Q_n} \lambda_q^{-1/(2q)} \right\} = o(1),$$

where λ_q is the oracle from the previous theorem. Let $\mathcal{M} = \mathcal{M}(L, N, \rho)$, $L > 0$, $N \in \mathbb{N}$,

s ≥ 2, corresponds the set of all square integrable sequences $b_{q,1}, b_{q,2}, \ldots$ such that

$$\frac{1}{n}\sum_{i=j}^{n} b_{q,i}^2 \leq \frac{L}{n}\sum_{i=j}^{sj} b_{q,i}^2, \quad N \leq j \leq n/s, \tag{14}$$

and define $\bar{\beta} = \max\left\{\beta > 1/2 : f \in \mathcal{W}_\beta(M) \cap \mathcal{M}\right\}$.

If for some $\beta > 1/2$, $f \in \mathcal{W}_\beta(M)$, then

$$\mathbb{P}\left\{\beta < \hat{q} \leq \log(n)\right\} \to 1, \quad n \to \infty.$$

If furthermore $\beta = \bar{\beta}$, and $f \in \mathcal{M}$, then

$$\hat{q} \xrightarrow{P} \bar{\beta}, \quad n \to \infty,$$

If for some $d \in \mathbb{N}$, $f \in \mathcal{P}_d$, the space of polynomials of degree strictly smaller than d, then

$$\mathbb{P}\left\{d \leq \hat{q} \leq \log(n)\right\} \to 1, \quad n \to \infty.$$

B Auxiliary results

In this appendix we collect a number of results that are used throughout this paper.

B.1 Demmler-Reinsch basis

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

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

(Rosales Marticorena, 2016) found explicit expressions for $\psi_{\beta,i}$ and $\nu_{\beta,i}$ as a solution to

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

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

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

$$\nu_{\beta,i} = \left\{ \pi \left(i - \frac{\beta+1}{2} \right) \right\}^{2\beta}, \quad i = \beta+1, \beta+2, \ldots$$

$$\psi_{\beta,i}(x) = \sqrt{2} \left[\cos \left\{ \pi \left(i - \frac{\beta+1}{2} \right) x + \pi \frac{\beta-1}{4} \right\} + T_i(x) \right] \tag{15}$$

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

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

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

$$\left| \frac{1}{n} \sum_{i=1}^{n} f(x_i) - \int_0^1 f(x)dx \right| \leq \frac{c}{n^2} \|f\|_{W_\beta},$$

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

B.2   Matrix identitites

We derive some matrix identities that are needed in our derivations. The smoother matrix $\boldsymbol{S}$ can also be written as

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

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 is also simple to check the scaling relation

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

The estimating equation for λ is obtained in a strightforward 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

$$\frac{\partial \boldsymbol{R}^{-1}(\boldsymbol{I}_n - \boldsymbol{S})}{\partial q} = \boldsymbol{R}^{-1}\boldsymbol{S}\frac{\partial \boldsymbol{S}^{-1}}{\partial q}\boldsymbol{S} = \frac{\lambda}{q}\boldsymbol{R}^{-1}\boldsymbol{S}\boldsymbol{R}\boldsymbol{\Phi}_q\mathrm{diag}\{n\boldsymbol{\eta}_q \circ \log(n\boldsymbol{\eta}_q)\}\boldsymbol{\Phi}_q^T\boldsymbol{S}$$

$$= \frac{\lambda}{q}\boldsymbol{S}^T\boldsymbol{\Phi}_q\mathrm{diag}\{n\boldsymbol{\eta}_q \circ \log(n\boldsymbol{\eta}_q)\}\boldsymbol{\Phi}_q^T\boldsymbol{S}$$

$$= \frac{1}{q}\boldsymbol{R}^{-1}(\boldsymbol{I} - \boldsymbol{S})\boldsymbol{\Phi}_q\mathrm{diag}\{\log(n\boldsymbol{\eta}_q)\}\boldsymbol{\Phi}_q^T\boldsymbol{S},$$

where $\circ$ denotes the Hadamard product. For the derivation of the estimating equations for the ρ_i we use the fact 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, we see that by using the chain rule

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

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

B.3 Traces and quadratic forms

We compute some traces and quadratic forms involving the smoother matrix $\boldsymbol{S}$. Let ρ be the spectral density corresponding to $\boldsymbol{R}$.

Define the matrix

$$\boldsymbol{S}^* = \boldsymbol{S}^*_{\lambda,q,\boldsymbol{R}} = \boldsymbol{\Phi}_q\left\{\boldsymbol{I}_n + \lambda\,\mathrm{diag}(n\boldsymbol{\eta}_q \circ \boldsymbol{\rho})\right\}^{-1}\boldsymbol{\Phi}_q^T, \tag{16}$$

where $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_n)$, for $\rho_i = \rho(\pi t_i)$, $i = 1, \ldots, n$, and $\circ$ represents the Hadamard

product. From the definition of $\boldsymbol{S}$ and $\boldsymbol{S}^*$,

$$\boldsymbol{S} = \left[\boldsymbol{I}_n + \lambda \boldsymbol{R}\boldsymbol{\Phi}_q \text{diag}(\boldsymbol{\rho})^{-1}\text{diag}(n\boldsymbol{\eta}_q\boldsymbol{\rho})\boldsymbol{\Phi}_q{}^T\right]^{-1} = \left[\boldsymbol{I}_n + \boldsymbol{R}\boldsymbol{\Phi}_q\text{diag}(\boldsymbol{\rho})^{-1}\boldsymbol{\Phi}_q{}^T(\boldsymbol{S}^{*-1} - \boldsymbol{I}_n)\right]^{-1}$$

$$= \boldsymbol{S}^* \left[\boldsymbol{I}_n + \boldsymbol{S}^* - \boldsymbol{I}_n + \boldsymbol{R}\boldsymbol{\Phi}_q\text{diag}(\boldsymbol{\rho})^{-1}\boldsymbol{\Phi}_q{}^T(\boldsymbol{I}_n - \boldsymbol{S}^*)\right]^{-1} = \boldsymbol{S}^*\left[\boldsymbol{I}_n + \boldsymbol{\Delta}(\boldsymbol{I}_n - \boldsymbol{S}^*)\right]^{-1},$$

where the perturbation matrix $\boldsymbol{\Delta}$ is given by

$$\boldsymbol{\Delta} = \boldsymbol{\Delta}(\boldsymbol{R}) = \boldsymbol{R}\boldsymbol{\Phi}_q\text{diag}(\boldsymbol{\rho})^{-1}\boldsymbol{\Phi}_q{}^T - \boldsymbol{I}_n. \tag{17}$$

With this notation, the relation between $\boldsymbol{S}$ and $\boldsymbol{S}^*$ is equivalent to

$$\boldsymbol{S}(\boldsymbol{I}_n + \boldsymbol{\Delta}) = (\boldsymbol{I}_n + \boldsymbol{S}\boldsymbol{\Delta})\boldsymbol{S}^*. \tag{18}$$

(Note that $\boldsymbol{S}$, $\boldsymbol{S}^*$, and $\boldsymbol{I}_n + \boldsymbol{\Delta}$ are positive definite, so that also $\boldsymbol{I}_n + \boldsymbol{S}\boldsymbol{\Delta}$ is positive definite.)

Let $\boldsymbol{M}$ be an arbitrary, positive definite $n \times n$ matrix with bounded singular values. Our goal is to compute traces and quadratic forms involving matrices of the type $\boldsymbol{M}(\boldsymbol{I}_n - \boldsymbol{S})^m\boldsymbol{S}^l$, $m \in \mathbb{N} \cup \{0\}$. We first show that if all but a finite number of singular values of $\boldsymbol{\Delta}$ are $o(1)$, then:

$$\text{tr}\left[\boldsymbol{M}\{\boldsymbol{S}(\boldsymbol{I}_n + \boldsymbol{\Delta})\}^m\right] = \text{tr}\left[\boldsymbol{S}^r\boldsymbol{M}\{\boldsymbol{S}(\boldsymbol{I}_n + \boldsymbol{\Delta})\}^{m-r}\right]\{1 + o(1)\}, \quad r = 0, 1, \ldots, m.$$

The equality in the previous display is trivial for $r = 0$. Let us then assume that equality above holds for a certain $r = p \geq 0$ and show that it also holds for $r = p + 1$, so that the

previous display follows by induction. By Von Neumann's trace inequality,

$$\left|\mathrm{tr}\left[\boldsymbol{S}^p\boldsymbol{M}\{\boldsymbol{S}(\boldsymbol{I}_n+\boldsymbol{\Delta})\}^{m-p}\right] - \mathrm{tr}\left[\boldsymbol{S}^{p+1}\boldsymbol{M}\{\boldsymbol{S}(\boldsymbol{I}_n+\boldsymbol{\Delta})\}^{m-(p+1)}\right]\right| =$$

$$= \left|\mathrm{tr}\left[\boldsymbol{S}^p\boldsymbol{M}\{\boldsymbol{S}(\boldsymbol{I}_n+\boldsymbol{\Delta})\}^{m-(p+1)}\boldsymbol{S}\boldsymbol{\Delta}\right]\right| \le \mathrm{tr}\left[\boldsymbol{S}^{p+1}\boldsymbol{M}\{\boldsymbol{S}(\boldsymbol{I}_n+\boldsymbol{\Delta})\}^{m-(p+1)}\right]o(1) + O(1).$$

The singular values of all the matrices above are bounded, so we conclude that as long as the trace above converges to infinity, then

$$\mathrm{tr}\left[\boldsymbol{S}^p\boldsymbol{M}\{\boldsymbol{S}(\boldsymbol{I}_n+\boldsymbol{\Delta})\}^{m-p}\right] = \mathrm{tr}\left[\boldsymbol{S}^{(p+1)}\boldsymbol{M}\{\boldsymbol{S}(\boldsymbol{I}_n+\boldsymbol{\Delta})\}^{m-(p+1)}\right]\{1+o(1)\},$$

so that $\mathrm{tr}\left[\boldsymbol{M}\{\boldsymbol{S}(\boldsymbol{I}_n+\boldsymbol{\Delta})\}^m\right] = \mathrm{tr}\left[\boldsymbol{S}^{p+1}\boldsymbol{M}\{\boldsymbol{S}(\boldsymbol{I}_n+\boldsymbol{\Delta})\}^{m-(p+1)}\right]\{1+o(1)\}.$

By taking $r = m$ above, and using the identity (18),

$$\mathrm{tr}\left[\{(\boldsymbol{I}_n+\boldsymbol{S}\boldsymbol{\Delta})\boldsymbol{S}^*\}^m\boldsymbol{M}\right] = \mathrm{tr}\left[\boldsymbol{M}\{\boldsymbol{S}(\boldsymbol{I}_n+\boldsymbol{\Delta})\}^m\right] = \mathrm{tr}\left[\boldsymbol{M}\boldsymbol{S}^m\right]\{1+o(1)\}.$$

Using exactly the same argument with $\boldsymbol{\Delta}$ replaced with $\boldsymbol{S}\boldsymbol{\Delta}$, and using the fact that the singular values of $\boldsymbol{S}$ belong to $(0, 1]$, we conclude that the left-hand-side of the previous display is equal to $\mathrm{tr}\{\boldsymbol{M}(\boldsymbol{S}^*)^m\}\{1+o(1)\}$, so that

$$\mathrm{tr}\left(\boldsymbol{M}\boldsymbol{S}^m\right) = \mathrm{tr}\{\boldsymbol{M}(\boldsymbol{S}^*)^m\}\{1+o(1)\}.$$

(Note that the $o(1)$ term can be taken uniform in m.)

Using the previous display and the linearity of the trace,

$$\text{tr}\{\boldsymbol{M}(\boldsymbol{I}_n - \boldsymbol{S})^m \boldsymbol{S}^l\} = \sum_{k=0}^{m}(-1)^k \binom{m}{k} \text{tr}(\boldsymbol{M}\boldsymbol{S}^{k+l}) =$$

$$= \sum_{k=0}^{m}(-1)^k \binom{m}{k} \text{tr}\{\boldsymbol{M}(\boldsymbol{S}^*)^{k+l}\}\{1 + o(1)\} = \text{tr}\{\boldsymbol{M}(\boldsymbol{I}_n - \boldsymbol{S}^*)^m (\boldsymbol{S}^*)^l\}\{1 + o(1)\}.$$

Let the superscript $^\circ$ represent the Hadamard power. Define the matrices

$$\boldsymbol{\Delta}_{t,s} = \boldsymbol{\Delta}_{t,s}(\boldsymbol{\mathcal{R}}, \boldsymbol{R}) = \boldsymbol{\mathcal{R}}^t \boldsymbol{R}^s \boldsymbol{\Phi}_q \text{diag}(\boldsymbol{\varrho}^{\circ t} \circ \boldsymbol{\rho}^{\circ s})^{-1} \boldsymbol{\Phi}_q^T - \boldsymbol{I}_n, \tag{19}$$

where $\boldsymbol{\mathcal{R}}$ is a Toeplitz correlation matrix with underlying spectral density ϱ, and $\boldsymbol{\varrho}$ is defined in analogy to $\boldsymbol{\rho}$. Note that in particular $\boldsymbol{\Delta}(\boldsymbol{R}) = \boldsymbol{\Delta}_{0,1}(\boldsymbol{I}_n, \boldsymbol{R})$. We can also write $\boldsymbol{\mathcal{R}}^t \boldsymbol{R}^s (\boldsymbol{I}_n - \boldsymbol{S})^m \boldsymbol{S}^l = (\boldsymbol{\Delta}_{t,s} + \boldsymbol{I}_n)\boldsymbol{\Phi}_q \text{diag}(\boldsymbol{\varrho}^{\circ t} \circ \boldsymbol{\rho}^{\circ s})\boldsymbol{\Phi}_q^T(\boldsymbol{I}_n - \boldsymbol{S})^m \boldsymbol{S}^l$, so that if we assume that all but a finite number of singular values of $\boldsymbol{\Delta}_{t,s}$ are $o(1)$, we can use the argument from before to get rid of the $\boldsymbol{\Delta}_{t,s}$ perturbation; setting $\boldsymbol{M} = \boldsymbol{\Phi}_q \text{diag}(\boldsymbol{\varrho}^{\circ t} \circ \boldsymbol{\rho}^{\circ s})\boldsymbol{\Phi}_q^T$ we get

$$\text{tr}\{\boldsymbol{\mathcal{R}}^t \boldsymbol{R}^s (\boldsymbol{I}_n - \boldsymbol{S})^m \boldsymbol{S}^l\} = \sum_{i=q+1}^{n} \frac{\varrho_i^t \rho_i^{m+s}(\lambda n \eta_{q,i})^m}{\left(1 + \lambda n \eta_{q,i}\rho_i\right)^{m+l}}\{1 + o(1)\}.$$

In the same way, if we now set $\boldsymbol{M} = \boldsymbol{\Phi}_q \text{diag}(\boldsymbol{\varrho}^{\circ t} \circ \boldsymbol{\rho}^{\circ s})\boldsymbol{\Phi}_q^T \boldsymbol{f} \boldsymbol{f}^T$, we get

$$\boldsymbol{f}^T \boldsymbol{\mathcal{R}}^t \boldsymbol{R}^s (\boldsymbol{I}_n - \boldsymbol{S})^m \boldsymbol{S}^l \boldsymbol{f} = \sum_{i=q+1}^{n} \frac{\varrho_i^t \rho_i^{m+s}(\lambda n \eta_{q,i})^m B_{q,i}^2}{\left(1 + \lambda n \eta_{q,i}\rho_i\right)^{m+l}}\{1 + o(1)\},$$

where $\boldsymbol{B}_q = (B_{q,1}, \ldots, B_{q,n})^T$ denotes $\boldsymbol{\Phi}_q^T \boldsymbol{f}$.

We proceed like in Lemma 1 of (Serra and Krivobokova, 2017). To deal with the trace, let

$$g_{q,n}(y, \lambda) = \rho\left(\frac{\{y/\lambda\}^{1/\{2q\}} + (q+1)\pi/2}{\pi(n-1)}\right), \quad h_{q,n}(y, \lambda) = \varrho\left(\frac{\{y/\lambda\}^{1/\{2q\}} + (q+1)\pi/2}{\pi(n-1)}\right),$$

and define for $m \in \mathbb{N} \cup \{0\}$, $l \in \mathbb{N}$, $t, s \in \mathbb{Z}$,

$$\kappa_q(m, l, t, s, \varrho, \rho) = \frac{1}{2\pi q} \lim_{n \to \infty} \int_{\lambda \pi^{2q}}^{\lambda \{\pi(n-q)\}^{2q}} \frac{y^{\frac{1}{2q}+m-1} g_{q,n}(y, \lambda)^{m+s} h_{q,n}(y, \lambda)^t}{\{1 + y\, g_{q,n}(y, \lambda)\}^{m+l}} \, dy. \qquad (20)$$

Note that for any $\lambda > 0$, within the integration range the argument of ρ and ϱ above lies in $(0, 1/2]$. Note also that if the spectral densities satisfy $\delta \leq \varrho(t), \rho(t) \leq 1/\delta$, for some $0 < \delta < 1$, $t \in [0, 1]$, then

$$\delta^{2m+l+t+s} \kappa_q(m, l) \leq \kappa_q(m, l, t, s, \varrho, \rho) \leq \delta^{-(2m+l+t+s)} \kappa_q(m, l),$$

where $\kappa_q(m, l)$ are the constants from Lemma 1 of (Serra and Krivobokova, 2017), so that indeed the limit above is well defined.

With this notation we have

$$\mathrm{tr}\{\boldsymbol{\mathcal{R}}^t \boldsymbol{R}^s (\boldsymbol{I}_n - \boldsymbol{S})^m \boldsymbol{S}^l\} = \lambda^{-1/(2q)} \, \kappa_q(m, l, t, s, \varrho, \rho)\{1 + o(1)\}. \qquad (21)$$

We deal now with the quadratic forms. From (Serra and Krivobokova, 2017) it is known that for $m \in \mathbb{N} \cup \{0\}$, $l \in \mathbb{N}$,

$$\sum_{i=q+1}^n \frac{\lambda n \eta_{q,i} B_{q,i}^2}{(1 + \lambda n \eta_{q,i})^m} = n\lambda \|f^{(q)}\|^2 \{1 + o(1)\}, \qquad f \in \mathcal{W}_q,$$

$$\sum_{i=q+1}^n \frac{\lambda n \eta_{q,i} B_{q,i}^2}{(1 + \lambda n \eta_{q,i})^l} \leq n\lambda^{\beta/q} \|f^{(\beta)}\|^2 \{1 + o(1)\}, \qquad f \in \mathcal{W}_\beta, \quad q > \beta.$$

In the case where $\delta \leq \rho_i \leq 1/\delta$, for some $0 < \delta < 1$, exactly the same argument can be

used to show that for $m \in \mathbb{N} \cup \{0\}$, $l \in \mathbb{N}$,

$$\sum_{i=q+1}^{n} \frac{\lambda n \eta_{q,i} B_{q,i}^2}{\left(1 + \lambda n \eta_{q,i} \rho_i\right)^m} = n\lambda \|f^{(q)}\|^2 \{1 + o(1)\}, \qquad f \in \mathcal{W}_q,$$

$$\sum_{i=q+1}^{n} \frac{\lambda n \eta_{q,i} B_{q,i}^2}{\left(1 + \lambda n \eta_{q,i} \rho_i\right)^l} \leq n\lambda^{\beta/q} \|f^{(\beta)}\|^2 \{1 + o(1)\}, \qquad f \in \mathcal{W}_\beta, \quad q > \beta.$$

Combining the previous display with the representation for the quadratic form from before, we conclude in particular that for $m \in \mathbb{N} \cup \{0\}$,

$$\boldsymbol{f}^T \boldsymbol{R}^{-1}(\boldsymbol{I}_n - \boldsymbol{S})\boldsymbol{S}^m \boldsymbol{f} = n\lambda \|f^{(q)}\|^2 \{1 + o(1)\}, \qquad f \in \mathcal{W}_q,$$

$$\boldsymbol{f}^T \boldsymbol{R}^{-1}(\boldsymbol{I}_n - \boldsymbol{S})\boldsymbol{S}^m \boldsymbol{f} \leq n\lambda^{\beta/q} \|f^{(\beta)}\|^2 \{1 + o(1)\}, \quad f \in \mathcal{W}_\beta, \quad q > \beta, \tag{22}$$

which covers all the quadratic forms that we need to control in this paper.

B.4 $\boldsymbol{R}$ 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 we have that $r_{i-j} = r_{j-i}$ so that we can write

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

36

In practice we approximate r_k via

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

For more details see (Gray, 1972).

The following lemmas are crucial for our approach.

Lemma 2 *Let $\mathbf{R}$ be a $n \times n$ covariance matrix of a stationary process with a Lipschitz continuous spectral density ρ. If $R_{i,j} = 0$, for $|i-j| > m$, $m \in \mathbb{N}$, then*

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

Proof

Noting that $x_i = (i-1)/(n-1)$ and denoting $x_{j,q} = \{j - (q+1)/2)\}/(n-1)$ we get for $i,j = q+1, \ldots, n$

$$\begin{aligned}
\{\mathbf{\Phi}_q^T \mathbf{R} \mathbf{\Phi}_q\}_{i,j} &= \sum_{l=1}^{n} \phi_{q,i}(x_l) \sum_{k=1}^{n} \sqrt{\frac{2}{n}} \cos\{\pi(k-1)x_{j,q} + \pi(q-1)/4\} r_{|l-k|} \\
&+ \sum_{l=1}^{n} \phi_{q,i}(x_l) \sum_{k=1}^{n} \left[\phi_{q,j}(x_k) - \sqrt{\frac{2}{n}} \cos\{\pi(k-1)x_{j,q} + \pi(q-1)/4\}\right] r_{|l-k|} \\
&= \rho(\pi x_{j,q})\delta_{i,j} - \{1 + (-1)^{|i-j|}\}\sqrt{\frac{2}{n}} \sum_{l=1}^{m} \phi_{q,i}(x_l) \sum_{k=1}^{m-l+1} \cos(\pi k x_{j,q}) r_{k+l-1} \quad (23) \\
&- \rho(\pi x_{j,q}) \sum_{l=1}^{n} \phi_{q,i}(x_l) \left[\phi_{q,j}(x_l) - \sqrt{\frac{2}{n}} \cos\{\pi(l-1)x_{j,q} + \pi(q-1)/4\}\right] \\
&+ \sum_{l=1}^{n} \phi_{q,i}(x_l) \sum_{k=1}^{n} \left[\phi_{q,j}(x_k) - \sqrt{\frac{2}{n}} \cos\{\pi(k-1)x_{j,q} + \pi(q-1)/4\}\right] r_{|l-k|}.
\end{aligned}$$

We used that $R_{i,j} = r_{|i-j|} = 0$ for $|i-j| > m$, for some fixed m independent on n and that cosine and sine are even and odd functions, respectively. Three last terms in (23) vanish for $|i-j|$ odd; for the last two terms this follows since $\phi_{q,j}(x_k)$ has the same number of

37

sign changes as $\cos\{\pi(k-1)x_{j,q} + \pi(q-1)/4\}$ (see section B.1) and $\phi_{q,i}(x_k)$ is an even function for i odd and odd function for i even. If $|i-j|$ is even, then the second term is of order $O(n^{-1})$ since $\phi_{q,i}(x) = O(n^{-1/2})$ and m is fixed, small and independent on n. To see that the last two terms are of order $O(n^{-1})$ for $|i-j|$ even, use the Cauchy-Schwarz inequality, Lemma 1 and the result from Section B.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 (15), with the tail parts $T_i(x)$ vanishing exponentially away from the boundaries. Finally, since ρ is Lipschitz continuous, it follows $\rho(\pi x_{j,q}) = \rho(\pi x_j) + O(n^{-1}) = \rho_j + O(n^{-1})$.

Lemma 3 *Under assumptions of Lemma 2, as n goes to infinity, the perturbation matrices $\boldsymbol{\Delta}$ from (17) has a bounded number of eigenvalues of order $O(1)$, and the remaining eigenvalues are $o(1)$.*

Proof

Let $\boldsymbol{\Phi}_q^T \boldsymbol{R} \boldsymbol{\Phi}_q = \boldsymbol{U}\,\mathrm{diag}(\boldsymbol{\zeta})\boldsymbol{U}^t$. On the one hand, $\zeta_i = \rho_i + O(n^{-1})$, on the other hand, from Lemma 2 we know that $\{\boldsymbol{\Phi}_q^T \boldsymbol{R} \boldsymbol{\Phi}_q\}_{i,j} = \rho_j \delta_{i,j} + O(n^{-1})$, $i = q+1, \ldots, n$, so that

$$\{\boldsymbol{\Phi}_q^T \boldsymbol{R} \boldsymbol{\Phi}_q\}_{i,i} = \{\boldsymbol{U}\,\mathrm{diag}(\boldsymbol{\zeta})\boldsymbol{U}^t\}_{i,i} = U_{i,i}^2 \zeta_i + \sum_{j \neq i} U_{i,j}^2 \zeta_j = \rho_i + O(n^{-1}) = \zeta_i + O(n^{-1}).$$

Hence,

$$(U_{i,i}^2 - 1)\zeta_i + \sum_{j \neq i} U_{i,j}^2 \zeta_j = O(n^{-1}), \ i = q+1, \ldots, n.$$

Note that both ζ_i and ρ_i are positive and $\boldsymbol{U}$ is an orthogonal matrix. Therefore, for each j, it holds that $U_{i,i}^2 = 1 + O(n^{-1})$ and there is a bounded number of $U_{i,j} = O(n^{-1/2})$, while the remaining $n\{1 + o(1)\}$ elements $U_{i,j} = O(n^{-1})$. Hence, $\boldsymbol{U}$ is asymptotically a diagonal matrix.

We need to find the bound on the eigenvalues of $\boldsymbol{\Delta}$, which are the same as the eigenvalues

of

$$\text{diag}\left(\frac{1}{\sqrt{\boldsymbol{\rho}}}\right)\boldsymbol{\Phi}_q^T\boldsymbol{R}\boldsymbol{\Phi}_q\text{diag}\left(\frac{1}{\sqrt{\boldsymbol{\rho}}}\right) - \boldsymbol{I}_n.$$

Now,

$$\boldsymbol{\Phi}_q^T\boldsymbol{R}\boldsymbol{\Phi}_q = \boldsymbol{U}\text{diag}(\boldsymbol{\rho})\boldsymbol{U}^t + \boldsymbol{U}\text{diag}(\boldsymbol{\zeta} - \boldsymbol{\rho})\boldsymbol{U}^t.$$

Since the eigenvalues of the second matrix are $O(n^{-1})$, except for the first q eigenvalues, it is enough to consider

$$\boldsymbol{V} = \text{diag}\left(\frac{1}{\sqrt{\boldsymbol{\rho}}}\right)\boldsymbol{U}\text{diag}(\boldsymbol{\rho})\boldsymbol{U}^t\text{diag}\left(\frac{1}{\sqrt{\boldsymbol{\rho}}}\right).$$

The diagonal elements $V_{i,i}$, $i = q+1, \ldots, n$ are given by

$$\sum_{j:|i-j|\leq k} U_{i,j}^2\frac{\rho_i}{\rho_j} + \sum_{j:|i-j|>k} U_{i,j}^2\frac{\rho_i}{\rho_j} = \sum_{j:|i-j|\leq k} U_{i,j}^2\{1+O(n^{-1})\}+O(n^{-1}) = \sum_{j=1}^{n} U_{i,j}^2\{1+O(n^{-1})\},$$

for some fixed bounded k. We used that $\rho_i/\rho_j = 1+O(n^{-1})$ for $|i-j| \leq k$ due to Lipschitz continuity of the spectral density ρ, while for $|i-j| > k$ the ratio ρ_i/ρ_j is bounded by the assumption on $\boldsymbol{R}$. In the same way, the off-diagonal elements $V_{i,j}$ are given by

$$\sum_{l=1}^{n} U_{i,l}U_{j,l}\frac{\rho_l}{\sqrt{\rho_i\rho_j}} = \sum_{l=1}^{n} U_{i,l}U_{j,l}\{1 + O(n^{-1})\}.$$

Hence, we can write $V_{i,j} = \{\boldsymbol{U}\text{diag}\{1+O(n^{-1})\}\boldsymbol{U}^t\}_{i,j} = \delta_{i,j} + O(n^{-1})$, $i,j = q+1,\ldots,n$ and the i-th eigenvalue of $\boldsymbol{V}$ is $1 + O(n^{-1})$, $i = q+1,\ldots,n$. This completes the proof. A similar result holds for the matrices $\boldsymbol{\Delta}_{t,s}$ from (19) but we omit the details.

B.5 Derivation of the estimating equations

Exact estimating equations given in Section 2.2 follow from the expression for the log-likelihood in (5) and the identities for the derivatives from Appendix B.2. These exact expressions for the estimating equations become difficult to evaluate in a numerically stable way if n and/or q are large. We suggest to use the approximate expressions that can be evaluated in a numerically stable way and provide excellent approximations for the exact expressions even for small sample size.

The approximate estimating equations for $T_\lambda(\lambda, q, \boldsymbol{\rho})$, $T_q(\lambda, q, \boldsymbol{\rho})$, and $T_{\rho_i}(\lambda, q, \boldsymbol{\rho})$ are obtained by using the approximation (16) of the smoother $\boldsymbol{S}$.

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

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

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

where

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

and the logarithm is taken entry-wise, and with the convention that $0 \cdot \log(0) = 0$. The theoretical performance of the solutions to these equations can be found in Appendix C.

C Proofs

Here we collect the proofs for Proposition 1, Proposition 2, Proposition 3, Theorem 1, and Theorem 2 in Sections C.1 through C.5, respectively.

C.1 Uniform convergence of T_λ to its expectation

Note that $\boldsymbol{W} = \boldsymbol{\mathcal{R}}^{-1/2}\boldsymbol{Y} \sim N(\boldsymbol{\mathcal{R}}^{-1/2}\boldsymbol{f}, \sigma^2 \boldsymbol{I}_n)$ so that if $\boldsymbol{L}$ is any symmetric, positive semi-definite matrix, then $\mathbb{E}(\boldsymbol{W}^T \boldsymbol{L} \boldsymbol{W}) = \boldsymbol{f}^T \boldsymbol{\mathcal{R}}^{-1/2} \boldsymbol{L} \boldsymbol{\mathcal{R}}^{-1/2} \boldsymbol{f} + \sigma^2 \mathrm{tr}(\boldsymbol{L})$. We write

$$T_\lambda(\lambda, q, \boldsymbol{R}) = \frac{1}{n_{\lambda,q}} \boldsymbol{Z}^T \boldsymbol{L} \boldsymbol{Z} - \frac{1}{n_{\lambda,q}} \{\boldsymbol{Z}^T \boldsymbol{L}' \boldsymbol{Z} + 1\} \frac{\mathrm{tr}(\boldsymbol{S}) - q}{n+1},$$

where $\boldsymbol{Z} = \boldsymbol{\Phi}_q^T \boldsymbol{\mathcal{R}}^{-1/2} \boldsymbol{Y} \sim N(\boldsymbol{z}, \sigma^2 \boldsymbol{I}_n)$, for $\boldsymbol{z} = \boldsymbol{\Phi}_q^{T} \boldsymbol{\mathcal{R}}^{-1/2} \boldsymbol{f}$, and the symmetric, positive definite matrices $\boldsymbol{L}$ and $\boldsymbol{L}'$ are

$$\boldsymbol{L} = \boldsymbol{\Phi}_q^T \boldsymbol{\mathcal{R}}^{1/2} \boldsymbol{R}^{-1}(\boldsymbol{I}_n - \boldsymbol{S})\boldsymbol{S}\boldsymbol{\mathcal{R}}^{1/2}\boldsymbol{\Phi}_q, \qquad \boldsymbol{L}' = \boldsymbol{\Phi}_q^T \boldsymbol{\mathcal{R}}^{1/2} \boldsymbol{R}^{-1}(\boldsymbol{I}_n - \boldsymbol{S})\boldsymbol{\mathcal{R}}^{1/2}\boldsymbol{\Phi}_q.$$

Denote by Λ_n a finite collection of values for λ such that $\lambda \to 0$, and $n\lambda \to \infty$. For any $\boldsymbol{R}$, we can upper bound $\sup_{\lambda \in \Lambda_n} |T_\lambda(\lambda, q, \boldsymbol{R}) - \mathbb{E} T_\lambda(\lambda, q, \boldsymbol{R})|$ as the sum of two suprema, one involving $\boldsymbol{L}$, and one involving $\boldsymbol{L}'$.

Consider the term involving $\boldsymbol{L}$ (the term involving $\boldsymbol{L}'$ is controlled in exactly the same way, and is in fact smaller). By Lemma 3 all but a finite number of singular values of $\boldsymbol{\Delta}_{1/2,-1}$ and of $\boldsymbol{\Delta}_{1/2,0}$ are $o(1)$, the rest of which are $O(1)$ so then it suffices to control

$$\sup_{\lambda \in \Lambda_n} \left| \frac{1}{n_{\lambda,q}} \boldsymbol{Z}^T \mathrm{diag}\left[\left\{\frac{\lambda\, n\eta_{q,i}\varrho_i}{\left(1 + \lambda\, n\eta_{q,i}\rho_i\right)^2}\right\}_i\right]\boldsymbol{Z} - \mathbb{E}\frac{1}{n_{\lambda,q}} \boldsymbol{Z}^T \mathrm{diag}\left[\left\{\frac{\lambda\, n\eta_{q,i}\varrho_i}{\left(1 + \lambda\, n\eta_{q,i}\rho_i\right)^2}\right\}_i\right]\boldsymbol{Z} \right|.$$

To control this we use the following inequalities for quadratic form in Gaussian random variables; cf. (Enikeeva et al., 2018). The random variables Z_i^2/σ^2, $i = q+1, \dots, n$, are independent, and distributed like $\mathcal{X}_1(z_i^2)$, a non-central $\mathcal{X}^2$-distribution with non-centrality parameter z_i^2, and 1 degree of freedom. If we abbreviate $M = \sum_{i=q+1}^n c_{\lambda,q,i} Z_i^2/\sigma^2$, for

$c_{\lambda,q,i} \geq 0$, and if $c_{\lambda,q} = \max_{i=q+1,\ldots,n} c_{\lambda,q,i}$, then for any $x > 0$,

$$\mathbb{P}\left\{ M - \mathbb{E}M \leq -\sqrt{2\mathbb{V}M\,x} \right\} \leq \exp(-x),$$

$$\mathbb{P}\left\{ M - \mathbb{E}M > \sqrt{2\mathbb{V}M\,x} + 2c_{\lambda,q}\,x \right\} \leq \exp(-x),$$

where $\mathbb{E}M = \sum_{i=q+1}^{n} c_{\lambda,q,i}(1 + z_i^2)$, and $\mathbb{V}M = 2\sum_{i=q+1}^{n} c_{\lambda,q,i}^2(1 + 2z_i^2)$. Note that using the bounds on the eigenvalues of autocorrelation matrices,

$$2\mathrm{tr}\left[\boldsymbol{\mathcal{R}}^2 \boldsymbol{R}^{-2}(\boldsymbol{I}_n - \boldsymbol{S}^*)^2 \{\boldsymbol{S}^*\}^2 \right] \leq \mathbb{V}M \leq 2\mathrm{tr}\left[\boldsymbol{\mathcal{R}}^2 \boldsymbol{R}^{-2}(\boldsymbol{I}_n - \boldsymbol{S}^*)^2 \{\boldsymbol{S}^*\}^2 \right] + 4\delta^{-5}\,n\lambda^{1\wedge(\beta/q)},$$

such that if we assume that $\beta > 1/2$,

$$O\{\lambda^{-1/(2q)}\} \leq \mathbb{V}M \leq O\{\lambda^{-1/(2q)} + n\lambda^{1\wedge(\beta/q)}\} \leq O\{\lambda^{-1/(2q)} + n\lambda^{1/(2q)}\}.$$

Note that this holds, in fact, for the variance of all of the quadratic forms involved in the definition of the estimating equation for λ, hence the scaling used there.

Combining the probability bounds with a union bound we get that for any $\epsilon > 0$,

$$\mathbb{P}\left(\sup_{\lambda \in \Lambda_n} \frac{|M - \mathbb{E}M|}{n_{\lambda,q}} > \epsilon \right) \leq 2|\Lambda_n| \sup_{\lambda \in \Lambda_n} \exp(-x),$$

as long as $\epsilon\, n_{\lambda,q} \geq \sqrt{2\mathbb{V}M\,x} + 2c_{\lambda,q}\,x$. The largest x such that this inequality holds is

$$\frac{\mathbb{V}M}{2c_{\lambda,q}}\left(\sqrt{\frac{4\epsilon\, n_{\lambda,q}c_{\lambda,q}}{\mathbb{V}M} + 1} - 1 \right)^2 \geq O\{\lambda^{-1/(2q)}\},$$

where the lower bound (which comes from using the bounds on the variance of M, noting that $c_{\lambda,q} = 1 + o(1)$, and by definition of $n_{\lambda,q}$) holds for any $\epsilon > 0$.

The conclusion is that as long as $|\Lambda_n| \exp\{- \inf_{\lambda \in \Lambda_n} \lambda^{-1/(2q)}\} = o(1)$, i.e., if the grid Λ_n does not have too many elements, and if the largest element in Λ_n is not too large, then the estimating equation for λ converges uniformly over Λ_n to its expectation. Note that this holds true over all $\boldsymbol{R}$ that satisfy our eigenvalue bounds, so we conclude that

$$\sup_{\boldsymbol{R} \in \mathcal{M}_{n,\delta}} \mathbb{P}\left\{ \sup_{\lambda \in \Lambda_n} \left| T_\lambda(\lambda, q, \boldsymbol{R}) - \mathbb{E}T_\lambda(\lambda, q, \boldsymbol{R}) \right| > \epsilon \right\} \to 0, \quad \epsilon > 0, \quad n \to \infty. \tag{9}$$

C.2 Consistency of the preliminary estimate of λ

Convergence of the estimator to the respective oracle follows from Proposition 1 and Theorem 5.9 of (van der Vaart, 1998). It remains to compute the oracle.

Firstly, the expression from the expectation of $T_\lambda(\lambda, q, \boldsymbol{R})$ for arbitrary $\boldsymbol{R}$ is given in Appendix C.3. The identity matrix $\boldsymbol{I}_n$ can be quite off from the true correlation matrix, so the trace that features in the expectation of T_λ is not necessarily positive. By (21), this trace is $\kappa_q(\boldsymbol{I}_n) \lambda^{-1/(2q)}\{1 + o(1)\}$, where

$$\kappa_q(\boldsymbol{I}_n) = \kappa_q(0, 2, 1, 0, \varrho, 1) + \kappa_q(0, 1, 0, 0, \varrho, 1) - \kappa_q(0, 1, 1, 0, \varrho, 1). \tag{24}$$

To ensure that a solution to the estimating equation $\mathbb{E}T_\lambda(\lambda, q, \boldsymbol{I}_n) = 0$ exists, we assume that $\kappa_q(\boldsymbol{I}_n)$ is positive. Again using (21) and (22), and solving for λ, we conclude that if $\kappa_q(\boldsymbol{I}_n) > 0$, then the solution to the estimating equation when $f \in \mathcal{W}_q$, $\boldsymbol{R} = \boldsymbol{I}_n$, is

$$\lambda_{q,\boldsymbol{I}} = \left[\frac{n \|f^{(q)}\|^2}{\sigma^2 \, \kappa_q(\boldsymbol{I}_n)} \{1 + o(1)\} \right]^{-\frac{2q}{2q+1}}, \tag{10}$$

and if $f \in \mathcal{W}_\beta$, $q > \beta$, then

$$\lambda_{q,\boldsymbol{I}} \geq \left[\frac{n \|f^{(\beta)}\|^2}{\sigma^2 \, \kappa_q(\boldsymbol{I}_n)} \{1 + o(1)\} \right]^{-\frac{2q}{2\beta+1}}. \tag{11}$$

43

C.3 Consistency of the preliminary estimator of $\mathcal{R}$

The result follows by just combining known results. The estimators for the eigenvalues $\sigma^2 \boldsymbol{\varrho}$ of the covariance matrix $\sigma^2 \mathcal{R}$ are

$$Z_i = \hat{\sigma}^2 \tilde{\rho}_i = \frac{\lambda n \eta_{q,i} B_{q,i}^2}{1 + \lambda n \eta_{q,i}}, \qquad i = q+1, \ldots, n,$$

where $\boldsymbol{B} \sim N\big(\boldsymbol{\Phi}_q \boldsymbol{f}, \sigma^2 \boldsymbol{\Phi}_q \mathcal{R} \boldsymbol{\Phi}_q^T\big)$. Based on Lemma 2, as n grows, this is distributed like $N\big(\boldsymbol{b}_q, \sigma^2 \mathrm{diag}\{\boldsymbol{\varrho}\}\big)$, where $\boldsymbol{b}_q = \mathbb{E}\boldsymbol{B}_q = \boldsymbol{\Phi}_q \boldsymbol{f}$. Note then that

$$\mathbb{E}\big(Z_i - \sigma^2 \varrho_i\big) = \frac{b_i^2 \, \lambda n \eta_{q,i}}{1 + \lambda n \eta_{q,i}} - \frac{\sigma^2 \varrho_i}{1 + \lambda n \eta_{q,i}}, \quad \text{and} \quad \mathbb{V}Z_i = \left(\frac{\lambda n \eta_{q,i}}{1 + \lambda n \eta_{q,i}}\right)^2 \big(6\sigma^2 b_i^2 \varrho_i + 3\sigma^4 \varrho_i^2\big).$$

The Z_i, as estimates of $\sigma^2 \varrho_i$ are not consistent – they are biased and their variances do not converge to zero. To obtain consistent estimates we smooth the Z_i using a smoothing spline. Let $\bar{\boldsymbol{S}} = \boldsymbol{S}_{\xi,p,\boldsymbol{I}}$ be the spline smoother, and consider the estimates $\hat{\boldsymbol{\rho}} = \bar{\boldsymbol{S}} \tilde{\boldsymbol{\rho}}$.

The heteroscedasticity of the noise terms is not an issue. As remarked by Eggermont and LaRiccia (2009, p.233), as long as the variances of the noise terms are bounded uniformly over n, the usual smoothing spline estimator (for homoskedastic noise) attains the optimal rate. The authors remark that weighted spline estimators based on estimates of the variances of the noise terms do not lead to large differences in the estimator.

Let us assume that the autocorrelations of the stationary noise process in (1) satisfy $r_i = O(|i|^{-2\alpha})$, $\alpha > 1/2$, such that in particular $\sum_{i=0}^{\infty} |r_i|$ is finite. The bias can be shown to vanish by combining this with (22) and using the consistency of the preliminary estimator for λ based on the identity as working correlation. Then the risk of the estimates of the eigenvalues of the autocovariance matrix is

$$\mathbb{E}\|\hat{\sigma}^2 \hat{\boldsymbol{\rho}} - \sigma^2 \boldsymbol{\varrho}\|_{\infty} = o(1).$$

This follows from (Eggermont and LaRiccia, 2006) where the authors determine uniform rates for spline smoothers.

C.4 Consistency of estimates of λ

By Lemma 3 all but a finite number of singular values of $\mathbf{\Delta}_{1/2,-1}(\mathcal{R}, \hat{\mathbf{R}})$, $\mathbf{\Delta}_{1/2}(\mathcal{R}, \hat{\mathbf{R}})$, and $\mathbf{\Delta}(\hat{\mathbf{R}})$ are $o_P(1)$, the rest of which are $O_P(1)$. By construction, $\hat{\mathbf{R}} \in \mathcal{M}_{n,\delta}$, with probability 1, for some $0 < \delta < 1$. The same can be assumed without loss of generality for $\mathcal{R}$. To control $T_\lambda(\lambda, q, \hat{\mathbf{R}}) - \mathbb{E}T_\lambda(\lambda, q, \mathcal{R})$ uniformly over $\lambda \in \Lambda_n$ we write this difference as

$$\left\{T_\lambda(\lambda, q, \mathbf{R}) - \mathbb{E}T_\lambda(\lambda, q, \mathbf{R})\right\}|_{\mathbf{R}=\hat{\mathbf{R}}} + \left\{\mathbb{E}T_\lambda(\lambda, q, \mathbf{R})\right\}|_{\mathbf{R}=\hat{\mathbf{R}}} - \mathbb{E}T_\lambda(\lambda, q, \mathcal{R}), \qquad (25)$$

where in the first three terms $\hat{\mathbf{R}}$ is plugged in for $\mathbf{R}$. In Appendix C.2 we show that

$$\sup_{\mathbf{R} \in \mathcal{M}_{n,\delta}} \mathbb{P}\left\{\sup_{\lambda \in \Lambda_n} \left|T_\lambda(\lambda, q, \mathbf{R}) - \mathbb{E}T_\lambda(\lambda, q, \mathbf{R})\right| > \epsilon\right\} \to 0, \quad n \to \infty. \qquad (9)$$

Using this fact we can control the first difference in (25) by noting that

$$\mathbb{P}\left[\sup_{\lambda \in \Lambda_n} \left|\left\{T_\lambda(\lambda, q, \mathbf{R}) - \mathbb{E}T_\lambda(\lambda, q, \mathbf{R})\right\}|_{\mathbf{R}=\hat{\mathbf{R}}}\right| > \epsilon\right] \le$$
$$\le \mathbb{P}\left[\sup_{\lambda \in \Lambda_n} \left|T_\lambda(\lambda, q, \hat{\mathbf{R}}) - \mathbb{E}\left\{T_\lambda(\lambda, q, \mathbf{R})\right\}|_{\mathbf{R}=\hat{\mathbf{R}}}\right| > \epsilon \mid \hat{\mathbf{R}} \in \mathcal{M}_{n,\delta}\right] + \mathbb{P}\left(\hat{\mathbf{R}} \notin \mathcal{M}_{n,\delta}\right).$$

Both terms on the right-hand-side converge to zero, as $n \to \infty$, so it remains to bound the second difference in (25).

Let us denote $\mathbf{S} = \mathbf{S}_{\lambda,q} = S_{\lambda,q,\mathcal{R}}$. These are the ideal spline smoothers that we would have used if the correlation matrix $\mathcal{R}$ were known. First we look at the expectation of the estimates for the variance that feature in the expectation of the estimating equation

for λ. Denote

$$\boldsymbol{\nabla} = \boldsymbol{\nabla}(\mathcal{R}, \boldsymbol{R}) = \mathcal{R} - \boldsymbol{R}, \tag{26}$$

and let $\mathcal{N}_n = \mathcal{N}_n(\mathcal{R})$ be the collection of all matrices $\boldsymbol{R}$ such that the largest eigenvalue of $\boldsymbol{\nabla}$ is $o(1)$. Then, for any $\boldsymbol{R} \in \mathcal{N}_n \cap \mathcal{M}_{n,\delta}$, and any $\lambda \in \Lambda_n$,

$$\mathbb{E}\big(\hat{\sigma}^2_{\boldsymbol{R}}\big) = \frac{1}{n+1}\Big[\boldsymbol{f}^T\boldsymbol{R}^{-1}(\boldsymbol{I}_n - \boldsymbol{S})\boldsymbol{f} + 1 + \sigma^2\mathrm{tr}\big\{(\boldsymbol{\nabla}\boldsymbol{R}^{-1} + \boldsymbol{I}_n)(\boldsymbol{I}_n - \boldsymbol{S})\big\}\Big]$$

$$= \sigma^2 \frac{n}{n+1}\Big[1 + O\big(\lambda^{1 \wedge (\beta/q)}\big) + O\big\{\lambda^{-1/(2q)}/n\big\} + O\{\mathrm{tr}(\boldsymbol{\nabla})/n\}\Big] = \sigma^2\{1 + o(1)\},$$

so that $\mathbb{E}\big(\hat{\sigma}^2_{\boldsymbol{R}}\big) = \sigma^2\{1 + o(1)\} = \mathbb{E}\big(\hat{\sigma}^2_{\mathcal{R}}\big)$, if we assume w.l.g., that $\mathcal{R} \in \mathcal{M}_{n,\delta}$.

The estimating equation for λ can be written in matrix notation as

$$T_\lambda(\lambda, q, \boldsymbol{R}) = \frac{1}{n_{\lambda,q}}\Big[\boldsymbol{Y}^T\boldsymbol{R}^{-1}(\boldsymbol{I}_n - \boldsymbol{S})\boldsymbol{S}\boldsymbol{Y} - \hat{\sigma}^2\big\{\mathrm{tr}(\boldsymbol{S}) - q\big\}\Big],$$

so simple computations and use of (21) and (22) show that for each λ, q, $\boldsymbol{R}$,

$$\mathbb{E}T_\lambda(\lambda, q, \boldsymbol{R}) = \frac{1}{n_{\lambda,q}}\Big[\boldsymbol{f}^T\boldsymbol{R}^{-1}(\boldsymbol{I}_n - \boldsymbol{S})\boldsymbol{S}\boldsymbol{f} + \sigma^2\mathrm{tr}\big\{\mathcal{R}\boldsymbol{R}^{-1}(\boldsymbol{I}_n - \boldsymbol{S})\boldsymbol{S}\big\} - \mathbb{E}\hat{\sigma}^2\big\{\mathrm{tr}(\boldsymbol{S}) - q\big\}\Big],$$

For any $\boldsymbol{R}$, up to lower order terms,

$$\mathbb{E}T_\lambda(\lambda, q, \boldsymbol{R}) = \frac{1}{n_{\lambda,q}}\Big[\boldsymbol{f}^T\boldsymbol{R}^{-1}(\boldsymbol{I}_n - \boldsymbol{S})\boldsymbol{S}\boldsymbol{f} + \sigma^2\mathrm{tr}\big\{(\boldsymbol{\nabla}\boldsymbol{R}^{-1} + \boldsymbol{I}_n)(\boldsymbol{I}_n - \boldsymbol{S})\boldsymbol{S}\big\} - \sigma^2\mathrm{tr}(\boldsymbol{S})\Big].$$

The difference $\mathbb{E}T_\lambda(\lambda, q, \boldsymbol{R}) - \mathbb{E}T_\lambda(\lambda, q, \mathcal{R})$ involves the difference of two quadratic forms and differences of traces, which we treat separately. From (22), for any $\boldsymbol{R} \in \mathcal{M}_{n,\delta}$, the influence of $\boldsymbol{R}$ on the particular quadratic form above lies only in the lower order terms

so we can directly conclude that

$$\boldsymbol{f}^T\boldsymbol{R}^{-1}(\boldsymbol{I}_n - \boldsymbol{S})\boldsymbol{S}\boldsymbol{f} - \boldsymbol{f}^T\boldsymbol{\mathcal{R}}^{-1}(\boldsymbol{I}_n - \boldsymbol{\mathcal{S}})\boldsymbol{\mathcal{S}}\boldsymbol{f} = o\left(n\lambda^{1\wedge(\beta/q)}\right).$$

Now we control the differences of traces. If we assume that $\boldsymbol{\Delta}$ is $o(1)$, and that $\boldsymbol{R} \in \mathcal{N}_n \cap \mathcal{M}_{n,\delta}$, then this difference is

$$\text{tr}\{\boldsymbol{\nabla}\boldsymbol{R}(\boldsymbol{I}_n - \boldsymbol{S})\boldsymbol{S}\} + \text{tr}(\boldsymbol{S}^2) - \text{tr}(\boldsymbol{\mathcal{S}}^2) =$$

$$= o\left[\text{tr}\{\boldsymbol{R}(\boldsymbol{I}_n - \boldsymbol{S})\boldsymbol{S}\}\right] + \text{tr}(\{\boldsymbol{S}^*\}^2)\{1 + o(1)\} - \text{tr}(\{\boldsymbol{\mathcal{S}}^*\}^2)\{1 + o(1)\}$$

$$= \text{tr}\{(\boldsymbol{S}^* + \boldsymbol{\mathcal{S}}^*)(\boldsymbol{S}^* - \boldsymbol{\mathcal{S}}^*)\} + o\{\lambda^{-1/(2q)}\},$$

where $\boldsymbol{\mathcal{S}}^*$ is $\boldsymbol{S}^*$ where $\boldsymbol{R}$ was replaced with $\boldsymbol{\mathcal{R}}$, and since the eigenspaces of $\boldsymbol{S}^*$ and $\boldsymbol{\mathcal{S}}^*$ are the same so that these two matrices comute. The eigenvalues of $\boldsymbol{S}^* - \boldsymbol{\mathcal{S}}^*$ satisfy

$$\left|\frac{1}{1 + \lambda n\eta_{q,i}\rho_i} - \frac{1}{1 + \lambda n\eta_{q,i}\varrho_i}\right| \leq \frac{\lambda n\eta_{q,i}|\rho_i - \varrho_i|}{(1 + \lambda n\eta_{q,i}\rho_i)(1 + \lambda n\eta_{q,i}\varrho_i)}$$

$$\leq o\left\{\frac{1}{\delta}\min\left(\frac{1}{1 + \lambda n\eta_{q,i}\rho_i}, \frac{1}{1 + \lambda n\eta_{q,i}\varrho_i}\right)\right\},$$

since the largest eigenvalue of $\boldsymbol{\nabla}$ is $o(1)$.

We conclude that

$$\text{tr}\{(\boldsymbol{S}^* + \boldsymbol{\mathcal{S}}^*)(\boldsymbol{S}^* - \boldsymbol{\mathcal{S}}^*)\} = o\left[\text{tr}(\{\boldsymbol{\mathcal{S}}^*\}^2)\right] + o\left[\text{tr}(\{\boldsymbol{S}^*\}^2)\right] = o\{\lambda^{-1/(2q)}\},$$

whence uniformly over $\lambda \in \Lambda_n$, and for any $\boldsymbol{R} \in \mathcal{A}$, which denotes the collection of matrices such that the largest singular value of $\boldsymbol{\Delta}$ is $o(1)$, and so that $\boldsymbol{R} \in \mathcal{N}_n \cap \mathcal{M}_{n,\delta}$, it

holds that

$$\mathbb{E}T_\lambda(\lambda, q, \boldsymbol{R}) - \mathbb{E}T_\lambda(\lambda, q, \boldsymbol{\mathcal{R}}) = \frac{o\left(n\lambda^{1\wedge(\beta/q)}\right) + o(\lambda^{-1/(2q)})}{n_{\lambda,q}} = \frac{o(n_{\lambda,q})}{n_{\lambda,q}} = o(1),$$

because $\beta > 1/2$. Since the probability that $\hat{\boldsymbol{R}} \in \mathcal{A}$ converges to 1, we conclude that

$$\mathbb{P}\left\{\sup_{\lambda\in\Lambda_n} \left|\mathbb{E}\{T_\lambda(\lambda, q, \boldsymbol{R})\}\big|_{\boldsymbol{R}=\hat{\boldsymbol{R}}} - \mathbb{E}T_\lambda(\lambda, q, \boldsymbol{\mathcal{R}})\right| > \epsilon\right\} \leq$$

$$\leq \mathbb{P}\left\{\sup_{\lambda\in\Lambda_n} \left|\mathbb{E}\{T_\lambda(\lambda, q, \boldsymbol{R})\}\big|_{\boldsymbol{R}=\hat{\boldsymbol{R}}} - \mathbb{E}T_\lambda(\lambda, q, \boldsymbol{\mathcal{R}})\right| > \epsilon \mid \hat{\boldsymbol{R}} \in \mathcal{A}\right\} + \mathbb{P}(\hat{\boldsymbol{R}} \notin \mathcal{A}) = o(1).$$

From this we conclude that if $|\Lambda_n| \exp\{-\inf_{\lambda\in\Lambda_n} \lambda^{-1/(2q)}\} = o(1)$, then

$$\mathbb{P}\left\{\sup_{\lambda\in\Lambda_n} \left|\{T_\lambda(\lambda, q, \hat{\boldsymbol{R}})\} - \mathbb{E}T_\lambda(\lambda, q, \boldsymbol{\mathcal{R}})\right| > \epsilon\right\} \to 0, \quad n \to \infty.$$

It remains to compute the oracle for λ. We have

$$\mathbb{E}T_\lambda(\lambda, q, \boldsymbol{\mathcal{R}}) = \frac{1}{n_{\lambda,q}}\left\{\boldsymbol{f}^T\boldsymbol{\mathcal{R}}^{-1}(\boldsymbol{I}_n - \boldsymbol{S})\boldsymbol{S}\boldsymbol{f} - \sigma^2\text{tr}(\boldsymbol{S}^2)\right\}.$$

Contrary to the case of the preliminary estimator where $\boldsymbol{R} = \boldsymbol{I}_n$, there always exists a solution to $\mathbb{E}T_\lambda(\lambda, q, \boldsymbol{\mathcal{R}}) = 0$. Using (21) and (22), we obtain the oracle $\lambda_{q,\boldsymbol{\mathcal{R}}}$ which satisfies the following: when $f \in \mathcal{W}_q$, and $q > 1/2$, then

$$\lambda_{q,\boldsymbol{\mathcal{R}}} = \left[\frac{n\|f^{(q)}\|^2}{\sigma^2\kappa_q(\boldsymbol{\mathcal{R}})}\{1 + o(1)\}\right]^{-\frac{2q}{2q+1}}, \tag{12}$$

and if $f \in \mathcal{W}_\beta$, $q > \beta$, then

$$\lambda_{q,\boldsymbol{\mathcal{R}}} \geq \left[\frac{n\|f^{(\beta)}\|^2}{\sigma^2\kappa_q(\boldsymbol{\mathcal{R}})}\{1 + o(1)\}\right]^{-\frac{2q}{2\beta+1}}, \tag{13}$$

48

where

$$\kappa_q(\boldsymbol{\mathcal{R}}) = \kappa_q(0, 2, 0, 0, \varrho, \varrho). \tag{27}$$

By Theorem 5.9 of (van der Vaart, 1998), we conclude that since $T_\lambda(\lambda, q, \hat{\boldsymbol{R}})$ converges uniformly over λ to $\mathbb{E}T_\lambda(\lambda, q, \boldsymbol{\mathcal{R}})$, then the solution to $T_\lambda(\lambda, q, \hat{\boldsymbol{R}}) = 0$ converges to the solution of $\mathbb{E}T_\lambda(\lambda, q, \boldsymbol{\mathcal{R}}) = 0$. This concludes the proof.

C.5 Consistency of the estimate of q

Let $\hat{\boldsymbol{R}}$ be any operator norm consistent estimator for $\boldsymbol{\mathcal{R}}$. To avoid excessive use of subscripts in what follows, denote by $\hat{\lambda}_q$ the solution to $T_\lambda(\lambda, q, \hat{\boldsymbol{R}}) = 0$, and by λ_q the oracle that solves $\mathbb{E}T_\lambda(\lambda, q, \boldsymbol{\mathcal{R}}) = 0$. Denote $n'_{\lambda,q} = n_{\lambda,q}(\log \lambda)^2$.

Using matrix notation and the identitites from before, we can write $T_q(\lambda, q, \boldsymbol{R})$ as

$$\frac{1}{n'_{\lambda,q}}\left[\boldsymbol{Y}^T\boldsymbol{S}^T\boldsymbol{\Phi}_q\mathrm{diag}\{\lambda n\boldsymbol{\eta}_q\log(n\boldsymbol{\eta}_q)\}\boldsymbol{\Phi}_q^T\boldsymbol{S}\boldsymbol{Y} - \hat{\sigma}^2\mathrm{tr}\left[\boldsymbol{S}\boldsymbol{\Phi}_q\mathrm{diag}\{\log(n\boldsymbol{\eta}_q)\}\boldsymbol{\Phi}_q^T\right]\right].$$

(The justification for the scaling follows below.) If we use the fact that we can write $\log(n\boldsymbol{\eta}_q) = \log(1/\lambda) + \log(\lambda n\boldsymbol{\eta}_q)$, take expectations and note that since the eigenvalues of $\boldsymbol{\mathcal{R}}$ are bounded away from zero and from infinity, and since the largest singular value of $\boldsymbol{\Delta}(\boldsymbol{R})$ is $o(1)$, then both $\mathrm{tr}\left[\boldsymbol{\mathcal{R}}\boldsymbol{S}\boldsymbol{\Phi}_q\mathrm{diag}\{\lambda n\boldsymbol{\eta}_q\log(\lambda n\boldsymbol{\eta}_q)\}\boldsymbol{\Phi}_q^T\boldsymbol{S}\right]$ and also $\mathrm{tr}\left[\boldsymbol{S}\boldsymbol{\Phi}_q\mathrm{diag}\{\lambda n\boldsymbol{\eta}_q\log(\lambda n\boldsymbol{\eta}_q)\}\boldsymbol{\Phi}_q^T\right]$ are $o\{\log(1/\lambda)\lambda^{-1/(2q)}\}$ (cf. Lemma 2 of (Serra and Krivobokova, 2017)); consequently we have

$$\mathbb{E}T_q(\lambda, q, \boldsymbol{R}) = \frac{1}{n'_{\lambda,q}}\boldsymbol{f}^T\boldsymbol{S}\boldsymbol{\Phi}_q\mathrm{diag}\{\lambda n\boldsymbol{\eta}_q\log(\lambda n\boldsymbol{\eta}_q)\}\boldsymbol{\Phi}_q^T\boldsymbol{S}\boldsymbol{f} + \frac{1}{\log(1/\lambda)}\mathbb{E}T_\lambda(\lambda, q, \boldsymbol{R}).$$

This last expression is useful because it tells us that if we plug $\boldsymbol{\mathcal{R}}$ into $\boldsymbol{R}$, and λ_q into λ, then the second term on the right-hand-side of the previous display is zero, and so, since

49

the largest singular values of $\mathbf{\Delta}(\mathcal{R})$ is $o(1)$,

$$
n'_{\lambda,q} \cdot \mathbb{E} T_q(\lambda_q, q, \mathcal{R}) = \boldsymbol{f}^T \boldsymbol{\mathcal{S}} \boldsymbol{\Phi}_q \mathrm{diag}\{\lambda_q n \boldsymbol{\eta}_q \log(\lambda_q n \boldsymbol{\eta}_q)\} \boldsymbol{\Phi}_q^T \boldsymbol{\mathcal{S}} \boldsymbol{f} =
$$

$$
= \sum_{i=q+1}^n \frac{b_{q,i}^2 \lambda_q n \eta_{q,i} \log(\lambda_q n \eta_{q,i})}{(1 + \lambda_q n \eta_{q,i} \varrho_i)^2} \{1 + o(1)\} = \sum_{i=q+1}^n \frac{b_{q,i}^2 \lambda_q n \eta_{q,i} \log(\lambda_q n \eta_{q,i})}{(1 + \lambda_q n \eta_{q,i})^2} \{1 + o(1)\},
$$

since the eigenvalues of $\mathcal{R}$ are bounded away from zero and from infinity; the last equality follows from the proof of Lemma 3 of (Serra and Krivobokova, 2017). (It is only the order of $\lambda n \eta_{q,i}$ in the denominator that is relevant to compute the sum.) This lemma specifies, in fact, the behaviour of the last sum depending on the smoothness of f, which in turn characterises the oracle for q; cf. Theorem 2 of (Serra and Krivobokova, 2017).

It remains to show consistency. This follows the same steps as the the uniform consistency proof for T_λ (with slightly different weights in the quadratic forms) so we only highlight the differences. As before, to show that $\hat{q}$ is consistent for the respective oracle we control, uniformly over $q \in Q_n$, the difference $T_q(\hat{\lambda}_q, q, \hat{\boldsymbol{R}}) - \mathbb{E} T_q(\lambda_q, q, \mathcal{R})$ which we write as

$$
\{T_q(\lambda, q, \boldsymbol{R}) - \mathbb{E} T_q(\lambda, q, \boldsymbol{R})\}|_{(\lambda, \boldsymbol{R}) = (\hat{\lambda}_q, \hat{\boldsymbol{R}})} + \{\mathbb{E} T_q(\lambda, q, \boldsymbol{R})\}|_{(\lambda, \boldsymbol{R}) = (\hat{\lambda}_q, \hat{\boldsymbol{R}})} - \mathbb{E} T_q(\lambda_q, q, \mathcal{R}).
$$

The first difference is controlled in the same way as the respective difference in T_λ. In turn, this term can be written as the difference between two quadratic forms and their respective expectations. If $\mathbf{\Delta}_{1/2,-1}$ and $\mathbf{\Delta}_{1/2,0}$ are $o(1)$, then the first of these is, up to lower order terms, bounded by

$$
\frac{1}{n'_{\lambda,q}} \left| \boldsymbol{Z}^T \mathrm{diag}\left[\left\{ \frac{\lambda n \eta_{q,i} \varrho_i \log(n \eta_{q,i})}{\left(1 + \lambda n \eta_{q,i} \rho_i\right)^2} \right\}_i \right] \boldsymbol{Z} - \mathbb{E} \boldsymbol{Z}^T \mathrm{diag}\left[\left\{ \frac{\lambda n \eta_{q,i} \varrho_i \log(n \eta_{q,i})}{\left(1 + \lambda n \eta_{q,i} \rho_i\right)^2} \right\}_i \right] \boldsymbol{Z} \right|,
$$

where $\boldsymbol{Z} = \boldsymbol{\Phi}_q^T \mathcal{R}^{-1/2} \boldsymbol{Y} \sim N(\boldsymbol{z}, \sigma^2 \boldsymbol{I}_n)$. We want to control the supremum over $q \in Q_n$ of the previous display. (In analogy to the T_λ case, the remaining difference is of the same

type and is controlled in the same way.) If we denote by M the quadratic form in the previous display, then if we assume that $\beta > 1/2$,

$$O\{\lambda^{-1/(2q)}\} \leq \frac{\mathbb{V}M}{(\log \lambda)^2} \leq O\{\lambda^{-1/(2q)} + n\lambda^{1 \wedge (\beta/q)}\} \leq O\{\lambda^{-1/(2q)} + n\lambda^{1/(2q)}\},$$

where we use Lemmas 2 and 3 of (Serra and Krivobokova, 2017). (This justifies the scaling that we use in the criterium for q.) Furthermore, uniformly over λ such that $\lambda_q/\lambda \in [1 - \zeta, 1 + \zeta] = L_\zeta$, and $\boldsymbol{R}$ such that $\boldsymbol{R} \in \mathcal{N}_n \cap \mathcal{M}_{n,\delta}$, this variance satisfies

$$O\{(\log \lambda_q)^2 \lambda_q^{-1/(2q)}\} \leq \mathbb{V}M \leq O\{(\log \lambda_q)^2 \lambda_q^{-1/(2q)} + (\log \lambda_q)^2 n\lambda_q^{1/(2q)}\}.$$

Proceeding as before we have that uniformly over λ such that $\lambda_q/\lambda \in L_\zeta$ and $\boldsymbol{R} \in \mathcal{N}_n \cap \mathcal{M}_{n,\delta}$,

$$\mathbb{P}\left(\sup_{q \in Q_n} \frac{|M - \mathbb{E}M|}{n'_{\lambda,q}} > \epsilon\right) \leq 2|Q_n| \exp\left[-\inf_{q \in Q_n} (\log \lambda_q)^2 \lambda_q^{-1/(2q)}\right].$$

Let $\mathcal{A}$ denote the event that $\lambda_q/\hat{\lambda}_q \in L_\zeta$, $0 < \zeta < 1$, and that $\hat{\boldsymbol{R}} \in \mathcal{N}_n \cap \mathcal{M}_{n,\delta}$. From the consistency of $\hat{\lambda}_q$ and $\hat{\boldsymbol{R}}$ we known that the probability of this event converges to 1, as $n \to \infty$. The remaining term that comes from the variance is actually smaller, but can be controlled in the same way. From this we conclude that

$$\mathbb{P}\left[\sup_{q \in Q_n} \left|\{T_q(\lambda, q, \boldsymbol{R}) - \mathbb{E}T_\lambda(\lambda, q, \boldsymbol{R})\}|_{(\lambda, \boldsymbol{R}) = (\hat{\lambda}_q, \hat{\boldsymbol{R}})}\right|\right] \leq$$

$$\leq \mathbb{P}\left[\sup_{q \in Q_n} \left|\{T_q(\lambda, q, \boldsymbol{R}) - \mathbb{E}T_\lambda(\lambda, q, \boldsymbol{R})\}|_{(\lambda, \boldsymbol{R}) = (\hat{\lambda}_q, \hat{\boldsymbol{R}})}\right| \mid \mathcal{A}\right] + \mathbb{P}(\mathcal{A}^c) = o(1).$$

It remains to control $\{\mathbb{E}T_q(\lambda, q, \boldsymbol{R})\}|_{(\lambda, \boldsymbol{R}) = (\hat{\lambda}_q, \hat{\boldsymbol{R}})} - \mathbb{E}T_q(\lambda_q, q, \mathcal{R})$. As we saw before, under the event that the eigenvalues of $\boldsymbol{R}$ and of $\mathcal{R}$ are bounded away from zero and infinity

(which occurs with probability going to 1),

$$\mathbb{E}T_q(\lambda, q, \boldsymbol{R}) - \mathbb{E}T_q(\lambda_q, q, \boldsymbol{\mathcal{R}}) =$$

$$= \left[\frac{1}{n'_{\lambda,q}} \sum_{i=q+1}^{n} \frac{b_{q,i}^2 \lambda n \eta_{q,i} \log(\lambda n \eta_{q,i})}{(1 + \lambda n \eta_{q,i})^2} - \frac{1}{n'_{\lambda_q,q}} \sum_{i=q+1}^{n} \frac{b_{q,i}^2 \lambda_q n \eta_{q,i} \log(\lambda_q n \eta_{q,i})}{(1 + \lambda_q n \eta_{q,i})^2} \right] \{1 + o(1)\}.$$

By the fundamental theorem of calculus, we see that

$$|\mathbb{E}T_q(\lambda, q, \boldsymbol{R}) - \mathbb{E}T_q(\lambda_q, q, \boldsymbol{\mathcal{R}})| \leq \frac{1}{n'_{\lambda,q}} \left| \sum_{i=q+1}^{n} \frac{b_{q,i}^2 \lambda_q n \eta_{q,i} \log(\lambda_q n \eta_{q,i})}{(1 + \lambda_q n \eta_{q,i})^2} \right| \cdot \left| 1 - \frac{\lambda}{\lambda_q} \right|.$$

Under the event $\mathcal{A}$ this is indeed $o(1)$. With this we conclude the proof.

References

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.

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

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

Cai, T., Zhang, C.-H., and Zhou, H. (2010). Optimal rates of convergence for covariance matrix estimation. *Ann. Statist.*, 38:2118–2144.

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

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

Durban, M. and Currie, I. (2003). A note on P-spline additive models correlated errors. *Comput. Statist.*, 18:89–121.

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

Eggermont, P. P. B. and LaRiccia, V. N. (2006). Uniform error bounds for smoothing splines. *Lecture Notes-Monograph Series*, 51:220–237.

Enikeeva, F., Munk, A., and Werner, F. (2018). Bump detection in heterogeneous gaussian regression. *Bernoulli*, 24(2):1266–1306.

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

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. *Advances in Mathematics*, 10(2):300 – 316.

Gray, R. (1972). On the asymptotic eigenvalue distribution of toeplitz matrices. *IEEE Trans. Inform. Theory*, 18(6):725–730.

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.

Hart, J. and Wehrley, T. (1986). Kernel regression estimation using repeated measurements data. *J. Am. Statist. Ass.*, 81(396):1080–1088.

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.

Kohn, R., Ansley, C., and Wong, C. (1992). Nonparametric spline regression with autoregressive moving average errors. *Biometrika*, 79:335–346.

Kotz, S. and Nadarajah, S. (2004). *Multivariate t-distributions and their applications.* Cambridge University Press.

Krivobokova, T. and Kauermann, G. (2007). A note on penalized spline smoothing with correlated errors. *J. Am. Statist. Ass.*, 102(480).

Lee, Y. K., Mammen, E., and Park, B. U. (2010). Bandwidth selection for kernel regression

with correlated errors. *Statistics*, 44(4):327–340.

Opsomer, J., Wang, Y., and Yang, Y. (2001). Nonparametric regression with correlated errors. *Statist. Sci.*, 16(134–153).

Purahmadi, M. (2011). Covariance estimation: the GLM and regularized perspectives. *Statist. Sci.*, 26:369–387.

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.

Serra, P. and Krivobokova, T. (2017). Adaptive empirical Bayesian smoothing splines. *Bayesian Anal.*, 12(1):219–238.

Speckman, P. and Sun, D. (2003). Fully Bayesian spline smoothing and intrinsic autoregressive priors. *Biometrika*, 90(2):289–302.

Szabó, B., van der Vaart, A. W., van Zanten, J., et al. (2015). Frequentist coverage of adaptive nonparametric bayesian credible sets. *Ann. Statist.*, 43(4):1391–1428.

Utreras Diaz, F. (1980). Pur ie choix du paramétre d'ajustement dans le lissage par fonctions spline. *Numer. Math.*, 34:15–28.

van der Vaart, A. (1998). *Asymptotic Statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press.

Wang, Y. (1998). Smoothing spline models with correlated random errors. *J. Am. Statist. Ass.*, 93:341–348.

Wood, S. (2017). *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC.

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.