Statistics and Probability Letters 79 (2009) 733–740

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

Bandwidth selection for functional time series prediction

Anestis Antoniadis [a], Efstathios Paparoditis [b], Theofanis Sapatinas [b,*]

[a] University Joseph Fourier, France

[b] University of Cyprus, Cyprus

ARTICLE INFO

Article history:
Received 8 May 2007
Received in revised form 2 October 2008
Accepted 24 October 2008
Available online 5 November 2008

ABSTRACT

We propose a method to select the bandwidth for functional time series prediction. The idea underlying this method is to calculate the empirical risk of prediction using past segments of the observed series and to select as value of the bandwidth for prediction the bandwidth which minimizes this risk. We prove an oracle bound for the proposed bandwidth estimator showing that it mimics, asymptotically, the value of the bandwidth which minimizes the unknown theoretical risk of prediction based on past segments. We illustrate the usefulness of the proposed estimator in finite sample situations by means of a small simulation study and compare the resulting predictions with those obtained by a leave-one-curve-out cross-validation estimator used in the literature.

1. Introduction

Let $X = (X(t);\ t \in \mathbb{R})$ be a (real-valued) continuous-time stochastic process defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$. Motivated by applications to prediction it is supposed that the time domain of X is divided into intervals of constant-width $\delta > 0$. Therefore, from X, a sequence of (function-valued) random variables $(Z_s;\ s \in \mathbb{N})$ is constructed according to the representation $Z_s(t) = X(t + (s-1)\delta)$, $s \in \mathbb{N}$, for all $t \in [0, \delta)$. Note that δ is not a parameter to be included in the modeling formulation. For some specific examples at hand, where some periodicity is obvious in the observed phenomena, the parameter δ is directly tied to the period. On the other hand, δ does not need to be a period. For more details see, e.g., Chapter 9 of Bosq (2000), Besse and Cardot (1996), Antoniadis and Sapatinas (2003), Antoniadis et al. (2006) and Chapter 12 of Ferraty and Vieu (2006).

The above approach is attractive because of its ability to aid understanding of the whole evolution of the continuous-time stochastic process X. In the recent literature, practically all investigations to date for this prediction problem are for the case where one assumes that (an appropriately centered version of) the discrete-time stochastic process $Z = (Z_s;\ s \in \mathbb{N})$ is a (zero-mean) Hilbert-valued autoregressive (of order 1) process (ARH(1)). The best prediction of Z_{N+1} given its past history, that is given $Z_N, Z_{N-1}, \ldots, Z_1$, is then obtained by $\rho(Z_N)$, where ρ is a bounded linear operator associated with the ARH(1) process (see, e.g, Chapter 3 of Bosq (2000)).

In what follows, we assume that each curve Z_s is observed at P equidistant time points t_i, $i = 1, 2, \ldots, P$. Letting $Z_s(t_i)$, $i = 1, 2, \ldots, P$, be the corresponding observations, it is supposed that Z_s satisfies the following functional time series regression model

$$Z_{s+1}(t_i) = m(Z_s)(t_i) + \epsilon_s(t_i), \quad i = 1, 2, \ldots, P, \tag{1}$$

where

$$m(z)(t_i) = \mathbb{E}(Z_{s+1}(t_i) \mid Z_s = z), \quad i = 1, 2, \ldots, P, \tag{2}$$

* Corresponding address: Department of Mathematics and Statistics, University of Cyprus, P.O. Box 20537, CY 1678 Nicosia, Cyprus.
E-mail address: t.sapatinas@ucy.ac.cy (T. Sapatinas).

0167-7152/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.spl.2008.10.028

z is a fixed element of $C([0,\delta))$ (the space of continuous functions defined on the interval $[0,\delta)$) and $\epsilon = (\epsilon_s,\ s \in \mathbb{N})$ is a $C([0,\delta))$-valued strong Gaussian white noise, i.e., a sequence of independent and identically distributed (i.i.d) $C([0,\delta))$-valued Gaussian random variables with $\mathbb{E}(\epsilon) = 0$ and $\mathbb{E}(\|\epsilon\|^2) < \infty$. (Note that, in this case, the errors $\epsilon_s(t_i), i = 1, 2, \ldots, P$, forms a sequence of i.i.d. Gaussian random variables with mean 0 and finite variance, say σ_ϵ^2.) It is also assumed that ϵ_s is independent of Z_j for all $j \le s, s \in \mathbb{N}$.

Consider now the following functional kernel estimator of the conditional expectation $m(z)(t_i)$ given in (2), based on the "sample" $Z_1, Z_2, \ldots, Z_N$ $(N \ge 2)$,

$$\widehat{m}_h(z)(t_i) = \frac{\sum_{s=1}^{N-1} \mathcal{K}_h(\mathcal{D}(Z_s, z))Z_{s+1}(t_i)}{\frac{1}{N} + \sum_{s=1}^{N-1} \mathcal{K}_h(\mathcal{D}(Z_s, z))}, \quad i = 1, 2, \ldots, P, \tag{3}$$

where the (positive) scalar $\mathcal{D}(x, y)$ denotes a distance (a metric or a semi-metric) between x and y, $\mathcal{K}_h(\cdot) = h^{-1}\mathcal{K}(\cdot/h)$, $\mathcal{K}(\cdot)$ is the kernel function, and h is the bandwidth (a positive number) associated with it. The predicted value of $Z_{N+1}(\cdot)$ is then obtained by $\widehat{Z}_{N+1}(\cdot) = \widehat{m}_h(Z_N)(\cdot)$. (The factor $1/N$ in the denominator allows expression (3) to be properly defined.) The estimator (3) can be viewed as a generalization in the functional framework of the classical Nadaraya–Watson regression estimator. It can be seen as a weighted average of the past 'blocks', placing more weight on those 'blocks' the preceding of which is similar to the present one. In other words, the weights associated with the $Z_{s+1}(t_i)$ values increase with the closeness between z and the corresponding Z_s, in the distance sense. For a similar approach see, e.g., Ferraty et al. (2002), Antoniadis et al. (2006) and Chapter 11 of Ferraty and Vieu (2006).

However, as in any nonparametric smoothing approach, the estimator (3) depends on the choice of the smoothing parameter h (the bandwidth associated with the kernel part of the method), and this choice is of great importance. Intuitively, a large value of h will lead to an estimator that incurs large bias, while a small value might reduce the bias but the variability of the predicted curve could be large since only few segments are used in the estimation. A good choice of h should balance the bias-variance trade off. In the standard nonparametric regression setting, where the data are assumed independent and identically distributed, a popular means of choosing h is by leave-one-out cross-validation obtained by minimizing a mean-squared prediction error, where each observation is predicted by the value of an appropriate nonparametric estimator constructed from all the data except the one to be predicted. In the functional nonparametric regression setting, a leave-one-curve-out cross-validation has been proposed by Rachdi and Vieu (2007) for selecting h for the prediction of a scalar response given a functional variable. Although such a proposal has not been theoretically justified in the functional time series context, it has been recently applied to continuous-time stochastic process prediction (see, e.g., Ferraty et al. (2002) and Chapter 12 of Ferraty and Vieu (2006)).

In this paper, we propose a method to select the bandwidth h for functional time series prediction. This could be seen as an answer to the open question 10 posed in Section 11.7.3 of Ferraty and Vieu (2006) regarding the development of an automatical bandwidth selection procedure for functional dependent data. The idea underlying our method is to use past segments of the series in order to calculate the empirical risk of prediction for different values of the bandwidth. In particular, consider the last v_N segments $Z_{N-v_N+1}, Z_{N-v_N+2}, \ldots, Z_N$, $1 < v_N \ll N$. For each one of these segments, say $Z_{N-v_N+j}, j = 1, 2, \ldots, v_N$, use its previous $n = N - v_N$ segments, that is $Z_{N-v_N+j-1}, Z_{N-v_N+j-2}, \ldots, Z_{N-v_N+j-n}$, to predict segment Z_{N-v_n+j} for a range of bandwidths within a given grid of values. The value of the bandwidth which minimizes the empirical risk of prediction over the v_N segments predicted is then selected as the bandwidth used for the prediction of segment Z_{N+1}. We prove an oracle bound for the proposed bandwidth estimator showing that it mimics, asymptotically, the value of the bandwidth which minimizes the unknown theoretical risk of prediction based on N segments. The suggested method is appropriately designed for functional time series prediction since, unlike leave-one-curve-out cross-validation, every predicted segment used in the empirical risk calculation is obtained using its n preceding segments. A related method has been applied to ozone forecasting by Damon and Guillas (2002) without any theoretical justification.

The paper is organized as follows. In Section 2, we describe the suggested method to select the bandwidth for functional time series prediction and state an oracle bound for the proposed bandwidth estimator. In Section 3, we provide a numerical study in order to illustrate the performance of the proposed bandwidth estimator in finite sample situations. We also compare the resulting predictions with those obtained by a leave-one-curve-out cross-validation estimator used in the literature. Auxiliary results and proofs are compiled in the Appendix.

2. Bandwidth selection

As mentioned in the introduction, the choice of the bandwidth h is important for the construction of the functional kernel estimator (3). Under some regularity and mixing conditions, in order for this estimator to achieve an almost surely uniform convergence over a suitable increasing sequences of compact sets in $\mathbb{R}^P$, the bandwidth h has to be of order $\left(\log^2(N)/N\right)^{1/(P+4)}$ (see the discussion after Assumptions A1–A3 in the Appendix). A practical way of selecting h is then to choose it within a grid H_N given by

$$H_N = \left\{ \frac{1}{L}Kc_N, \frac{2}{L}Kc_N, \ldots, \frac{L-1}{L}Kc_N, Kc_N \right\}, \quad L \in \mathbb{N},$$

where $c_N = \left(\log^2(N)/N\right)^{1/(P+4)}$. Here, K is a positive constant, large enough so that the grid H_N covers the optimal bandwidth, while the reciprocal value of L controls the relative difference between two consecutive values within the grid H_N and depends on the smallest bandwidth one wants to try and the precision one wishes in order to find the optimal bandwidth within the interval $(0, Kc_N]$. Of course, in practice, the constant K is unknown but fixing it to some large value hardly matters from a practical point of view (see Section 3).

Recall that $n = N - v_N$, assume that $v_N/N \to 0$ as $N \to \infty$, and let

$$h_{l,N} = \frac{l}{L} K c_N \in H_N, \quad l = 1, 2, \ldots, L,$$

and define the empirical risk calculated over the last v_N segments predicted based on the previous n segments, i.e.,

$$R_l = \frac{1}{P v_N} \sum_{i=1}^{P} \sum_{s=1}^{v_N} \left(Z_{n+s}(t_i) - \widetilde{m}_l(Z_{n+s-1})(t_i)\right)^2, \tag{4}$$

where, for each $s = 1, 2, \ldots, v_N$,

$$\widetilde{m}_l(Z_{n+s-1})(t_i) = \frac{\sum_{r=2}^{n} \mathcal{K}_{h_{l,N}}(\mathcal{D}(Z_{r+s-2}, Z_{n+s-1})) Z_{r+s-1}(t_i)}{\frac{1}{n} + \sum_{r=2}^{n} \mathcal{K}_{h_{l,N}}(\mathcal{D}(Z_{r+s-2}, Z_{n+s-1}))}, \quad i = 1, 2, \ldots, P. \tag{5}$$

Also let

$$\hat{l} = \underset{h_{l,N} \in H_N}{\operatorname{argmin}} \{R_l\},$$

and define $\mathcal{R}_l$ to be the theoretical counterpart of R_l based on N segments, i.e.,

$$\mathcal{R}_l = \frac{1}{P} \sum_{i=1}^{P} \mathbb{E}\left(m(Z_1)(t_i) - \widehat{m}_l(Z_1)(t_i)\right)^2$$

and $\widetilde{\mathcal{R}}_l$ to be the theoretical counterpart of R_l based on n segments, i.e.,

$$\widetilde{\mathcal{R}}_l = \frac{1}{P} \sum_{i=1}^{P} \mathbb{E}\left(m(Z_1)(t_i) - \widetilde{m}_l(Z_1)(t_i)\right)^2.$$

The following theorem provides an oracle bound for the proposed bandwidth estimator. The bound relies upon a sequence of real numbers $\{\delta_N\}_{N \in \mathbb{N}}$ such that $\delta_N \to 0$ as $N \to \infty$ which is related to the rate at which the tail probabilities of the conditional densities of Z_s given Z_{s+k} for all $k \in \mathbb{Z} \setminus \{0\}$ approach zero outside a sequence of compact sets C_N (see Assumption A3 in the Appendix).

Theorem 2.1. *Suppose that the Assumptions A1–A3, given in the Appendix, are true and that the sequence v_N satisfies $v_N \to \infty$ such that $v_N/N \to 0$ as $N \to \infty$. Then, there exist a non-negative sequence of real numbers $\{\delta_N\}_{N \in \mathbb{N}}$, $\delta_N \to 0$ as $N \to \infty$, and a constant $\alpha < 1$ such that the following oracle bound is true*

$$\mathbb{E}\left\{\widetilde{\mathcal{R}}_{\hat{l}}\right\} \leq (1 + \delta_N) \left\{ \min_{h_{l,N} \in H_N} \{\mathcal{R}_l\} + \frac{\log(v_N)}{v_N^{1-\alpha}} \log(L) \right\}, \tag{6}$$

where $\hat{l} = \operatorname{argmin}_{h_{l,N} \in H_N} \{R_l\}$.

Theorem 2.1 shows that, as long as the initial interval is chosen so that it contains the optimal bandwidth, the bandwidth obtained by minimizing the empirical risk of prediction based on $n = N - v_N$ segments within this interval mimics, asymptotically, the value of the bandwidth that minimizes the unknown theoretical risk of prediction based on N segments within the same interval.

3. Numerical results

In this section, we illustrate the performance of the proposed bandwidth estimator discussed in Section 2 by means of a small simulation study. We compare the resulting predictions when choosing the bandwidth with the proposed method based on an empirical risk criterion (RM) and those obtained by selecting the bandwidth by a leave-one-curve-out cross-validation criterion (CV). This latter bandwidth is obtained as

$$\hat{l} = \underset{h_{l,N} \in H_N}{\operatorname{argmin}} \{CV_l\},$$

where

$$CV_l = \frac{1}{(N-1)P} \sum_{i=1}^{P} \sum_{s=2}^{N} \left(Z_s(t_i) - \widehat{m}_l^{(-s)}(Z_{s-1})(t_i)\right)^2,$$

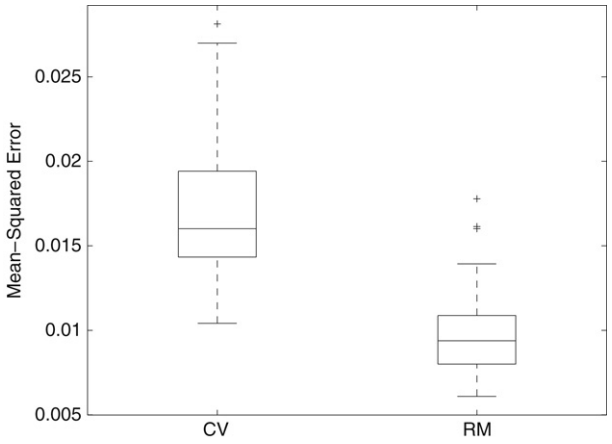

Fig. 3.1. Boxplots of the MSE for the prediction of the last segment for the CV and RM methods over 100 simulation runs.

and

$$\widehat{m}_l^{(-s)}(z)(t_i) = \frac{\sum_{j=1, j\neq s}^{N-1} \mathcal{K}_{h_{l,N}}(\mathcal{D}(Z_j, z))Z_{j+1}(t_i)}{\frac{1}{N} + \sum_{j=1, j\neq s}^{N-1} \mathcal{K}_{h_{l,N}}(\mathcal{D}(Z_j, z))}, \quad i = 1, 2, \ldots, P.$$

For both the RM and CV methods, the bandwidth h is selected within an interval $[h_{\min}, h_{\max}]$, where $h_{\min} = (K/L)c_N$ and $h_{\max} = Kc_N$ are selected according to the theoretical findings of Section 2, allowing us to control the permissible values of h. By default, the constant K is set equal to $4\hat{\sigma}$, where $\hat{\sigma}$ is the estimated standard deviation of the series, while L is set equal to 70. Finally, the kernel $\mathcal{K}(\cdot)$ is chosen to be the standard Laplace probability density function, and the distance $\mathcal{D}(x, y)$ is chosen to be the Mahalanobis distance, i.e., $\mathcal{D}(x, y) = ((x - y)'\hat{\Sigma}^{-1}(x - y))^{1/2}$, where $x, y \in \mathbb{R}^P$ and $\hat{\Sigma}$ is the sample estimator of the covariance matrix $\Sigma = \mathrm{Cov}\left((Z_s(t_1), Z_s(t_2), \ldots, Z_s(t_P))'\right)$, although other kernels $\mathcal{K}(\cdot)$ and/or distances $\mathcal{D}(x, y)$ could be used (see, e.g., Chapter 3 of Ferraty and Vieu (2006)).

The quality of prediction is measured by the mean-squared error (MSE) criterion, defined by

$$\mathrm{MSE} = \frac{1}{P}\sum_{i=1}^{P}(Z_{N_0}(t_i) - \hat{Z}_{N_0}(t_i))^2, \tag{7}$$

where Z_{N_0} is the N_0th element of the time series Z and $\hat{Z}_{N_0}$ is the corresponding prediction of Z_{N_0} given the past $Z_{N_0-1}, Z_{N_0-2}, \ldots, Z_1$ (see (3)), with the bandwidth h selected either by the RM method or the CV method. (The overall numerical study was carried out in the `Matlab 7.0.4` programming environment.)

3.1. A simulated example

We carried out a small simulation study to compare the RM and CV bandwidth selection criteria in terms of forecasting. We generated a series of observations as the superposition of two deterministic signals with different periods and a first order moving-average noise. More specifically, as in Antoniadis et al. (2006), we considered the following structure for X,

$$X(t) = \beta_1 m_1(t) + \beta_2 m_2(t) + \epsilon(t),$$

where $m_1(t) = \cos(2\pi t/64) + \sin(2\pi t/64)$, $m_2(t) = \cos(2\pi t/6) + \sin(2\pi t/6)$ and $\epsilon(t) = u(t) + \theta u(t-1)$, $u(t) \overset{i.i.d}{\sim} N(0, \sigma^2)$. The parameters were chosen as follows: $\beta_1 = 0.8$, $\beta_2 = 0.18$, $\theta = 0.8$ and $\sigma^2 = 0.005$. The motivation beyond this choice is to generate realizations containing a dominant component with a period of 64 observations, a less pronounced and more irregular component with a period of 6 observations, contaminated with an additive and correlated random component. The time period that we have analyzed runs 30 segments (i.e., $N = 30$), each one containing 64 observations (i.e., $P = 64$).

We simulated 100 realizations of the above stochastic process. For each realization, we have predicted its last segment given its past, using bandwidths h selected according to both the CV method and the RM method, the latter one based on various values of v_N, and the corresponding prediction errors have been evaluated. Our limited simulation study suggests $v_N = [\log(N)] + 1$, and we have taken this choice as a rule of thumb in our analysis. Hence, in this example, $v_N = 4$.

The boxplots of the MSE of each prediction (i.e., we have taken $N_0 = 30$ in (7)) made with the corresponding selected bandwidth over the 100 simulations are displayed in Fig. 3.1. As observed in the figure, the predictions by the RM method are better than those obtained by the CV method, in terms of MSE.

Acknowledgements

Anestis Antoniadis was supported by the 'Cyprus-France CY-FR/0907/02 Zenon Program'. Efstathios Paparoditis and Theofanis Sapatinas were supported by the 'Cyprus-France CY-FR/0907/02 Zenon Program' and the Electricity Authority of Cyprus. The authors would like to thank an anonymous referee for useful suggestions on improvements to this paper.

Appendix. Proofs

The results obtained below are based on the following set of assumptions, which we detail below before proceeding to the proofs. In what follows, Z_r refers to the P-dimensional vector $(Z_r(t_1), Z_r(t_2), \ldots, Z_r(t_P))$ for a generic index $r \in \mathbb{N}$.

Assumption A1. The conditional expectation m is bounded, i.e., there exists a positive constant A such that $\sup_{z\in\mathbb{R}^P} |m(z)| \leq A$.

Assumption A2. There exists a sequence of compact sets $\{C_N;\ N \in \mathbb{N}\}$, $C_N \subset \mathbb{R}^P$, such that, for all $n \to \infty$ satisfying $n/N \to 1$ as $N \to \infty$,

$$N^{\vartheta} \sup_{z\in C_N} |\widetilde{m}_l(z) - m(z)| \to 0 \quad \text{almost surely,}$$

for some $0 < \vartheta < 1/2$.

Assumption A3. There exists a sequence of compact sets $\{C_N;\ N \in \mathbb{N}\}$, $C_N \subset \mathbb{R}^P$, such that $\rho_N = O\left(N^{-\nu}\right)$ for some finite constant $\nu > 0$, where

$$\rho_N = \sup_{k\in\mathbb{Z}\setminus\{0\}} \sup_{z\in\mathbb{R}^P} \int_{\mathbb{R}^P\setminus C_N} g_{Z_s|Z_{s+k}}(x \mid z)\mathrm{d}x,$$

with $g_{Z_s|Z_{s+k}}$ denoting the conditional densities of Z_s given Z_{s+k} for all $k \in \mathbb{Z} \setminus \{0\}$.

Let us now comment on the nature of the above assumptions. Assumption A1 is rather restrictive, especially for the case of stochastic processes in discrete-time (e.g., classical discrete-time stationary autoregressive process); it has been used in order to simplify the technical arguments. We believe that the assumption on the boundness of $m(z)$, $z \in \mathbb{R}^P$, can be relaxed by appropriately controlling the rate at which $|m(z)|$ increases outside the compact sets $C_N \subset \mathbb{R}^P$, at the cost of more involved technical arguments. Assumption A2 is satisfied under some general conditions on the underlying stochastic process $Z = (Z_s,\ s \in \mathbb{N})$, the kernel function $\mathcal{K}(\cdot)$ and the bandwidth h. Theorem 3.3 of Bosq (1998) provides regularity and mixing conditions under which a kernel regression estimator on $\mathbb{R}^P$ achieves the above almost surely uniform convergence over a suitable increasing sequences of compact sets, requiring also that the bandwidth h is chosen to be of order $\left(\log^2(N)/N\right)^{1/(P+4)}$. On the other hand, Assumption A3 controls the behavior, outside the compact sets $C_N \subset \mathbb{R}^P$, of the tail probabilities of the conditional densities $g_{Z_s|Z_{s+k}}$ for all $k \in \mathbb{Z}\setminus\{0\}$. Note that, under the assumptions of the stochastic process $\epsilon = (\epsilon_s,\ s \in \mathbb{N})$, model (1) implies that the underlying stochastic process $Z = (Z_s,\ s \in \mathbb{N})$ is Gaussian, meaning that all conditional densities $g_{Z_s|Z_{s+k}}$, $k \in \mathbb{Z} \setminus \{0\}$, exist and are Gaussian. It is well-known that the tail probabilities of Gaussian densities approach zero power exponentially fast. Assumption A3 requires a weaker condition to hold, uniformly over $k \in \mathbb{Z} \setminus \{0\}$.

The proof of Theorem 2.1 is based on the following two lemmas. For $l \in \mathcal{L} = \{1, 2, \ldots, L\}$ and $i \in \mathcal{I} = \{1, 2, \ldots, P\}$, consider the following quantities

$$\widetilde{Q}_l = \frac{1}{Pv_N} \sum_{i=1}^{P} \sum_{s=1}^{v_N} \left[(Z_{n+s}(t_i) - \widetilde{m}_l(Z_{n+s-1})(t_i))^2 - \epsilon_{n+s}^2(t_i)\right]$$

and

$$\xi_{s,l}(t_i) = m(Z_s)(t_i) - \widetilde{m}_l(Z_s)(t_i).$$

Lemma A.1. *Suppose that the Assumptions A1–A3 are true. Let $c > 0$ and assume that $v_N > [(2c\sigma_\epsilon^2)/P] + 1$. Then, for each $l \in \mathcal{L}$, the following bound is true*

$$\mathbb{E}\left\{\mathrm{e}^{-c\,\widetilde{Q}_l}\right\} \leq \mathrm{e}^{-\tau v_N\{1-2\tau A^2(1+Mv_N^{\alpha})\}\widetilde{\mathcal{R}}_l}, \tag{8}$$

where $\alpha < 1$, A and M are some positive constants, independent of N and l, and

$$\tau = \frac{c}{v_N} - \frac{2c^2\sigma_\epsilon^2}{Pv_N^2} > 0. \tag{9}$$

Proof. After some simple algebra, we get

$$\mathbb{E}\left\{e^{-c\,\widetilde{Q}_l}\right\} = \mathbb{E}\left\{e^{-\frac{c}{Pv_N}\sum_{i=1}^{P}\sum_{s=1}^{v_N}\left(\xi^2_{n+s-1,l}(t_i)+2\xi_{n+s-1,l}(t_i)\epsilon_{n+s}(t_i)\right)}\right\}.$$

Using conditional expectation arguments, and the expression

$$\mathbb{E}\left(\exp\left\{-\sum_{i=1}^{m}a_iY_i\right\}\right) = \exp\left\{\frac{1}{2}\sum_{i=1}^{m}a_i\sigma_\epsilon^2\right\}$$

if $Y_1, Y_2, \ldots, Y_m$ are i.i.d. $N(0, \sigma_\epsilon^2)$ random variables, we get

$$\mathbb{E}\left\{e^{-c\,\widetilde{Q}_l}\right\} = \mathbb{E}\left\{e^{-\frac{\tau}{P}\sum_{i=1}^{P}\sum_{s=1}^{v_N}\xi^2_{n+s-1,l}(t_i)}\right\},$$

where τ is given by (9). Noting that $\epsilon_{n+s}(t_i)$, $i = 1, 2, \ldots, P$, is a sequence of i.i.d. Gaussian random variables with mean 0 and finite variance σ_ϵ^2, and using the inequalities $e^{-x} \le 1 - x + \frac{1}{2}x^2$ and $e^x \ge 1 + x$ for all $x \in \mathbb{R}$, we see that

$$\begin{aligned}\mathbb{E}\left\{e^{-\frac{\tau}{P}\sum_{i=1}^{P}\sum_{s=1}^{v_N}\xi^2_{n+s-1,l}(t_i)}\right\} \le \exp\Bigg\{&-\frac{\tau}{P}\sum_{i=1}^{P}\sum_{s=1}^{v_N}\mathbb{E}\left(\xi^2_{n+s-1,l}(t_i)\right)\\ &+\frac{1}{2}\frac{\tau^2}{P^2}\sum_{i_1=1}^{P}\sum_{i_2=1}^{P}\sum_{s_1=1}^{v_N}\sum_{s_2=1}^{v_N}\mathbb{E}\left(\xi^2_{n+s_1-1,l}(t_{i_1})\xi^2_{n+s_2-1,l}(t_{i_2})\right)\Bigg\}. \qquad (10)\end{aligned}$$

Consider first the second term in the exponent on the right-hand side of (10). We have

$$\begin{aligned}&\sum_{s_1=1}^{v_N}\sum_{s_2=1}^{v_N}\mathbb{E}\left(\xi^2_{n+s_1-1,l}(t_{i_1})\xi^2_{n+s_2-1,l}(t_{i_2})\right)\\ &\quad=\sum_{s=1}^{v_N}\mathbb{E}\left(\xi^2_{n+s-1,l}(t_{i_1})\xi^2_{n+s-1,l}(t_{i_2})\right)+\sum_{\substack{1\le s_1,s_2\le v_N\\ s_1\ne s_2}}\sum\mathbb{E}\left(\xi^2_{n+s_1-1,l}(t_{i_1})\xi^2_{n+s_2-1,l}(t_{i_2})\right)\\ &\quad:= T_{1,N}+T_{2,N}.\end{aligned}$$

We now study each of the terms $T_{1,N}$ and $T_{2,N}$ separately. For the first term, we have

$$\begin{aligned}T_{1,N} &= \sum_{s=1}^{v_N}\mathbb{E}\left[\left(\widetilde{m}_l(Z_{n+s-1})(t_{i_1})-m(Z_{n+s-1})(t_{i_1})\right)^2\left(\widetilde{m}_l(Z_{n+s-1})(t_{i_2})-m(Z_{n+s-1})(t_{i_2})\right)^2\right]\\ &\le \sum_{s=1}^{v_N}\left(\sup_{z\in\mathbb{R}^P}|\widetilde{m}_l(z)(t_{i_1})-m(z)(t_{i_1})|\right)^2\mathbb{E}\left(\widetilde{m}_l(Z_{n+s-1})(t_{i_2})-m(Z_{n+s-1})(t_{i_2})\right)^2\\ &\le 4v_NA^2\,\mathbb{E}\left(\widetilde{m}_l(Z_1)(t_{i_2})-m(Z_1)(t_{i_2})\right)^2,\end{aligned}$$

since, under Assumption A1, $\sup_{z\in\mathbb{R}^P}|\widetilde{m}_l(z)(t_i)-m(z)(t_i)| \le 2A$ for all t_i, $i \in \mathcal{I}$. For the second term, we have

$$\begin{aligned}T_{2,N} &= \sum_{\substack{1\le s_1,s_2\le v_N\\ s_1\ne s_2}}\sum\int_{\mathbb{R}^P}\int_{\mathbb{R}^P}\left(\widetilde{m}_l(z_1)(t_{i_1})-m(z_1)(t_{i_1})\right)^2\\ &\quad\times\left(\widetilde{m}_l(z_2)(t_{i_2})-m(z_2)(t_{i_2})\right)^2 g_{Z_{n+s_1-1},Z_{n+s_2-1}}(z_1,z_2)\,dz_1z_2\\ &\le \sum_{\substack{1\le s_1,s_2\le v_N\\ s_1\ne s_2}}\sum\int_{\mathbb{R}^P}\left(\widetilde{m}_l(z_1)(t_{i_1})-m(z_1)(t_{i_1})\right)^2\Bigg[\int_{C_n}\left(\widetilde{m}_l(z_2)(t_{i_2})-m(z_2)(t_{i_2})\right)^2\\ &\quad+\int_{\mathbb{R}^P\setminus C_n}\left(\widetilde{m}_l(z_2)(t_{i_2})-m(z_2)(t_{i_2})\right)^2\Bigg]g_{Z_{n+s_1-1},Z_{n+s_2-1}}(z_1,z_2)\,dz_1z_2\\ &:= T_{2,N}^{(1)}+T_{2,N}^{(2)}.\end{aligned}$$

We again study each of the terms above separately. For the first term, we have

$$\begin{aligned}
T_{2,N}^{(1)} &\le \sum_{\substack{1\le s_1,s_2\le v_N\\ s_1\neq s_2}}\sum \left(\sup_{z\in C_n}|\widetilde{m}_l(z)(t_{i_2})-m(z)(t_{i_2})|\right)^2\\
&\quad\times\int_{\mathbb{R}^P}\int_{C_n}\left(\widetilde{m}_l(z_1)(t_{i_1})-m(z_1)(t_{i_1})\right)^2 g_{Z_{n+s_1-1},Z_{n+s_2-1}}(z_1,z_2)\,\mathrm{d}z_1z_2\\
&= \sum_{\substack{1\le s_1,s_2\le v_N\\ s_1\neq s_2}}\sum \left(\sup_{z\in C_n}|\widetilde{m}_l(z)(t_{i_2})-m(z)(t_{i_2})|\right)^2\\
&\quad\times\int_{\mathbb{R}^P}\int_{C_n} g_{Z_{n+s_2-1}|Z_{n+s_1-1}}(z_2\mid z_1)\left(\widetilde{m}_l(z_1)(t_{i_1})-m(z_1)(t_{i_1})\right)^2 g_{Z_{n+s_1-1}}(z_1)\,\mathrm{d}z_1\mathrm{d}z_2\\
&\le v_N^2\left(\sup_{z\in C_n}|\widetilde{m}_l(z)(t_{i_2})-m(z)(t_{i_2})|\right)^2\mathbb{E}\left(\widetilde{m}_l(Z_1)(t_{i_1})-m(Z_1)(t_{i_1})\right)^2.
\end{aligned}$$

For the second term, we have

$$\begin{aligned}
T_{2,N}^{(2)} &\le \sum_{\substack{1\le s_1,s_2\le v_N\\ s_1\neq s_2}}\sum\int_{\mathbb{R}^P}\left(\widetilde{m}_l(z_1)(t_{i_1})-m(z_1)(t_{i_1})\right)^2 g_{Z_{n+s_2-1}}(z_2)\\
&\quad\times\int_{\mathbb{R}^P\setminus C_n}\left(\widetilde{m}_l(z_2)(t_{i_2})-m(z_2)(t_{i_2})\right)^2 g_{Z_{n+s_1-1}|Z_{n+s_2-1}}(z_1\mid z_2)\mathrm{d}z_1\mathrm{d}z_2\\
&\le \left(\sup_{z\in\mathbb{R}^P}|\widetilde{m}_l(z)(t_{i_1})-m(z)(t_{i_1})|\right)^2\sum_{\substack{1\le s_1,s_2\le v_N\\ s_1\neq s_2}}\sum\int_{\mathbb{R}^P}\left(\widetilde{m}_l(z_2)(t_{i_2})-m(z_2)(t_{i_2})\right)^2 g_{Z_{n+s_2}-1}(z_2)\\
&\quad\times\int_{\mathbb{R}^P\setminus C_n} g_{Z_{n+s_1-1}|Z_{n+s_2-1}}(z_1\mid z_2)\mathrm{d}z_1\mathrm{d}z_2\\
&\le 4A^2\,\mathbb{E}\left(\widetilde{m}_l(Z_1)(t_{i_2})-m(Z_1)(t_{i_2})\right)^2\sum_{\substack{1\le s_1,s_2\le v_N\\ s_1\neq s_2}}\sum\left(\sup_{z\in\mathbb{R}^P}\mathbb{P}\left(Z_{n+s_1-1}\in\mathbb{R}^P\setminus C_n\mid Z_{n+s_2-1}=z\right)\right)\\
&\le 4A^2v_N^2\rho_n\,\mathbb{E}\left(\widetilde{m}_l(Z_1)(t_{i_2})-m(Z_1)(t_{i_2})\right)^2,
\end{aligned}$$

using Assumption A3.

Combining the above bounds for $T_{2,N}^{(1)}$ and $T_{2,N}^{(2)}$, we arrive at

$$T_{2,N}\le v_N^2\left[\left(\sup_{z\in C_n}|\widetilde{m}_l(z)(t_{i_1})-m(z)(t_{i_1})|\right)^2+4A^2\rho_n\right]\mathbb{E}\left(\widetilde{m}_l(Z_1)(t_{i_2})-m(Z_1)(t_{i_2})\right)^2.$$

Using Assumptions A1–A2, it is easily seen that

$$T_{2,N}\le 4A^2Mv_N^{1+\alpha}\,\mathbb{E}\left(\widetilde{m}_l(Z_1)(t_{i_2})-m(Z_1)(t_{i_2})\right)^2,$$

where $\alpha=1-\min\{2\vartheta,\nu\}<1$ for some positive constant M.

Consider now the first term in the exponent on the right-hand side of (10). It is immediate that

$$T_{0,N}:=\sum_{s=1}^{v_N}\mathbb{E}\left\{\xi_{n+s-1,l}^2(t_i)\right\}=v_N\mathbb{E}\left(\widetilde{m}_l(Z_1)(t_i)-m(Z_1)(t_i)\right)^2.$$

By combining the above bounds $T_{0,N}$, $T_{1,N}$ and $T_{2,N}$, summing over $i\in\mathcal{I}$ in $T_{0,N}$ and over $i_1,i_2\in\mathcal{I}$ in $T_{1,N}$ and $T_{2,N}$, and using (10), we arrive at (8), thus completing the proof. □

Lemma A.2. *Assume that constants $b,c>0$ and $0<\lambda<1$ exist such that, for each $l\in\mathcal{L}$, the following bound is true*

$$\lambda\,\mathbb{E}\left\{\mathrm{e}^{-c\,\widetilde{Q}_l}\right\}\le b\,\mathrm{e}^{-c\lambda\widetilde{\mathcal{R}}_l}. \tag{11}$$

Then, the following oracle bound is true

$$\lambda\,\mathbb{E}\left\{\widetilde{\mathcal{R}}_{\hat{l}}\right\}\le\min_l\{\widetilde{\mathcal{R}}_l\}+\frac{1}{c}\log\left(\frac{bL}{\lambda}\right), \tag{12}$$

where $\hat{l}=\mathrm{argmin}_{l\in\mathcal{L}}\{\widetilde{Q}_l\}$.

Proof. Let $m = \operatorname{argmin}_{l\in\mathcal{L}}\{\widetilde{\mathcal{R}}_l\}$. Since $\widetilde{Q}_m - \widetilde{Q}_{\hat{l}} \geq 0$, and using the fact that $\mathbb{E}\left\{\widetilde{Q}_m\right\} = \widetilde{\mathcal{R}}_m$, it is easily seen that

$$\lambda\,\mathbb{E}\left\{\widetilde{\mathcal{R}}_{\hat{l}}\right\} \leq \widetilde{\mathcal{R}}_m + \mathbb{E}\left\{\max_{l\in\mathcal{L}}\{\lambda\widetilde{\mathcal{R}}_l - \widetilde{Q}_l\}\right\}.$$

Using the inequalities $c(x-a) \leq \mathrm{e}^{c(x-a)} - 1$, for each $c > 0$ and for all $a, x \in \mathbb{R}$, and $\mathbb{E}(\max_{l\in\mathcal{L}}\{X_l\}) \leq L\,\max_{l\in\mathcal{L}}\mathbb{E}\{X_l\}$, with $\{X_l\}_{l\in\mathcal{L}}$ a sequence of non-negative random variables with $|\mathcal{L}| = L$, we get

$$\lambda\,\mathbb{E}\left\{\widetilde{\mathcal{R}}_{\hat{l}}\right\} \leq \widetilde{\mathcal{R}}_m + a + \frac{1}{c}\left[L\,\max_{l\in\mathcal{L}}\mathbb{E}\left\{\mathrm{e}^{c(\lambda\,\widetilde{\mathcal{R}}_l - \widetilde{Q}_l - a)}\right\} - 1\right],$$

and by (11), we get

$$\lambda\,\mathbb{E}\left\{\widetilde{\mathcal{R}}_{\hat{l}}\right\} \leq \widetilde{\mathcal{R}}_m + a + \frac{1}{c}\left[L\,\frac{b}{\lambda}\mathrm{e}^{-c\,a} - 1\right].$$

Taking $a = (1/c)\log(bL/\lambda)$, we arrive at (12), thus completing the proof. □

We are now in a position to prove the main result.

Proof of Theorem 2.1. Standard calculations, the Cauchy–Schwarz inequality and the consistency property of both $\widehat{m}_l(\cdot)$ and $\widetilde{m}_l(\cdot)$ estimators, yield that

$$|\mathcal{R}_l - \widetilde{\mathcal{R}}_l| = O\left(\frac{v_N}{N}\right),\ \text{uniformly in } h_{l,N} \in H_N. \tag{13}$$

To bound $\mathbb{E}\left\{\widetilde{\mathcal{R}}_{\hat{l}}\right\}$, let $c = v_N^{1-\alpha}q/\log(v_N)$ for some positive constant q and some constant $\alpha < 1$. Now applying Lemmas A.1 and A.2 with this c, $b = \lambda$ and $\lambda = \tau v_N\{1 - 2\tau A^2(1 + v_N^{\alpha}M)\}/c$, and substituting the values of c and τ, we get $\lambda := \lambda_N = 1 - 2Mq/\log(v_N) + o(1/\log(v_N))$. For large enough N, as long as $1 \leq q < v_N^{\alpha}\log(v_N)P/(4\sigma_\epsilon^2)$, it follows by simple algebra that $0 < \lambda < 1$ and, noting that $\operatorname{argmin}_{l\in\mathcal{L}}\{\widetilde{Q}_l\} = \operatorname{argmin}_{h_{l,N}\in H_N}\{\widetilde{R}_l\}$, we arrive at

$$(1 - a_N)\mathbb{E}\left\{\widetilde{\mathcal{R}}_{\hat{l}}\right\} \leq \left\{\min_{h_{l,N}\in H_N}\{\widetilde{\mathcal{R}}_l\} + \frac{\log(v_N)}{v_N^{1-\alpha}}\log(L)\right\}, \tag{14}$$

where $a_N = 2Mq/\log(v_N) + o(1/\log(v_N))$. Hence, (6) follows by combining (13) and (14), on noting that $v_N \to \infty$ such that $v_N/N \to 0$ as $N \to \infty$, thus completing the proof. □

References

Antoniadis, A., Sapatinas, T., 2003. Wavelet methods for continuous-time prediction using Hilbert-valued autoregressive processes. Journal of Multivariate Analysis 87, 133–158.

Antoniadis, A., Paparoditis, E., Sapatinas, T., 2006. A functional wavelet-kernel approach for time series prediction. Journal of the Royal Statistical Society, Series B 68, 837–857.

Besse, P.C., Cardot, H., 1996. Approximation spline de la prévision d'un processus fonctionnel autorégressif d'ordre 1. Canadian Journal of Statistics 24, 467–487.

Bosq, D., 1998. Nonparametric Statistics for Stochastic Processes. In: Lecture Notes in Statistics, vol. 110. Springer-Verlag, New York.

Bosq, D., 2000. Linear Processes in Function Spaces. In: Lecture Notes in Statistics, vol. 149. Springer-Verlag, New York.

Damon, J., Guillas, S., 2002. The inclusion of exogenous variables in functional autoregressive ozone forecasting. Environmetrics 13, 759–774.

Ferraty, F., Vieu, P., 2006. Nonparametric Functional Data Analysis. Springer-Verlag, New York.

Ferraty, F., Goia, A., Vieu, P., 2002. Functional nonparametric model for time series: A fractal approach for dimension reduction. Test 11, 317–344.

Rachdi, M., Vieu, P., 2007. Nonparametric regression for functional data: Automatic smoothing parameter selection. Journal of Statistical Planning and Inference 137, 2784–2801.