Forecasting Methods for Functional Time Series[*]

Nazarii Salish [**]
BGSE and University of Cologne

Alexander Gleim
Statkraft AS

November 16, 2015

Abstract

This paper develops statistical tools for forecasting functional times series which for example can be used to analyze big data sets. To tackle the issue of time dependence we introduce the notion of functional dependence through scores of the spectral representation. We investigate the impact of time dependence thus quantified on the estimation of functional principal components. The rate of mean squared convergence of the estimator of the covariance operator is derived under long range dependence of the functional time series. After that, we suggest two forecasting techniques for functional time series satisfying our measure of time dependence and derive the asymptotic properties of their predictors. The first is the functional autoregressive model which is commonly used to describe linear processes. As our notion of functional dependence covers a broader class of processes we also study the functional additive autoregressive model and construct its forecasts by using the k-nearest neighbors approach. The accuracy of the proposed tools is verified through Monte Carlo simulations. Empirical relevance of the theory is illustrated through an application to electricity consumption in the Nordic countries.

JEL classification: C51(C55); C53; C14
AMS 2000 classification: Primary 62M10, 62G20; secondary 62P05

Keywords: Functional time series; Forecasting; Additive autoregressive process; K-nearest neighbors; Principal components

[*] We are thankful for the extensive advise and support of Alois Kneip as well as to Joerg Breitung, Matei Demetrescu and Dominik Liebl for helpful comments and suggestions. Further, we would like to thank the participants of the German Statistical Week 2014 and the Nordic Econometric Meeting 2015. The first author gratefully acknowledges the financial support from the German Research Foundation (Deutsche Forschungsgemeinschaft) through grant DE 1617/2-1.

[**] Corresponding author: University of Bonn, BGSE, Kaiserstrasse 1, 53113 Bonn, Germany. Email: salish@uni-bonn.de.

1 Introduction

In recent years advances in data collection and storage led to the possibility of recording many real life processes at increasingly high accuracy. Examples include high frequency data such as financial transactions, environmental data such as ozone or insolation maps and economic data such as income distributions or yield curves. The availability of large amounts of data offers manifold opportunities for researchers to obtain a better understanding of the underlying processes. However, to make use of this growing information and efficiently handle big data sets, suitable statistical tools are required to describe, model and forecast the relevant characteristics of this data. Functional data analysis (FDA) has emerged as a response to this request and has consequently been growing into an important field of statistical research.

In FDA, where large data sets are utilized in the form of functional observations (or curves), the focus has been mostly on independent and identically distributed observations. In many empirical applications data is collected sequentially over time. Consequently, we expect that the functional observations in a given time period are affected by past observations. Therefore, additional tools are required to analyze data that is given in the form of a functional time series (FTS). This paper studies the problem of describing and forecasting FTS and consists of two main parts. In a first step we provide a simple yet broad framework to quantify time dependencies in FTS. Second, we develop forecasting techniques for FTS under the given definition of time dependency.

Stochastic processes with time dependencies have been considered in the statistical literature. In the context of classical (i.e., finite dimensional) time series analysis, ergodicity and various mixing conditions are well established and frequently used (see, e.g. Hamilton (1994) and Davidson (1994) for a review). In the functional context, however, only few concepts are available when dealing with time-dependent observations. A key reference is Hörmann and Kokoszka (2010) who introduce a moment based notion of weak dependence using m-dependence. In this paper we complement the approach of Hörmann and Kokoszka (2010) by suggesting an alternative concept of time dependencies for FTS. Using the spectral Karhunen-Loeve representation functional observations can be represented by their functional principal component (FPC) scores. Therefore, the dependence between functional observations can be quantified through their respective FPC scores. This approach allows us to adapt various concepts of dependence available in the time series literature to the functional context. In particular, we consider dependence based on the autocovariances and cumulants of FPC scores. Further, since FPCs play a major role in explaining time dependencies it is necessary to establish the consistency of their estimates. We derive the convergence rates for the estimators of the FPCs under quite general serial dependence that allows for the long range dependence of the FPC scores.

This in turn extends the result in Hörmann and Kokoszka (2010).

In the second part of the paper we discuss forecasting methods for FTS. Most work dedicated to the prediction of (FTS) has focused on the functional autoregressive model of order one (FAR(1)) suggested in Bosq (2000). In particular, Bosq (2000) derives the estimator and the predictor for the FAR model using the Yule Walker equation and shows their consistency. Besse et al. (2000) propose a local adaptation of the FAR(1) model by introducing a nonparametric weighted kernel estimator. The issue of weak convergence for estimates of the FAR(1) model is addressed in Mas (2007). Kargin and Onatski (2008) develop a predictive factor technique for the estimation of the autoregressive operator. Park and Qian (2012) apply the FAR(1) framework to model FTS of distributions. Didericksen et al. (2012) provide a small sample simulation study of the performance of the FAR(1) model and several competing prediction techniques. More recently, Kokoszka and Reimherr (2013) suggest a testing procedure to determine the lag order for more general FAR(p) processes. Aue et al. (2015) suggest a simple alternative procedure to transform the FAR model into a vector autoregressive model of functional principal scores, where standard multivariate techniques can be used to model and predict FTS.

In order to forecast FTS that follow our concept of time dependence we discuss two forecasting techniques. First, FTS processes that have a linear response to the past functional observations can be forecasted by the FAR model. We show that the autocovariance estimator given in Bosq (2000) is consistent under our notion of time dependence and derive its convergence rate. However, the concept of time dependence we introduce covers a broader class of processes than described by FAR. More precisely, the behavior of the autocovariances of the FPC scores is less restrictive (in particular we can allow for long range dependence) and non-linear responses are possible. For this reason we generalize the FAR model to the functional additive autoregressive model (FAAR). The idea of functional additive models was introduced by Müller and Yao (2008) in the context of functional linear regressions. This approach gives rise to a more flexible and essentially nonparametric model and allows us to consider the problem of prediction as a problem of nonlinear response of the FPC scores. To estimate the nonlinear responses we propose a k-nearest neighbors classification approach that is simple to implement and in the finite-dimensional setting well understood. As this approach has been successfully applied to classical time series analysis (see, e.g., Cover and Hart (1967), Stone (1977), Stute (1984) and Yakowitz (1987)), we can use the available theoretical results to derive the convergence rate of our predictor in the FAAR model.

To assess the performance of the proposed forecasting methods in small samples we provide a Monte Carlo simulation study. In particular, we compare the accuracy of the prediction of the FAAR model to the FAR model, the multivariate score model suggested by Aue et al. (2015) and benchmark models such as mean predictor, naive predictor and prediction of VAR for discrete observations. Further, we compare the performance of the above mentioned FTS models in forecasting electricity consumption in Denmark, Finland, Norway and Sweden. Our results show that FAAR models and multivariate score models provide the most accurate forecasts.

The remainder of this paper is organized as follows. Section 2 introduces the notion of dependence for functional time series. Section 3 discusses the impact of time dependence on the estimators of the functional principal components. In Section 4 we address the FAR model, while a generalization of the FAR model to FAAR, its estimation and asymptotic properties are presented in Section 5. A supporting small sample study is presented in Section 6. An empirical application to electricity consumption is described in Section 7 and concluding remarks are given in Section 8. All proofs, figures and tables are collected in the Appendix.

2 Methodology and Assumptions

We shall assume that we observe a series of functional observations $\{X_i(t)\}$ for $t \in [a, b]$ and $i = 1, ..., N$, where the interval $[a, b]$ is normalized to $[0, 1]$. For each i the observation X_i belongs to the Hilbert space $H = L^2([0, 1], \|\cdot\|)$ of square integrable functions which is equipped with a norm $\|\cdot\|$ induced by the inner product $\langle x, y\rangle \equiv \int_0^1 x(t)y(t)\mathrm{d}t$. The object $\{X_i(t)\}_{i=1}^N$ is referred to as functional time series (see e.g., Horváth and Kokoszka, 2012, Chapter 13-16 and Bosq, 2000 for a survey on FTS analysis) and we refer to i as the time index. In what follows the data $\{X_i\}$ are assumed to be given in a functional form since the problem of data representation in functional form has been extensively studied in the literature (see, e.g., Ramsay and Silverman, 2005 for a review of the available techniques and general description of FDA).

Our attention is restricted to weakly stationary processes allowing for the standard time series representation

$$X_i = G(\varepsilon_i, \varepsilon_{i-1}, \ldots), \tag{1}$$

where $\{\varepsilon_i\}$ denotes the series of errors or innovations which are i.i.d elements from Hilbert space H, and G is a measurable function $G : H^\infty \to H$. In this paper two cases of representation (1) are considered. The first is the functional autoregressive (FAR) model that models linear responses of a FTS to its lags (see Section 4). Second, To account for possible nonlinear responses we extend the FAR framework to more general settings using the functional additive approach suggested in Müller and Yao (2008) for functional regressions (see Section 5). Representation (1) can also be extended to non-stationary sequences $\{X_i\}$. We do not pursue this topic in our paper and refer the interested reader to Horváth et al. (2014) for additional insights. For future reference, $\mathcal{S}$ denotes the space of Hilbert-Schmidt operators from H to H and is equipped with the operator norm

$\|\cdot\|_{\mathcal{S}}$ (i.e., for some $\Psi \in \mathcal{S}$, $\|\Psi\|_{\mathcal{S}} = (\sum_{h=1}^{\infty} \|\Psi(e_h)\|^2)^{1/2}$ for any orthonormal basis $\{e_h\}_{h\geq 1}$) and the space of bounded linear operators on H is denoted by $\mathcal{L}$ with the norm $\|\Psi\|_{\mathcal{L}} = \sup_{\|x\|\leq 1} \{\|\Psi(x)\|,\ x \in H\}$.

We begin by describing the concept of time dependency in functional time series. It is founded on the spectral decomposition of random functions as follows. All random functions are defined on a common probability space $(\Omega, \mathcal{A}, P)$. Let $L_H^p(\Omega, \mathcal{A}, P)$ denote the space of H valued random variables X such that for $p \geq 1$, $\mathbb{E}\|X\|^p < \infty$. Every function $X \in L_H^2$ possesses a mean function $\mu := \mathbb{E}(X)$ and a covariance operator $C(x) := \mathbb{E}[\langle X - \mu, x\rangle X - \mu]$, where $x \in L^2$ and C admits the spectral decomposition. That is,

$$C(x) = \sum_{\ell=1}^{\infty} \lambda_\ell \langle \psi_\ell, x\rangle \psi_\ell, \tag{2}$$

where $\{\lambda_\ell\}_{\ell\geq 1}$ is the strictly positive decreasing sequence of eigenvalues and $\{\psi_\ell\}_{\ell\geq 1}$ denotes the corresponding sequence of eigenfunctions (i.e., $C(\psi_\ell) = \lambda_\ell \psi_\ell$) which forms an orthonormal basis system of H. It follows that X admits the Karhunen-Loève representation

$$X(t) = \mu(t) + \sum_{\ell=1}^{\infty} \theta_\ell \psi_\ell(t), \tag{3}$$

where $\theta_\ell = \langle X, \psi_\ell\rangle$ denotes the ℓ-th functional principal component *score* of X. By construction, the sequence of functional principal component scores $\{\theta_\ell\}_{\ell\geq 1}$ is such that the elements θ_ℓ are uncorrelated across the spectral dimension ℓ, have mean zero and variance λ_ℓ. Then for a given weakly stationary FTS $\{X_i\}$ (such that for each $i = 1, ..., N$, $X_i \in L_H^2$) X_i admits a Karhunen-Loève decomposition which in turn yields a sequence of scores $\{\theta_{i,\ell}\}$, and the corresponding sequences of eigenvalues $\{\lambda_\ell\}$ and eigenfunctions $\{\psi_\ell\}_{\ell\geq 1}$.

The following assumption formalizes how time dependencies between functional observations $\{X_i\}$ are translated into their score series. Let $\kappa_{\ell_1,\ldots,\ell_q}(0,\tau_1,\ldots,\tau_{q-1})$ denote the q-th order cumulant of $(\theta_{i,\ell_1}, \theta_{i+\tau_1,\ell_2}, \ldots, \theta_{i+\tau_{q-1},\ell_q})$, where $\tau_1, \ldots, \tau_{q-1} \in \mathbb{N}$ are integers (see, e.g., Brillinger, 2001, p.19 for a more detailed description of cumulants). Then we shall assume:

Assumption 1

(i) *For some $\alpha > 2$ and all $\ell \geq 1$,*

$$\lambda_\ell - \lambda_{\ell+1} \sim \ell^{-\alpha-1}.$$

(ii) *Define $B_{\ell,s}^{(h)} := \sup_i |\mathbb{E}[\theta_{i,\ell}\theta_{i-h,s}]|$. Then there exists a constant $B > 0$ and some*

$\beta > 0$ *such that*

$$B_{\ell,s}^{(h)} \leq B\, h^{-\beta} \sqrt{\lambda_\ell \lambda_s}.$$

(iii) *For fixed* $q \geq 3$ *and some constant* $B > 0$*, the joint* q*-th order cumulants are absolutely summable*

$$\sum_{\tau_1,\ldots,\tau_{q-1}=-\infty}^{\infty} \cdots \sum \left|\kappa_{\ell_1,\ldots,\ell_q}(0,\tau_1,\ldots,\tau_{q-1})\right| \leq B \prod_{j=1}^{q} \lambda_{\ell_j}^{1/2}.$$

Part (**i**) of Assumption 1 is the standard assumption that prevents the spacing between adjacent eigenvalues λ_ℓ from being too small. It also implies that $\lambda_\ell \sim \ell^{-\alpha}$. The importance of spacing property (**i**) will become particularly apparent from the results of Corollary 2, where the asymptotic properties of eigenfunction estimators are studied.

Part (**ii**) and (**iii**) of Assumption 1 describe the form of time dependencies that we allow for the scores $\{\theta_{i,\ell}\}_{i,\ell\geq 1}$. The assumed behavior of $B_{\ell,s}^{(h)}$, which represents a measure of absolute covariances between score series $\{\theta_{i,\ell}\}$ and lagged series $\{\theta_{i-h,s}\}$, is only a mild restriction. In particular, part (**ii**) implies an intuitive restriction on the absolute summability of the h-th autocovariances of the score series $\{\theta_{i,\ell}\}_i$ across the spectral dimension ℓ, since $\sum_{\ell\geq 1} |\mathbb{E}[\theta_{i,\ell}\theta_{i-h,\ell}]| \leq \sum_{\ell \geq 1} B_{\ell,\ell}^{(h)} \leq Ch^{-\beta}$. However, absolute summability of the autocovariances of the score series is not required across the time dimension i and fixed spectral dimention ℓ. More precisely, for $0 < \beta < 1$ one can conclude that $\sum_{h=1}^{N} \mathbb{E}[\theta_{i,\ell}\theta_{i-h,\ell}] \leq \sum_{h=1}^{N} B_{\ell,\ell}^{(h)}$ is of order $N^{1-\beta}\lambda_\ell$ which diverges for fixed ℓ and large N. In what follows we refer to this as a long range dependence property. A similar restriction holds for the covariances of the score series across time dimension with fixed the spectral dimensions $\ell \neq s$, i.e.,

$$\sum_{h=1}^{N} |\mathbb{E}[\theta_{i,\ell}\theta_{i-h,s}]| = O\left(N^{1-\beta}\sqrt{\lambda_\ell \lambda_s}\right)$$

Finally, Assumption 1 (**iii**) requires absolute summability of the joint cumulants of $\{\theta_{i,\ell}\}$ up to q-th order. This allows us to control the temporal dependencies in the q-th moments of the score series across spectral and time dimension. In particular, condition (**iii**) for one fixed spectral direction ℓ, $\sum_{\tau_1,\ldots,\tau_{q-1}=-\infty}^{\infty} |\kappa_{\ell,\ldots,\ell}(0,\tau_1,\ldots,\tau_{q-1})| \leq C\lambda_\ell^{q/2}$, implies the finiteness of the q-th moment, i.e., $\mathbb{E}\|X_i\|^q < \Delta < \infty$ for all i. For more details on how moments are related to cumulants see Appendix A equation (A.1). In general this cumulant condition is standard for the time series literature (see, e.g. Andrews, 1991, Brillinger, 2001, and Demetrescu et al., 2008) and provides us with a useful measure of the joint statistical dependence of higher order moments and a convenient tool for deriving the rates of convergence. It should be noted that the value of q is method-specific and as we shall see in the sequel relaxing linear structure of the model may require strengthening the restrictions on the moments.

Furthermore, note that the concept of α-mixing is closely related to the form of time dependencies assumed in **(ii)**-**(iii)**. In fact, α-mixing together with finite sixth moments implies absolute summability of the joint cumulants up to sixth order (see, e.g. Andrews, 1991 or Gonçalves and Kilian, 2007). Hence, the main difference between the two approaches lies in the way autocovariances are handled. In general we find that conditions **(ii)** and **(iii)** have several advantages in a functional setting. First, they allow for a broader scope of time dependencies (in that absolutely summable autocovariances are not necessary which can be controlled through parameter β). Second, incorporating decay across the spectral dimension ℓ is straighforward, which is crucial for the analysis. Third, the stated conditions have an intuitive interpretation of the time dependence concept for functional data when compared to various mixing properties. Moreover, using standard time series techniques it can be easily verified in practice if there is time dependence between the scores of the FTS.

3 Properties of Functional Principal Components

The fundamental ingredients for describing time dependence in functional data are principal component scores. However, in practice scores and other FPC (C and its eigenvalues and eigenfunctions) are not known and must be estimated. Therefore, before developing forecasting methods that rely on Assumption 1, it is crucial to verify the convergence of the estimated FPC to their population counterparts. Consistency results for the FPC are available for independent observations (see, e.g., Dauxois et al., 1982) and for L^4-m-dependent functional data (see e.g., Hörmann and Kokoszka, 2010). In this section we show that consistency of the corresponding estimators extends to our time dependency settings.

We start with the preliminaries. Suppose we observe $X_1, ..., X_N$. The standard estimators for the mean function, μ, and the covariance operator, $C(x)$, are given by the following sample averages

$$\hat{\mu}(t) = \frac{1}{N}\sum_{i=1}^{N} X_i(t), \tag{4}$$

$$\widehat{C}_N(x) = \frac{1}{N}\sum_{i=1}^{N} \langle X_i - \hat{\mu}, x\rangle (X_i(t) - \hat{\mu}(t)), \quad x \in L^2. \tag{5}$$

Further, we denote the estimators of eigenvalues and eigenfunctions as $\{\hat{\lambda}_\ell\}_{\ell=1}^{L}$ and $\{\widehat{\psi}_\ell\}_{\ell=1}^{L}$, respectively. Using $\widehat{C}_N(t)$, they are computed from the eigenequation

$$\widehat{C}_N(\widehat{\psi}_\ell) = \widehat{\lambda}_\ell \widehat{\psi}_\ell.$$

Typically estimates of eigenelements ($\hat{\lambda}_\ell$ and $\widehat{\psi}_\ell$) can be obtained for an arbitrary fixed level L such that $L < N$. The asymptotic results in Section 4 and 5 provide a discussion of this issue, where L is set to be a function of N, such that $L \to \infty$ as $N \to \infty$. Ramsay and Silverman (2005, Section 6.4) discuss practical/computational methods for solving eigenequations.

Remark 1 *In what follows we shall assume without loss of generality that X_i have means equal to zero for all $i = 1, ..., N$. For any practical application the methodology introduced in this paper remains unchanged if data are centered prior to the forecasting exercise. For the completeness of the discussion we state the following result for the estimator of μ. For the weakly stationary FTS $\{X_i\}_{i=1}^N$ that fulfills Assumption 1* (**i**)-(**ii**) *we have*

$$\mathbb{E} \left\| \hat{\mu}_N - \mu \right\|^2 = O\left(\max\left\{N^{-\beta}, N^{-1}\right\}\right).$$

The following result establishes the consistency of estimator (5).

Theorem 1 *If a weakly stationary FTS $\{X_i\}_{i=1}^N$ fulfills Assumption 1 with joint cumulants up to order* 4 *then*

$$\mathbb{E} \left\| \widehat{C}_N - C \right\|_{\mathcal{S}}^2 = O\left(N^{-2\beta^*}\right),$$

where $\beta^* := \min\{\beta, 1/2\}$.

Theorem 1 implies that the fastest convergence speed that can be achieved for the empirical estimator of the covariance operator is N^{-1} when $\beta \geq 1/2$. This extends previously obtained results in Bosq (2000) and Hörmann and Kokoszka (2010) showing that the fastest convergence can also be achieved for processes that potentially posses long range dependencies. In other words, the absolute summability of the autocovariances of the functional principal component score series $\{\theta_{i,\ell}\}_{i\geq 1}$ across the time dimension i, is not necessary to get rate N^{-1}. If one is only interested in establishing the consistency of the covariance operator estimator, part (**ii**) of Assumption 1 can be relaxed to $B_{\ell,s}^{(h)} \leq Bb_h\sqrt{\lambda_\ell, \lambda_s}$ with $\sum_{h=1}^{\infty} h^{-1} b_h < \infty$. This condition allows for a slow decay of the time dependencies represented by component b_h that can even be of logarithmic order $b_h = O\left(\ln(h)^{-1-\beta}\right)$ for $\beta > 0$ (see, e.g., Davidson, 1994, Theorem 2.31).

The autocovariance operator defined as

$$\Gamma_h = \mathbb{E}\left[\langle X_i, x\rangle X_{i-h}\right], \tag{6}$$

for $i = 1, ..., N$ and some h, can estimated similarly by the sample analogue

$$\widehat{\Gamma}_{h,N} = \frac{1}{N-1} \sum_{i=1}^{N-1} \langle X_i, x \rangle \left(X_i \left(t \right) \right). \tag{7}$$

Furthermore, the following holds for any autocovariance operator of order h.

Corollary 1 *If a weakly stationary FTS $\{X_i\}_{i=1}^N$ fulfills Assumption 1 with joint cumulants up to order 4 then*

$$\mathbb{E} \left\| \widehat{\Gamma}_{h,N} - \Gamma_h \right\|_{\mathcal{S}}^2 = O\left(N^{-2\beta^*} \right).$$

Our next result gives explicit bounds for the mean squared error of the eigenelement estimators.

Corollary 2 *If a weakly stationary FTS $\{X_i\}_{i=1}^N$ fulfills Assumption 1 with joint cumulants up to order 4 then*

$$\textbf{(i)} \qquad \mathbb{E}\left(\sup_{\ell \geq 1} \left| \hat{\lambda}_\ell - \lambda_\ell \right|^2 \right) = O\left(N^{-2\beta^*} \right),$$

$$\textbf{(ii)} \qquad \mathbb{E}\left(\sup_{1 \leq \ell \geq L} \left\| a_\ell \widehat{\psi}_\ell - \psi_\ell \right\|^2 \right) = O\left(\delta_\ell^2 N^{-2\beta^*} \right),$$

where $a_\ell := sign(\langle \widehat{\psi}_\ell, \psi_\ell \rangle)$, $\delta_\ell := \max_{1 \leq k \leq \ell} (\lambda_k - \lambda_{k+1})^{-1}$.

The results in Corollary 2 indicate that, as ℓ increases, it becomes more difficult to estimate the eigenfunctions ψ_ℓ associated with λ_ℓ since the expected L^2 error is proportional to δ_ℓ^2. As a consequence, the spacing between adjacent eigenvalues $\{\lambda_\ell\}_{\ell \geq 1}$ cannot decrease too fast. In particular, by Assumption 1(**i**) $\mathbb{E}\left(\sup_{1 \leq \ell \geq L} \left\| a_\ell \widehat{\psi}_\ell - \psi_\ell \right\|^2 \right) = O\left(L^{2(1+\alpha)} N^{-2\beta^*} \right)$. Therefore, restriction $L = o\left(N^{\beta^*/(1+\alpha)} \right)$ has to hold for estimators $\{\widehat{\psi}_\ell\}_{\ell=1}^L$ to be consistent. Further, the estimator $\widehat{\psi}_l$ of ψ_l is only identified up to a change in sign. As is standard in the literature, we shall tacitly assume that the sign of $\widehat{\psi}_l$ is chosen such that $\int \widehat{\psi}_l \psi_l \geq 0$.

Note, recently Hörmann and Kidziński (2015) proofed that for the consistency of FPCs estimators the spacing property given in Assumption 1(**i**) can be relaxed to more general settings. However, our subsequent analysis of the forecasting techniques in Sections 4 and 5 requires explicit rates of convergence for the estimators $\widehat{\lambda}_\ell$ and $\widehat{\psi}_\ell$ and consequently the spacing property.

4 Forecasting Linear FTS

In this section we discuss estimation and forecasting techniques for FAR models. As pointed out in the introduction the FAR(1) model is the model most commonly used in the FTS analysis and it is natural to use it as the main linear FTS benchmark model. The theory of FAR(1) processes in Hilbert and Banach spaces is studied in Bosq (2000) to which we refer the reader for a general overview. In this section we study the estimator suggested in Bosq (2000) and derive its convergence rate under the time dependency assumption stated in Section 2. For simplicity of exposition we consider the FAR model of order one.[1] The model takes the form

$$X_i = \rho\left(X_{i-1}\right) + \varepsilon_i, \tag{8}$$

where ε_i is a strong white noise in L^2_H, i.e., ε_i is a zero mean iid sequence in L^2_H with the covariance operator $C_\varepsilon(x) := \mathbb{E}\left[\langle \varepsilon_i, x\rangle \varepsilon_i\right]$ being a positive definite Hilbert-Schmidt operator. The autoregressive operator ρ is a assumed to be Hilbert-Schmidt operator satisfying

$$\|\rho^k\|_{\mathcal{L}} < 1 \text{ for some } k \geq 1. \tag{9}$$

This condition assures strict stationarity for process X_i (see, e.g., Bosq, 2000, Theorem 3.1). In other words, if (9) holds then function $G(\cdot)$ in FTS representation (1) takes an additive linear form

$$X_i = \sum_{h=1}^{\infty} \rho^h(\varepsilon_{i-h}).$$

To formulate the estimator of $\rho(\cdot)$ and derive its convergence rate we first address the well known issue often referred to as an ill-posed inverse problem. Recall that $C(x) = \mathbb{E}\left[\langle X_i, x\rangle X_i\right]$ and $\Gamma_h(x) = \mathbb{E}\left[\langle X_i, x\rangle X_{i-h}\right]$, and both operators allow for spectral representations

$$C(x) = \sum_{\ell=1}^{\infty} \lambda_\ell \langle \psi_\ell, x\rangle \psi_\ell, \tag{10}$$

$$\Gamma_h(x) = \sum_{\ell=1}^{\infty}\sum_{s=1}^{\infty} \mathbb{E}\left[\theta_{i,\ell}, \theta_{i-h,s}\right] \langle \psi_\ell, x\rangle \psi_s. \tag{11}$$

It follows from (8) that operator equation $\Gamma_1 = \rho C$ holds and formally gives the solution $\rho = \Gamma_1 C^{-1}$. However, the operator C does not have a bounded inverse on the entire space H. It follows from (10) that $C^{-1} = \sum_{\ell=1}^{\infty} \lambda_\ell^{-1}\langle \psi_\ell, x\rangle \psi_\ell$, where $\lambda_\ell^{-1} \to \infty$ as $\ell \to \infty$ and the domain of C^{-1} is restricted to $\mathcal{D}\left(C^{-1}\right) = \{y \in H \,|\sum_{\ell=1}^{\infty}\langle y, \psi_\ell\rangle^2/\lambda^2 < \infty\}$. The standard method in the literature to circumvent this problem is to use only the first L functional components. That is, for $\lambda_1 > \lambda_2 > \ldots > 0$ we define H_L, a subspace of H spanned by

[1] See, e.g., Bosq (2000, Section 5) and Horváth and Kokoszka (2012, Chapter 15.1) for the review on how to estimate higher order FAR models

the L-eigenvectors $\psi_1, ..., \psi_L$ associated with $\lambda_1 > \ldots > \lambda_L$, and consider

$$C_L^{-1} = \sum_{\ell=1}^{L} \lambda_\ell^{-1} \langle \psi_\ell, x \rangle \psi_\ell, \tag{12}$$

where C_L^{-1} is the inverse of C on H_L and L is the function of N such that $L \to \infty$ as $N \to \infty$. Then the estimator of ρ is based on (7), the sample analog of (12) and can be formulated as

$$\widehat{\rho}_N (x) = \frac{1}{N-1} \sum_{i=1}^{N} \sum_{\ell,s=1}^{L} \widehat{\lambda}_\ell^{-1} \langle \widehat{\psi}_\ell, x \rangle \widehat{\theta}_{i,\ell} \widehat{\theta}_{i+1,s} \widehat{\psi}_s. \tag{13}$$

Remark 2 *Note that the FAR process (8)-(9) satisfies the time dependence notion discussed in Section 2, however it impose stricter conditions on the autocovariances of the FPC scores:*

1. *The FAR process (8)-(9) does not posses the long range dependence property (i.e., $\beta > 1$). Indeed, condition (9) implies $\sum_{h=1}^{\infty} \left\| \rho^h \right\|_{\mathcal{L}} < \infty$ which in turn implies $\sum_{h=1}^{\infty} \left\| \Gamma_h \right\|_{\mathcal{L}} < \infty$. Using expression (11) one can conclude that $\sum_{h=1}^{\infty} \left\| \Gamma_h \right\|_{\mathcal{L}} < \infty$ if $\beta > 1$.*

2. *The autocovariances of the FPC scores $\mathbb{E}[\theta_{i,\ell}\theta_{i-h,\ell}]$ decay faster then the variances $\mathbb{E}[\theta_{i,\ell}\theta_{i,\ell}]$ across spectral dimension ℓ. To see this note that the autoregressive operator ρ admits the representation*

$$\rho(x) = \sum_{\ell=1}^{\infty} \sum_{s=1}^{\infty} a_{\ell,s} \langle \psi_\ell, x \rangle \psi_s, \;\; \textit{with } x \in H, \tag{14}$$

where $a_{\ell,s} = \mathbb{E}\left[\theta_{i,\ell}, \theta_{i-1,s}\right] \lambda_\ell^{-1}$ denote the spectral coefficients. Further, we adopt the approach of Hall and Horowitz (2007) for functional linear regressions and substitute Assumption 1 (**ii**) *with one, that allows us to control the decrease of the spectral coefficients $a_{\ell,s}$ with more flexibility (see Assumption 3.3 in Hall and Horowitz, 2007). That is, instead of Assumption 1* (**ii**) *assume there exists a constant $B > 0$, some $\beta > 1$ and $\gamma > 1/2 + \alpha$ such that for all $\ell \geq 1$,*

$$B_{\ell,s}^{(h)} \leq B \; h^{-\beta} \ell^{-\gamma} s^{-\gamma}. \tag{15}$$

Then, since ρ is the Hilbert-Schmidt operator we have $\sum_{s=1}^{\infty} \sum_{\ell=1}^{\infty} a_{\ell,s}^2 < \infty$. The squared summability of $a_{\ell,s}$ is assured if and only if $\gamma > 1/2 + \alpha$. In turn, the autocovariances of the FPC scores behave as $\mathbb{E}[\theta_{i,\ell}\theta_{i-h,\ell}] = O(\ell^{2\gamma})$ and decay faster then the variances $\mathbb{E}[\theta_{i,\ell}\theta_{i,\ell}] = O(\ell^{\alpha})$.

The following result shows the consistency of $\widehat{\rho}_N$ and its speed of convergence.

Theorem 2 *If a FAR process* (8)-(9) *satisfies Assumption 1* (**i**) *and* (**iii**) *with joint cumulants up to order* 4*, and condition (15) then*

$$\|\widehat{\rho}_N - \rho\|_{\mathcal{L}} = O_p\left(\max\left\{\frac{L^{2\alpha+\frac{3}{2}}}{\sqrt{N}}, L^{1+2(\alpha-\gamma)}\right\}\right). \tag{16}$$

The rate of convergence for the estimator of the autoregressive operator consists of two parts. The first one, $\frac{L^{2\alpha+\frac{3}{2}}}{\sqrt{N}}$, characterizes the convergence of estimator $\widehat{\rho}_N$ to the truncated true operator $\rho_L = \Gamma_1 C_L^{-1}$. Moreover, it restricts L for the estimator $\widehat{\rho}_N$ to be consistent such that $L = o\left(N^{1/(4\alpha+3)}\right)$ and $L \to \infty$ as $N \to \infty$. The second part, $L^{1+2(\alpha-\gamma)}$, describes asymptotic behaviour of the reminder $\|\rho_L - \rho\|_{\mathcal{L}}$, which converge in probability to zero since $1+2(\alpha-\gamma) < 0$. Note that the fastest convergence rate $O_p\left(N^{-1/2}\right)$ can be achieved when space H is finite dimensional which is inline with the results for the OLS estimator of stationary multivariate autoregressive models (such as VAR for instance).

5 Forecasting Nonlinear FTS

As the correct model specification for FTS is not known in practice it might be too restrictive to assume a linear modeling framework, as for instance, FAR model. For this reason, in this section we propose a simple, yet robust and versatile approach to tackle potential nonlinearity in FTS. We use the functional additive approach of Müller and Yao (2008) to generalize FAR(1) model (8) and rewrite it as a functional additive autoregressive model. Using equation (14) the FAR model can be rewritten as standard linear regression model with infinitely many FPC score as predictors,

$$\mathbb{E}\left[X_{i+1}|X_i\right] = \sum_{s=1}^{\infty}\sum_{\ell=1}^{\infty} a_{\ell,s}\theta_{i,\ell}\psi_s,$$

In particular, the relationship between the response and predictor scores is modeled linearly as $\mathbb{E}\left[\theta_{i+1,s}|X_i\right] = \sum_{s=1}^{\infty} a_{\ell,s}\theta_{i,\ell}$. Furthermore, the linear framework of the FAR model and the uncorrelatedness of the FPS scores imply that $\mathbb{E}\left[\theta_{i+1,s}|\theta_{i,\ell}\right] = a_{\ell,s}\theta_{i,\ell}$. As suggested in Müller and Yao (2008), this model can be generalized by replacing the linear terms $a_{\ell,s}\theta_{i,\ell}$ by functional counterparts $m_{\ell,s}(\theta_\ell)$. This transforms the FAR model into a functional additive autoregressive model (FAAR)

$$\mathbb{E}\left[X_{i+1}|X_i\right] = \sum_{s=1}^{\infty}\sum_{\ell=1}^{\infty} m_{\ell,s}(\theta_{i,\ell})\psi_s, \tag{17}$$

where it is assumed that $\mathbb{E}[m_{\ell,s}(\theta_{i,\ell})] = 0$ for all $\ell, s \geq 1$ to assure identifiability. We impose a mild restriction on the model (17). Let the random principal component scores

$\theta_{i,\ell}$ have unconditional probability density function $f_\ell(\theta_{i,\ell})$, and write $f_{\ell,s}(\theta_{i+1,s}|\theta_{i,\ell})$ for the conditional probability density of $\theta_{i+1,s}$ given $\theta_{i,\ell}$.

Assumption 2 *$m_{\ell,s}(\cdot)$, $f_\ell(\cdot)$ and $f_{\ell,s}(\cdot)$ are twice continuously differentiable and $f_\ell(\cdot)$, and $f_\ell(\cdot)$ are bounded. Furthermore, the functional principal component scores $\theta_{i,\ell}$ and $\theta_{i,s}$ are independent for $\ell \neq s$.*

That is, the only requirement for functions $m_{\ell,s}(\cdot)$ is smoothness. Further, Assumption 2 strengthens contemporaneous uncorrelatedness of the FPC scores to independency. This in turn implies that

$$\mathbb{E}\left[\theta_{i+1,s}|\theta_{i,\ell}\right] = \mathbb{E}\left[\mathbb{E}\left[\theta_{i+1,s}|X_i\right]|\theta_{i,\ell}\right] = \mathbb{E}\left[\sum_{q=1}^{\infty} m_{q,s}(\theta_{i,q})|\theta_{i,\ell}\right] = m_{\ell,s}\left(\theta_{i,\ell}\right).$$

The simple and flexible framework of model (17) provides us with a non-linear alternative to the FAR model. In particular, representation (17) motivates a straightforward forecasting scheme to predict the expected value of X_{N+1} through estimates of the conditional means $m_{\ell,s}(\theta_{N,\ell})$. Define the predictor $M(X_N) := \mathbb{E}\left[X_{N+1}|X_N\right]$. Then using the approximation $\widehat{X}_{i,L} = \sum_{\ell=1}^{L} \widehat{\theta}_{i,\ell}\widehat{\psi}_\ell$ instead of real functions X_i the estimator of $M(X_N)$ can be constructed as

$$\widehat{M}_{N,L}(\widehat{X}_{N,L}) = \sum_{\ell=1}^{L}\sum_{s=1}^{L} \widehat{m}_{\ell,s}(\widehat{\theta}_{N,\ell})\widehat{\psi}_s, \tag{18}$$

where L is set to be a function of N such that $L \to \infty$ as $N \to \infty$. While the estimation of the functional principal components ψ_ℓ and $\theta_{i,\ell}$ has already been discussed in Section 3, we propose in the following section an estimator for the conditional means $m_{\ell,s}(\theta_{i,\ell})$.

5.1 k-Nearest Neighbors Estimator

In this section a simple method based on the k-nearest neighbors approach (KNN) is suggested to estimate predictor $M(X_N)$. The main idea behind forecasting with KNN is to identify the past observations of the time series that are most similar (in terms of some distance) to the last onservation and use a combination of their future values to predict the next value of the series.

If FTS satisfies model (17) and Assumptions 1 and 2 then the KNN method can be adopted directly to the series of the FPC scores. The estimation procedure consists of three basic steps:

1. Use data $X_1, ..., X_N$ and the FPC analysis to compute estimates $\widehat{\psi}_\ell$, $\widehat{\lambda}_\ell$ and FPC scores $\{\widehat{\theta}_{i,\ell}\}_{i=1}^{N}$ for $\ell = 1, ..., L$ (as described in Section 3).

2. Compute the distance between the most recent FPC score $\widehat{\theta}_{N,\ell}$ and each element in the rest of the score series $\{\widehat{\theta}_{i,\ell}\}_{i=1}^{N-1}$. A typical choice for this task Minkowski distance. Denote the index set of the k_N closest neighbors to the feature score component $\widehat{\theta}_{N,\ell}$ by $\mathcal{I}(k_N; \widehat{\theta}_{N,\ell})$, where the number of neighbors depends on sample size N such that $k_N \to \infty$ as $N \to \infty$.

3. Once the k_N closest elements are identified their subsequent values are averaged to obtain the final estimator, i.e.,

$$\widehat{m}_{\ell,s}(\widehat{\theta}_{N,\ell}) := \frac{1}{k_N} \sum_{i \in \widehat{\mathcal{I}}\left(k_N; \hat{\theta}_{N,\ell}\right)} \hat{\theta}_{i+1,\ell}, \tag{19}$$

   for $\ell, s = 1, ..., L$.

Substituting estimates $\widehat{m}_{\ell,s}(\widehat{\theta}_{N,\ell})$ and $\widehat{\psi}_s$ where $\ell, s = 1, ..., L$ back to (18) gives the functional predictor. Note that KNN estimator (19) is presented with equal weights $1/k_N$. Alternative weighting schemes can be considered as well. For instance, weights can be set to be inversely proportional to the distance between the last observation $\widehat{\theta}_{N,\ell}$ and a neighbor from $\widehat{\mathcal{I}}\left(k_N; \hat{\theta}_{N,\ell}\right)$, i.e.,

$$w_i = \frac{\frac{1}{d_i}}{\sum\limits_{j=1}^{k_N} \frac{1}{d_j}},$$

where d_i is a distance between $\widehat{\theta}_{N,\ell}$ and a neighbor $i \in \widehat{\mathcal{I}}\left(k_N; \hat{\theta}_{N,\ell}\right)$.

5.2 Asymptotic properties of FKNN

We split the investigation of the asymptotic properties of predictor (18)-(19) for FAAR model into two parts as follows. Consider the *infeasible* estimator of $m_{\ell,s}(\theta_\ell)$ given by

$$\widetilde{m}_{\ell,s}(\theta_{N,\ell}) := \frac{1}{k_N} \sum_{i \in \mathcal{I}\left(k_N; \theta_{N,\ell}\right)} \theta_{i+1,\ell}.$$

where all quantities of spectral decomposition, λ_ℓ, ψ_ℓ and $\theta_{i,\ell}$ are assumed to be known. Consequently, the *infeasible* functional predictor $M_{N,L}(x_L)$ with the additional smoothing step based on a approximation $X_{i,L}(t) = \sum_{\ell=1}^{L} \theta_{i,\ell}\psi_\ell(t)$ is defined by

$$M_{N,L}\left(X_{N,L}\right) := \sum_{\ell=1}^{L} \sum_{s=1}^{L} \widetilde{m}_{\ell,s}(\theta_{N,\ell})\psi_s.$$

Then to obtain the convergence rate of the estimator (18)-(19) to the true predictor it suffices to obtain the convergence rate of infeasible estimator to the true predictor,

$\mathbb{E}\|M_{N,L}(X_{N,L}) - M(X_N)\|^2$, and convergence rate of the feasible estimator (18)-(19) to infeasible one, $\mathbb{E}\left\|\widehat{M}_{N,L}(\hat{x}_L) - M_{N,L}(x_L)\right\|^2$. The following theorems present the respective convergence rates.

Theorem 3 *Let a weakly stationary FTS* $\{X_i\}_{i=1}^N$ *fulfills Assumption 1 with joint cumulants up to order* 4*, Assumption 2 and follows model* (17)*. Moreover, it is assumed that* $L^{\alpha-1}\sum_{\ell=L}^{\infty}\mathbb{E}\left[m_{\ell,s}^2(\theta_{i,\ell})\right] = O\left(\lambda_s\right)$*. Then we have*

$$\mathbb{E}\|M_{N,L}(X_{N,L}) - M(X_N)\|^2 = O\left(\max\left\{k_N^{-1}, L^{1-\alpha}\right\}\right),$$

where $k_N \sim N^{4/5}$*.*

Theorem 4 *If a weakly stationary FTS* $\{X_i\}_{i=1}^N$ *fulfills Assumption 1 with joint cumulants up to order* 6*, Assumption 2 and follows model* (17) *then*

$$\mathbb{E}\left\|\widehat{M}_{N,L}(\hat{x}_L) - M_{N,L}(x_L)\right\|^2 = O\left(\frac{L^{3+2\alpha}\log(N)}{N^{2\beta^*}}\right), \tag{20}$$

where $\beta^* = \min\{\beta, 1/2\}$*.*

The result of Theorem 3 implies that the infeasible estimator is consistent and its convergence rate consists of two parts. The first part, k_N^{-1}, describes the convergence of the infeasible estimator to the truncated true predictor $M_L(X_{N,L}) = \sum_{s,\ell=1}^{L} m_{\ell,s}(\theta_{N,\ell})\psi_s$. It also shows that the consistency result requires the number of neighbors to be the function of the sample size such that $k_N \sim N^{4/5}$. The second one characterizes the convergence of the remainder $\mathbb{E}\|M_L(X_{N,L}) - M(X_N)\|^2$ which is of order $O\left(L^{1-\alpha}\right)$.

Theorem 4 delivers the convergence between feasible and infeasible estimators. One benefit of this result is that it allows us to state the restrictions on the principal component cutoff L. It is required that $L = o\left(N^{2\beta^*/(2\alpha+3)}/\log(N)^{1/(2\alpha+3)}\right)$ and $L \to \infty$ as $N \to \infty$ to obtain the consistent FAAR predictor.

6 Small Sample Performance

We now turn to study the small-sample properties of the proposed models. The objective of this section is twofold. The first objective is to evaluate the forecasting performance of the FAR and the FAAR frameworks in different setups, relating to the asymptotic results obtained in Sections 4 and 5. The second one is to conduct a comparison of the proposed models with other alternatives available in the related forecasting/functional literature. The last aspect is covered by examining the comparative forecast performance of the FAR model and FAAR approach with that of the

1. *VAR model.* It is natural to investigate when functional settings provide an advantage compared to standard multivariate techniques. For this reason we include the VAR method, where functional observations X_i are treated as $T \times 1$ vectors $\mathbf{X}_i = [X_i(t_1), ..., X_i(t_K)]'$. These vectors are obtained by evaluating the original functions at T equidistant points $t_s = \frac{s-1}{T-1}$, $s = 1, ..., T$ and $i = 1, ..., N$;

2. *Improved FAR* [iFAR]. This approach is suggested by Kokoszka and Zhang (2010) to control for possibly small values of $\widehat{\lambda}_\ell$ that potentially can be translated into large errors in $\widehat{\lambda}_\ell^{-1}$. It is suggested to add a positive baseline to $\widehat{\lambda}_\ell$ in (13) for $\ell \geq 2$;

3. *Multivariate score model.* This model is recently suggested by Aue et al. (2015) and is based on the standard multivariate techniques applied to the vector of scores. Here we employ the VAR model for the score series which provides a simplified and elegant alternative for the FAR model.. In what follows this method will be referred to as MSM method.

We also supplement our comparative analysis with two standard benchmarks commonly employed in functional data analysis (see, e.g., Didericksen et al., 2012). The first is *Mean prediction* [MP], where predictors are obtained as the mean of the sample $\widehat{X}_{N+1} = \frac{1}{N}\sum_{i=1}^{N} X_i$, and the second is *Naive Prediction* [NP] given as $\widehat{X}_{N+1} = X_N$.

We use the FAR(1) model as the *main benchmark* design for FTS processes

$$X_i(t) = \int_0^1 \rho(t,s) X_{i-1}(s) \mathrm{d}s + \varepsilon_i(t), \tag{21}$$

for $i = 1, ..., N$. The error terms are generated as Brownian bridges

$$\varepsilon_i(t) = W(t) - tW(1), \tag{22}$$

where $W(\cdot)$ is the standard Wiener process generated as $W(\frac{k}{K}) = \frac{1}{\sqrt{K}} \sum_{j=1}^{k} Z_j$ for $k = 1, ..., K$ and Z_j are independent standard normals.

Three different forms of the kernel $\rho(t,s)$ are used: $\rho(t,s) = Ce^{\frac{-(t^2+s^2)}{2}}$, $\rho(t,s) = C$ and $\rho(t,s) = Ct$. In all cases the constant C is chosen such that $\|\rho\|_{\mathcal{S}} = 0.5$. Samples of size $N = 50, 100$ and 200 have been generated with a burn-in period of 100 functional observations. In all cases $N-1$ observations where used for the estimation and on the last observation a one-step ahead forecast was computed. All results were repeated $N_r = 1000$ times. For the FAAR model, the number of nearest neighbors k_N was set to $N^{4/5}$ as suggested by Theorem 3. To estimate and forecast with the VAR model the size of the grid has to be specified and the following rule was applied $T = 0.1N$. Finally, to measure the forecasts performance, the mean squared error (MSE) and the mean median error (MME) were computed, i.e.,

$$MSE \equiv \frac{1}{N_r}\sum_{j=1}^{N_r}\|X^j_{N+1}-\widehat{X}^j_{N+1}\|^2, \tag{23}$$

$$MME \equiv \frac{1}{N_r}\sum_{j=1}^{N_r}\int_0^1 \left|X^j_{N+1}(s)-\widehat{X}^j_{N+1}(s)\mathrm{d}s\right|, \tag{24}$$

where X^j_{N+1} and $\widehat{X}^j_{N+1}$ represent real observations and obtained forecasts, respectively, for j's replication. It should be mentioned that we used two approaches to estimat the number of FPC L. First, L is selected such that FPCs explain at least 99% or 95% of the variability in the sample. Second, we apply the selection criteria suggested in Aue et al. (2015). We report that the second approach provides forecasts with smaller MSE and MME errors. Therefore, the results based on the first approach are omitted here and are available upon request.

We report our results in the form of boxplots of the errors MSE and MME for different sample sizes and kernels. Figures 1, 2 and 3 present the results for the case when the kernel is given as $\rho(t,s)=Ce^{\frac{-(t^2+s^2)}{2}}$, $\rho(t,s)=C$ and $\rho(t,s)=Ct$, respectively. All models based on functional observations (e.g., FAAR, FAR, iFAR and MSM) perform significantly better than the benchmark predictors and the VAR model, except for the special case when $\rho(t,s)=C$. In this setup, the mean predictor provides the best forecasting results due to the structure of the DGP. In general, none of the FAR, iFAR and MSM dominates the others, while the FAAR model has marginally higher median and variance of the forecast errors. This stems from the fact that the aim of the FAAR model is to forecast general autoregressive processes while FAR, iFAR and MSM are explicitly tailored for the considered FAR DGP.

7 Forecasting electric load demand in the Nordic countries

In this section we are considering the prediction of daily electric load demand curves in the Nordic countries from a functional perspective. This problem has been of high interest to decision makers in the energy sector and has seen numerous contributions in the statistical literature. Traditionally, parametric time series models have been applied to this problem - both classical time series methods and machine learning type methods such as artificial neural networks and support vector machines (see, e.g., Kyriakides and Polycarpou, 2007, Feinberg and Genthliou, 2005, Hippert et al. (2001) and Chen et al. (2004) among others). This section describes the implementation and comparison of the

FTS models discussed in Section 6.

The data that is used in this application has been provided by Nord Pool Spot AS, the energy exchange of the Nordic and Baltic countries in Oslo, Norway [2]. Hourly demand data is made available for Denmark, Finland, Norway and Sweden since 2013. The time stamps of the raw data are converted to UTC such that every day has always 24 hours. That is, our sample for each country consist of $N = 987$ daily observations from January 1, 2013 till September 15, 2015, where each one is observed at 24 equidistant time points (e.g., hourly). Figure 4 plots a typical daily observation in a summer period. Further, a visual inspection of the data reveals that the level of the electricity demand significantly changes between different seasons of the year. Therefore, the data was centered and adjusted for monthly seasonality by subtracting from each observation the corresponding monthly average. Figure 5 plots the seasonal monthly components for each country.

Since we treat discrete observations as realizations of continuous functions, a preliminary smoothing step is required to reconstruct the underlying functional observations. For reconstruction of the deseasonalized load demand functions we consider a basis representation in terms of fourth-order B-splines with knots placed at each observed hour. Thus, the number of employed basis functions is 24 per curve. This amount of basis functions leads inevitably to overfitting the data and we thus penalize the sum of squared errors for roughness (as measured through the squared second derivative). The optimal choice of the smoothing parameter λ can be determined through minimizing a generalized cross-validation criterion (GCV). The FDA package offered by Ramsay et al. (2009) for the Matlab was applied here.[3]

We start with the report on the estimation of the functional principal components. For each country the first three principle components combined account for more than 90% of the total variation in the sample. Figure 6 plots the eigenfunctions and their respective percentages. Further, an analysis of the estimated score series provides evidence of the time dependencies for each sample. In particular, we verify the presence of the dependencies by looking at autocovariances and partial autocovariances of the score series. Figure 7 illustrates our findings for the first FPC score series.

We apply FAAR, FAR, iFAR, MSM, VAR models and benchmark models such as the naive prediction and the mean prediction to obtain forecasts for the deseasonalized electric load demand functions. The original sample is split into two parts. The first one from January 1, 2013 till December 31, 2014 is reserved for the estimation and learning purposes and the second for the evaluation of the one step ahead forecast performance. Finally, MSE and MME given in (24) and (24), respectively, are used for the comparison of the quality of the competing procedures. The number of principal components and

[2] http://www.nordpoolspot.com/historical-market-data/

[3] http://www.psych.mcgill.ca/misc/fda/downloads/FDAfuns/

lags is selected according to the selection criteria suggested in Aue et al. (2015). Further, more attention is paid to choosing the number of neighbors for the predictor in the FAAR model. More precisely, we forecast the last observation in the estimating part of the sample using (18)-19 with different values of $k_N = 1, ..., N^{4/5}$. Then the number k_N is selected to minimize the MSE between the obtained predictors and the last observation.

The results are reported in Figure 7 in the form of boxplots of the MSE and the MME errors. In general the MSM model is the best framework for forecasting electricity demand in Nordic countries except Denmark. In the case of Denmark the FAAR model provides forecasts with smaller errors when compared to MSM and for other cases is a runner-up. This finding indicates that there is a nonlinear response of the FPC score series to the past observations. This statement is also supported by the evidence from scatter plots illustrated in Figure 9. The bold lines show the best polynomial fit of order 3. In all countries but Denmark we can see that the relationship between the current first FPC score value and its lag is linear. Finally, FAR, iFAR and VAR models deliver equally good results and in general are able to outperform the naive predictors.

8 Conclusion

In this paper a time dependence concept for functional observations is proposed. It is based on the idea of the Karhunen-Loève decomposition of functional observations which gives us the vector valued time series of FPC scores. In particular, time dependence in FTS is quantified through the autocovariances and cumulants of its FPC scores series. To operate with this concept in practice we show that the estimates of the FPCs are consistent under the described dependencies. Further, two forecasting techniques for functional time series are discussed. The first one is the FAR model for processes that have a linear relation with the past observations. We then extend this linear framework using the functional additive approach suggested in Müller and Yao (2008) and offer a simple forecasting technique based on the kNN approach. Asymptotic consistency is derived. Further our simulations indicate that the loss of efficiency against the FAR model when the true underlying DGP is linear is only marginal.

A Appendix: Auxiliary results

To economize notations we use $\sum_{i,j=1}^{N}$ and $\sum_{i\neq j=1}^{N}$ instead of full expressions $\sum_{i=1}^{N}\sum_{j=1}^{N}$ and $\sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N}$ throughout this appendix. Further, the following combinatorial representation of p-th order moments in terms of joint cumulants is often used for proofs and is stated here for future reference. For a set of random variables $x_1, \ldots, x_p$ one has

$$\mathbb{E}\left[x_1 \cdot \ldots \cdot x_p\right] = \sum_{\pi} \prod_{B \in \pi} \kappa(x_i : i \in B), \tag{A.1}$$

were π cycles through all possible partitions of the set $\{1, 2, \ldots, p\}$ and B cycles through all blocks of partition π. For instance, zero mean random variables satisfies the following expressions: $\kappa(x_1,x_2) = \mathbb{E}\left[x_1, x_2\right]$ for $p = 2$, $\kappa(x_1,x_2,x_3) = \mathbb{E}\left[x_1, x_2, x_3\right]$ for $p = 3$ and

$$\begin{aligned} \kappa(x_1,x_2,x_3,x_4) &= \mathbb{E}\left[x_1, x_2, x_3, x_4\right] - \mathbb{E}\left[x_1, x_2\right]\mathbb{E}[x_3, x_4] \\ &- \mathbb{E}\left[x_1, x_3\right]\mathbb{E}[x_2, x_4] - \mathbb{E}\left[x_1, x_4\right]\mathbb{E}[x_2, x_3]\,. \end{aligned}$$

To facilitate understanding of the following proofs we collect intermediate steps into auxiliary Lemmas.

Lemma A.1 *Let a weakly stationary FTS $\{X_i\}_{i=1}^{N}$ satisfies Assumption 1 with $q = 4$ then*

$$\sup_{\ell \geq 1} \left|\hat{\lambda}_\ell - \lambda_\ell\right| \leq \left\|\widehat{C}_N - C\right\|_{\mathcal{L}}, \tag{A.2}$$

$$\left\|c_\ell \widehat{\psi}_\ell - \psi_\ell\right\| \leq C\delta_\ell \left\|\widehat{C}_N - C\right\|_{\mathcal{L}}, \text{ for } 2 \leq \ell \leq L \tag{A.3}$$

where $c_\ell = sign\left(\langle \widehat{\psi}_\ell, \psi_\ell\rangle\right)$, $\delta_\ell = \max_{1\leq k\leq \ell}\left(\lambda_k - \lambda_{k+1}\right)^{-1}$, and C is some positive constant.

Proof. Both results (A.2) and (A.3) follow from Bosq (2000, Lemma 4.2 and 4.3), respectively. ■

Lemma A.2 *A FAR process (8)-(9) satisfies Assumption 1* **(i)** *and* **(iii)** *with joint cumulants up to order 4, and condition (15) then:*

(i) $\frac{1}{N}\sum_{i=1}^{N}\|X_i\|^2 = \sum_{\ell=1}^{\infty}\lambda_\ell + O_p\left(N^{-1/2}\right)$;

(ii) $\widehat{\lambda}_L^{-1} = O_p\left(L^{\alpha}\right)$ *as* $N \to \infty$, $L \to \infty$ *and* $\frac{L^{\alpha}}{N^{1/2}} \to 0$;

(iii) $\left\|\widehat{\Gamma}_{1,N}\right\|_{\mathcal{L}} = O_p(1)$;

(iv) $\left\|\widehat{\Gamma}_{1,N}\left(\widehat{\psi}_\ell\right)\right\| \leq 2\widehat{\lambda}_\ell^{1/2}\left(\frac{1}{N}\sum_{i=1}^{N}\|X_i\|^2\right)^{1/2}$;

(**v**) $\sum_{\ell=L}^{\infty}\left\|\rho\left(\widehat{\psi}_\ell\right)\right\|^2 = O_p\left(\max\left\{\frac{L^{2+\alpha}}{N^{1/2}}, L^{1+2(\alpha-\gamma)}\right\}\right)$;

Proof.

Proof of item (i): To establish item (**i**) we show that $\mathbb{E}\left|\frac{1}{N}\sum_{i=1}^{N}\|X_i\|^2 - \sum_{\ell=1}^{\infty}\lambda_\ell\right|^2 = O\left(N^{-1}\right)$ and then by Chebyshev inequality (**i**) will follow. First, notice that $\frac{1}{N}\sum_{i=1}^{N}\|X_i\| = \frac{1}{N}\sum_{i=1}^{N}\sum_{\ell=1}^{\infty}\theta_{i,\ell}^2$, and denote $Z_i = \sum_{\ell=1}^{\infty}\theta_{i,\ell}^2$, $\overline{Z}_N = \frac{1}{N}\sum_{i=1}^{N}Z_i$ and $m = \sum_{\ell=1}^{\infty}\lambda_\ell$. Then

$$\begin{aligned} Var\left(\overline{Z}_N\right) &= \frac{1}{N^2}\sum_{i,j=1}^{N}\sum_{\ell,s=1}^{\infty}\mathbb{E}\left[\theta_{i,\ell}^2\theta_{j,s}^2\right] - m^2 \\ &= \frac{1}{N^2}\sum_{i,j=1}^{N}\sum_{\ell,s=1}^{\infty}\left(\kappa_{\ell,\ell,s,s}(0,0,|i-j|,|i-j|) + 2\mathbb{E}\left[\theta_{i,\ell}\theta_{j,s}\right]^2\right), \end{aligned}$$

where the last equality comes from relation (A.1). For the first term by Assumption 1(**iii**) we have

$$\frac{1}{N^2}\sum_{i,j=1}^{N}\sum_{\ell,s=1}^{\infty}\kappa_{\ell,\ell,s,s}(0,0,|i-j|,|i-j|) \le \frac{B}{N^2}\sum_{i=1}^{N}\sum_{\ell,s=1}^{\infty}\lambda_\ell\lambda_s = O\left(N^{-1}\right),$$

and for the second

$$\begin{aligned} \frac{2}{N^2}\sum_{i,j=1}^{N}\sum_{\ell,s=1}^{\infty}\mathbb{E}\left[\theta_{i,\ell}\theta_{j,s}\right]^2 &= \frac{2}{N^2}\sum_{i\neq j=1}^{N}\sum_{\ell,s=1}^{\infty}\mathbb{E}\left[\theta_{i,\ell}\theta_{j,s}\right]^2 + \frac{2}{N}\sum_{\ell=1}^{\infty}\lambda_\ell^2 \\ &\le \frac{4}{N^2}\sum_{h=1}^{N-1}\sum_{i=h+1}^{N}\sum_{\ell,s=1}^{\infty}\left(B_{\ell,s}^{(h)}\right)^2 + \frac{2}{N}\sum_{\ell=1}^{\infty}\lambda_\ell^2 \\ &\le \frac{B}{N}\sum_{h=1}^{N-1}h^{-2\beta}\sum_{\ell,s=1}^{\infty}\ell^{-\gamma}s^{-\gamma} + \frac{2}{N}\sum_{\ell=1}^{\infty}\lambda_\ell^2 = O\left(N^{-1}\right), \end{aligned}$$

where the last result comes from Assumption 1 (**i**) and (**iii**) and condition 15.

Proof of item (ii): It follows immediately from Corollary 2 and Chebyshev inequality $\widehat{\lambda}_\ell = O_p\left(\max\left\{L^{-\alpha}, N^{-1/2}\right\}\right)$ and $\widehat{\lambda}_\ell^{-1} = O_p\left(\frac{1}{\max\left\{L^{-\alpha},N^{-1/2}\right\}}\right)$. The item (**ii**) will follow from the fact $N^{-1/2}$ will go to zero faster then $L^{-\alpha}$ since $L^{\alpha}/N^{1/2} \to 0$.

Proof of item (iii): Follows from Corollary 1 and Chebyshev inequality.

Proof of item (iv): Follows from Lemma 8.3 in Bosq (2000).

Proof of item (v): Item (**v**) is obtained by using the proof from Lemma 8.2 in Bosq (2000) and the facts that $\left\|\widehat{C}_N - C\right\|_{\mathcal{L}} = O_p(N^{-1/2})$, $\sum_{\ell=1}^{L}\delta_\ell = O(L^{2+\alpha})$ and $\sum_{\ell=L}^{\infty}\|\rho\left(\psi_\ell\right)\|^2 = O\left(L^{1+2(\alpha-\gamma)}\right)$ ■

B Appendix: Proofs

Proof of Remark 1

We have

$$\begin{aligned}\mathbb{E}\left\|\hat{\mu}-\mu\right\|^2 &= \frac{1}{N^2}\sum_{i,j=1}^{N}\mathbb{E}\left\langle X_i-\mu, X_j-\mu\right\rangle = \frac{1}{N^2}\sum_{i,j=1}^{N}\sum_{\ell,s=1}^{\infty}\mathbb{E}\left[\theta_{i,\ell},\theta_{j,s}\right] \\ &= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{\ell=1}^{\infty}\mathbb{E}\left[\theta_{i,\ell},\theta_{i,\ell}\right]+\frac{1}{N^2}\sum_{i\neq j=1}^{N}\sum_{\ell,s=1}^{\infty}\mathbb{E}\left[\theta_{i,\ell},\theta_{j,s}\right].\end{aligned}$$

As a consequence of Assumption 1 part (**i**) $\sum_{\ell=1}^{\infty}\lambda_\ell<\infty$ such that the first term in the last equation above behaves as $O\left(N^{-1}\right)$. Rearranging the second term and invoking Assumption 1 (**ii**) gives

$$\begin{aligned}\frac{1}{N^2}\sum_{i\neq j=1}^{N}\sum_{\ell,s=1}^{\infty}\mathbb{E}\left[\theta_{i,\ell},\theta_{j,s}\right] &= \frac{2}{N^2}\sum_{h=1}^{N-1}\sum_{i=h+1}^{N}\sum_{\ell,s=1}^{\infty}\mathbb{E}\left[\theta_{i,\ell},\theta_{j,s}\right] \\ &\leq \frac{2}{N^2}\sum_{h=1}^{N-1}\sum_{i=h+1}^{N}\sum_{\ell,s=1}^{\infty}B_{\ell,s}^{(h)} \\ &\leq \frac{C}{N^2}\sum_{h=1}^{N-1}(N-h)h^{-\beta}\sum_{\ell,s=1}^{\infty}\sqrt{\lambda_\ell\lambda_s}=O\left(\max\left\{N^{-\beta},N^{-1}\right\}\right).\end{aligned}$$

The last equality uses Davidson (1994, Theorem 2.27) and the fact that $\sum_{\ell=1}^{\infty}\sqrt{\lambda_\ell}<\infty$ which follows from Assumption 1.

Proof of Theorem 1

We have,

$$\begin{aligned}\mathbb{E}\left\|\widehat{C}_N-C\right\|_{\mathcal{S}}^2 &= \sum_{\ell=1}^{\infty}\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\left(\left\langle X_i,\psi_\ell\right\rangle X_i-\mathbb{E}\left[\left\langle X_i,\psi_\ell\right\rangle X_i\right]\right)\right\|^2 \\ &= \frac{1}{N^2}\sum_{i,j=1}^{N}\sum_{\ell=1}^{\infty}\left(\sum_{s=1}^{\infty}\mathbb{E}\left[\theta_{i,\ell}\theta_{j,\ell}\theta_{i,s}\theta_{j,s}\right]-\lambda_\ell^2\right) && \text{(A.4)} \\ &= \frac{1}{N^2}\sum_{i,j=1}^{N}\sum_{\ell=1}^{\infty}\left(\mathbb{E}\left[\theta_{i,\ell}^2\theta_{j,\ell}^2\right]-\lambda_\ell^2\right) \\ &+ \frac{1}{N^2}\sum_{i,j=1}^{N}\sum_{\ell\neq s=1}^{\infty}\mathbb{E}\left[\theta_{i,\ell}\theta_{j,\ell}\theta_{i,s}\theta_{j,s}\right]:=a+b. && \text{(A.5)}\end{aligned}$$

It follows from relation (A.1) that

$$a = \frac{1}{N^2} \sum_{i,j=1}^{N} \sum_{\ell=1}^{\infty} \left(\kappa_{\ell,\ell,\ell,\ell}(0,0,|i-j|,|i-j|) + 2\mathbb{E}\left[\theta_{i,\ell}\theta_{j,\ell}\right]^2 \right),$$

where $\frac{1}{N^2}\sum_{i,j=1}^{N}\sum_{\ell=1}^{\infty}\kappa_{\ell,\ell,\ell,\ell}(0,0,|i-j|,|i-j|) = O\left(N^{-1}\right)$ by Assumption 1(**iii**) and

$$\begin{aligned}
\frac{2}{N^2}\sum_{i,j=1}^{N}\sum_{\ell=1}^{\infty}\mathbb{E}\left[\theta_{i,\ell}\theta_{j,\ell}\right]^2 &= \frac{2}{N^2}\sum_{i\neq j=1}^{N}\sum_{\ell=1}^{\infty}\mathbb{E}\left[\theta_{i,\ell}\theta_{j,\ell}\right]^2 + \frac{2}{N}\sum_{\ell=1}^{\infty}\lambda_\ell^2 \\
&\leq \frac{4}{N^2}\sum_{h=1}^{N-1}\sum_{i=h+1}^{N}\sum_{\ell=1}^{\infty}\left(B_{\ell,\ell}^{(h)}\right)^2 + \frac{2}{N}\sum_{\ell=1}^{\infty}\lambda_\ell^2 \\
&\leq \frac{B}{N}\sum_{h=1}^{N-1}h^{-2\beta}\sum_{\ell=1}^{\infty}\lambda_\ell^2 + \frac{2}{N}\sum_{\ell=1}^{\infty}\lambda_\ell^2 \\
&= O\left(\max\left\{N^{-2\beta}, N^{-1}\right\}\right),
\end{aligned}$$

where the last equality comes from Assumption 1(**i**) and (**ii**).

Similar arguments apply to term b, i.e.,

$$\frac{1}{N^2}\sum_{i,j=1}^{N}\sum_{\ell\neq s=1}^{\infty}\mathbb{E}\left[\theta_{i,\ell}\theta_{j,\ell}\theta_{i,s}\theta_{j,s}\right] = \frac{1}{N^2}\sum_{i,j=1}^{N}\sum_{\ell\neq s=1}^{\infty}\left(\kappa_{\ell,\ell,s,s}(0,0,|i-j|,|i-j|)+\right. \tag{A.6}$$

$$\left. + \mathbb{E}\left[\theta_{i,\ell}\theta_{j,\ell}\right]\mathbb{E}\left[\theta_{i,s}\theta_{j,s}\right] + \mathbb{E}\left[\theta_{i,\ell}\theta_{j,s}\right]\mathbb{E}\left[\theta_{i,s}\theta_{j,\ell}\right]\right) \tag{A.7}$$

by relation (A.1). The first terms on the r.h.s of (A.6) is $O\left(N^{-1}\right)$ by Assumption 1(**iii**). The second and the third terms on the r.h.s of (A.6) are $O\left(\max\{N^{-2\beta}, N^{-1}\}\right)$ by the same arguments as above. In particular, for the third term we have

$$\begin{aligned}
\frac{1}{N^2}\sum_{i,j=1}^{N}\sum_{\ell\neq s=1}^{\infty}\mathbb{E}\left[\theta_{i,\ell}\theta_{j,s}\right]\mathbb{E}\left[\theta_{i,s}\theta_{j,\ell}\right] &\leq \frac{1}{N^2}\sum_{i,j=1}^{N}\sum_{\ell\neq s=1}^{\infty}\left(B_{\ell,s}^{(i-j)}\right)^2 = \frac{2}{N^2}\sum_{h=1}^{N-1}\sum_{i=h+1}^{N}\sum_{\ell\neq s=1}^{\infty}\left(B_{\ell,s}^{(h)}\right)^2 \\
&\leq \frac{B}{N}\sum_{h=1}^{N-1}h^{-2\beta}\sum_{\ell\neq s=1}^{\infty}\lambda_\ell\lambda_s = O\left(\max\left\{N^{-2\beta}, N^{-1}\right\}\right).
\end{aligned}$$

Putting together rates for a and b yields the statement of the theorem.

Proof of Theorem 2

Recall that $H_L = span\{\psi_1, ..., \psi_L\}$ and let $\widehat{H}_L = span\{\widehat{\psi}_1, ..., \widehat{\psi}_L\}$ and denote π_L and $\widehat{\pi}_L$ projections on H_L and $\widehat{H}_L$, respectively. Then we can consider the following decomposition

$$\begin{aligned}(\widehat{\rho}_N - \rho)(x) &= (\widehat{\rho}_N - \rho\pi_L(x)) + (\rho\pi_L(x) - \rho\widehat{\pi}_L(x)) + (\rho\widehat{\pi}_L(x) - \rho(x)) \\ &:= a_N(x) + b_N(x) + c_N(x).\end{aligned}$$

Further, denote $a_N(x) = \sum_{k=1}^{4} a_{k,N}(x)$, where

$$\begin{aligned}a_{1,N}(x) &= \widehat{\Gamma}_{1,N}\left(\sum_{\ell=1}^{L}\left(\widehat{\lambda}_\ell^{-1} - \lambda_\ell^{-1}\right)\langle x, \widehat{\psi}_\ell\rangle\widehat{\psi}_\ell\right), \\ a_{2,N}(x) &= \widehat{\Gamma}_{1,N}\left(\sum_{\ell=1}^{L}\lambda_\ell^{-1}\left(\langle x, \widehat{\psi}_\ell\rangle - \langle x, \psi'_\ell\rangle\right)\widehat{\psi}_\ell\right), \\ a_{3,N}(x) &= \widehat{\Gamma}_{1,N}\left(\sum_{\ell=1}^{L}\lambda_\ell^{-1}\langle x, \psi'_\ell\rangle\left(\widehat{\psi}_\ell - \psi'_\ell\right)\right), \\ a_{4,N}(x) &= \left(\widehat{\Gamma}_{1,N} - \Gamma\right)\left(\sum_{\ell=1}^{L}\lambda_\ell^{-1}\langle x, \psi'_\ell\rangle\psi'_\ell\right).\end{aligned}$$

For the first term we have

$$\|a_{N,1}(x)\| \leq \sum_{\ell=1}^{L}\frac{|\widehat{\lambda}_\ell - \lambda_\ell|}{\widehat{\lambda}_\ell\lambda_\ell}|\langle x, \widehat{\psi}_\ell\rangle|\left\|\widehat{\Gamma}_{1,N}(\widehat{\psi}_\ell)\right\|.$$

Using (A.2), Cauchy-Schwartz inequality and item (**iv**) of Lemma A.2 we obtain

$$\|a_{N,1}\|_{\mathcal{L}} \leq 2\left(\frac{1}{N}\sum_{i=1}^{N}\|X_i\|^2\right)^{1/2}\|C_N - C\|_{\mathcal{L}}\left(\sum_{\ell=1}^{L}\widehat{\lambda}_\ell^{-1/2}\lambda_\ell^{-1}\right).$$

From Theorem 1 and Chebyshev inequality $\|C_N - C\|_{\mathcal{L}} = O_p(N^{-1/2})$. Assume for now that $L^\alpha/N^{1/2} \to 0$, then by using item (**i**) and (**ii**) of Lemma A.2 one gets

$$\|a_{N,1}\|_{\mathcal{L}} = O_p\left(\frac{L^{\frac{3}{2}\alpha+1}}{N^{1/2}}\right). \tag{A.8}$$

Finally, to archive the consistency it is required that $L^{\frac{3}{2}\alpha+1}/N^{1/2} \to 0$ which in turn implies the condition $L^\alpha/N^{1/2} \to 0$ has to hold. That is, $L^\alpha/N^{1/2} \to 0$ is necessary but not sufficient to obtain the statement of the theorem.

Turning to $a_{N,2}(x)$, from item (**iv**) of Lemma A.2 and Cauchy-Schwartz inequality we have

$$\|a_{N,2}\|_{\mathcal{L}} \leq 2\left(\frac{1}{N}\sum_{i=1}^{N}\|X_i\|^2\right)^{1/2}\sum_{\ell=1}^{L}\widehat{\lambda}_\ell^{1/2}\lambda_\ell^{-1}\left\|\widehat{\psi}_\ell-\psi_\ell\right\|,$$

where (A.3) together with and the fact that $\sum_{\ell=1}^{L}\delta_\ell = O(L^{\alpha+2})$ yield

$$\|a_{N,2}\|_{\mathcal{L}} = O_p\left(\frac{L^{\frac{3}{2}\alpha+2}}{N^{1/2}}\right). \tag{A.9}$$

Concerning $a_{N,3}(x)$, Cauchy-Schwartz inequality and orthogonality of $\widehat{\psi}_\ell$ and ψ_ℓ yield the bound

$$\|a_{N,3}\|_{\mathcal{L}} \leq \left\|\widehat{\Gamma}_{1,N}\right\|_{\mathcal{L}}\left(\sum_{\ell=1}^{L}\lambda_\ell^{-2}\langle x,\widehat{\psi}_\ell\rangle^2\left\|\widehat{\psi}_\ell-\psi_\ell\right\|^2\right)^{1/2}.$$

Then using item (**iii**) of Lemma A.2 and the fact that $\left(\sum_{\ell=1}^{L}\sigma_\ell^2\right)^{1/2} = O(L^{\alpha+3/2})$ yield

$$\|a_{N,3}\|_{\mathcal{L}} = O_p\left(\frac{L^{2\alpha+\frac{3}{2}}}{N^{1/2}}\right). \tag{A.10}$$

Finally,

$$\|a_{N,4}\|_{\mathcal{L}} = \left\|\widehat{\Gamma}_{1,N}-\Gamma\right\|_{\mathcal{L}}\left(\sum_{\ell=1}^{L}\lambda_\ell^{-2}\langle x,\psi_\ell\rangle^2\right)^{1/2}.$$

Then Corollary 1 entail

$$\|a_{N,4}\|_{\mathcal{L}} = O_p\left(\frac{L^{\alpha+\frac{1}{2}}}{N^{1/2}}\right). \tag{A.11}$$

Next we turn to $b_N(x)$ and $c_N(x)$. First observe that

$$\|b_N\|_{\mathcal{L}} \leq C\left(\sum_{\ell=L}^{\infty}\left\|\rho\left(\widehat{\psi}_\ell\right)\right\|^2+\sum_{\ell=L}^{\infty}\|\rho\left(\psi_\ell\right)\|^2\right). \tag{A.12}$$

which behave as $O_p\left(\max\left\{\frac{L^{2+\alpha}}{N^{1/2}}, L^{1+2(\alpha-\gamma)}\right\}\right)$ by item (**v**) of Lemma A.2. For $c_N(x)$ we have $\|c_N\|_{\mathcal{L}} = \sum_{\ell=L}^{\infty}\|\rho\left(\psi_\ell\right)\|^2 = O_p\left(L^{1+2(\alpha-\gamma)}\right)$ and statement of the theorem is proofed.

Proof of Theorem 3

First, define $M_L(X_{N,L}) := \sum_{s,\ell=1}^{L} \mathbb{E}\left[\theta_{N+1,s}|\theta_{N,\ell}\right]\psi_s = \sum_{s,\ell=1}^{L} m_{\ell,s}(\theta_{N,\ell})\psi_s$, where in comparison to $M_{N,L}\left(x_L\right)$ the k_N-NN estimators of the scores have been replaced by the corresponding conditional population means. Since our interest is in analyzing $\mathbb{E}\|M_{N,L}(X_{N,L}) - M(X_N)\|^2$, it suffices, upon adding and subtracting $M_L(X_{N,L})$ in the argument of our object of interest, to consider the two terms

$$\mathbb{E}\|M_L(X_{N,L}) - M(X_N)\|^2 \text{ and } \mathbb{E}\|M_{N,L}(X_{N,L}) - M_L(X_{N,L})\|^2 \tag{A.13}$$

For simplicity of notation let θ_ℓ denote $\theta_{N,\ell}$. Then for the first term in (A.13) by using the orthonormality of the $\{\psi_\ell\}$ we have

$$\begin{aligned}\mathbb{E}\|M_L(X_{N,L}) - M(X_N)\|^2 &= \mathbb{E}\left\|\sum_{s,\ell=1}^{L} m_{\ell,s}(\theta_\ell)\psi_\ell - \sum_{s,\ell=1}^{\infty} m_{\ell,s}(\theta_\ell)\psi_\ell\right\|^2 \\ = \sum_{s,\ell=L+1}^{\infty} \mathbb{E}\left[m_{\ell,s}(\theta_\ell)^2\right] &+ \sum_{s=L+1}^{\infty}\sum_{\ell=1}^{L} \mathbb{E}\left[m_{\ell,s}(\theta_\ell)^2\right] + \sum_{s=1}^{L}\sum_{\ell=L+1}^{\infty} \mathbb{E}\left[m_{\ell,s}(\theta_\ell)^2\right].\end{aligned} \tag{A.14}$$

Now observe that from $L^{\alpha-1}\sum_{\ell=L}^{\infty} \mathbb{E}\left[m_{\ell,s}^2(\theta_{i,\ell})\right] = O\left(\lambda_s\right)$ it follows immediately that $\sum_{s,\ell=L+1}^{\infty} \mathbb{E}\left[m_{\ell,s}(\theta_\ell)^2\right] = O(L^{2(1-\alpha)})$, $\sum_{s=L+1}^{\infty}\sum_{\ell=1}^{L} \mathbb{E}\left[m_{\ell,s}(\theta_\ell)^2\right] = O(L^{1-\alpha})$ and $\sum_{s=1}^{L}\sum_{\ell=L+1}^{\infty} \mathbb{E}\left[m_{\ell,s}(\theta_\ell)^2\right] = O(L^{1-\alpha})$

Now we consider the second term in (A.13) which can be written as

$$\begin{aligned}\mathbb{E}\|M_{N,L}(X_{N,L}) - M_L(X_{N,L})\|^2 &= \mathbb{E}\left\|\sum_{s,\ell=1}^{L}\left(\widetilde{m}_{\ell,s}(\theta_l) - m_{\ell,s}(\theta_l)\right)\psi_l\right\|^2 \\ &= \sum_{s,\ell=1}^{L} \mathbb{E}\left[\left(\widetilde{m}_{\ell,s}(\theta_\ell) - m_{\ell,s}(\theta_\ell)\right)^2\right],\end{aligned} \tag{A.15}$$

where the second equality follows from the orthonormality of the sequence of eigenfunctions $(\psi_\ell)_{\ell=1}^{L}$. For fixed $\ell = 1, \ldots, L$, rates of convergence of the mean squared error in (A.15) can be derived by following results in Yakowitz (1987). A careful inspection of the proofs in Yakowitz (1987) reveals that analyzing the second moment of the distance between (the given) θ_l and its farthest (of the k_N) neighbor is of key importance. Denote this farthest neighbor to θ_l by $\theta_{N(k_N),l}$ and write $R_{i,l}(\theta_l) := |\theta_{i,l} - \theta_l|$ such that $R_{(k_N),l}(\theta_l) := |\theta_{N(k_N),l} - \theta_l|$ denotes the k_N-th order statistic of the $R_{i,l}(\theta_l)$. Results in Yakowitz (1987) indicate that $\mathbb{E}[R_{(k_N),l}(\theta_l)^2] \leq C_1(l)k_N^{-1/2}$, where $C_1(l)$ is some constant that depends only on l. While this holds true for fixed l, we have to consider asymptotics where L goes to infinity. Now observe that

$$\mathbb{E}\left[R_{(k_N),l}(\theta_l)^2\right] = \mathbb{E}\left[|\theta_{N(k_N),l} - \theta_l|^2\right] \leq C_2(N)\lambda_l$$

for fixed N, where $C_2(N)$ is some constant only depending on N. Combining these results gives us $\mathbb{E}[R_{(k_N),l}(\theta_l)^2] \leq C_3 k_N^{-1/2}\lambda_l$, where now C_3 is a constant that is independent of both l and N. Moreover, Yakowitz (1987) shows that the number of neighbors k_N has to grow with the sample size where $k_N \sim \lfloor N^{4/5} \rfloor$.

The desired result now follows from the Theorem 2.1 Yakowitz (1987) and the arguments presented above.

Proof of Theorem 4

Denote, for $i = 1, \ldots, k_N$, by $N(i) \in \mathcal{I}(k_N; \theta_\ell)$ the index of the i-th nearest neighbor to θ_ℓ. Then upon adding and subtracting $\sum_{\ell,s=1}^{L} \widetilde{m}_{\ell,s}(\theta_\ell)\widehat{\psi}_s$ to the argument of $\mathbb{E}\left\|\widehat{M}_{N,L}(\hat{x}_L) - M_{N,L}(x_L)\right\|^2$ it suffices to analyze the quantities

$$\mathbb{E}\left\|\sum_{\ell,s=1}^{L} \widetilde{m}_{\ell,s}(\theta_\ell)\left(\widehat{\psi}_s - \psi_s\right)\right\|^2 \quad \text{and} \quad \mathbb{E}\left\|\sum_{\ell,s=1}^{L}\left(\widehat{m}_{\ell,s}(\hat{\theta}_\ell) - \widetilde{m}_{\ell,s}(\theta_\ell)\right)\widehat{\psi}_s\right\|^2.$$

For the first term we have

$$\begin{aligned}
\mathbb{E}\ &\left\|\sum_{\ell,s=1}^{L} \widetilde{m}_{\ell,s}(\theta_\ell)\left(\widehat{\psi}_s - \psi_s\right)\right\|^2 = \mathbb{E}\left[\sum_{\ell,s=1}^{L}\sum_{k,\tau=1}^{L} \widetilde{m}_{\ell,s}(\theta_\ell)\widetilde{m}_{k,\tau}(\theta_k)\left\langle \widehat{\psi}_s - \psi_s, \widehat{\psi}_\tau - \psi_\tau \right\rangle\right] \\
\leq\ & \mathbb{E}\left[\sum_{\ell,s=1}^{L}\sum_{k,\tau=1}^{L} \widetilde{m}_{\ell,s}(\theta_\ell)\widetilde{m}_{k,\tau}(\theta_k)\left\|\widehat{\psi}_s - \psi_s\right\|\left\|\widehat{\psi}_\tau - \psi_\tau\right\|\right] \\
\leq\ & \frac{1}{k_N^2}\sum_{\ell,s=1}^{L}\sum_{k,\tau=1}^{L}\sum_{i,j=1}^{k_N} \mathbb{E}\left[\theta_{N(i)+1,\ell}\theta_{N(j)+1,k}\delta_s\delta_\tau\left\|\widehat{C}_N - C\right\|_{\mathcal{S}}^2\right], \qquad \text{(A.16)}
\end{aligned}$$

where the last inequality follows from Lemma A.1. As already discussed in the proof of Theorem 1 we have

$$\begin{aligned}
\left\|\widehat{C}_N - C\right\|_{\mathcal{S}}^2 \ =\ & \frac{1}{N^2}\sum_{n,m=1}^{N}\left(\sum_{h_1,h_2=1}^{\infty}\theta_{n,h_1}\theta_{n,h_2}\theta_{m,h_1}\theta_{m,h_2}\right. \\
& \left. + \sum_{h_1=1}^{\infty}\lambda_{h_1}^2 - \sum_{h_1=1}^{\infty}\lambda_{h_1}\theta_{n,h_1}^2 - \sum_{h_1=1}^{\infty}\lambda_{h_1}\theta_{m,h_1}^2\right).
\end{aligned}$$

Thus the expression in (A.16) can be rewritten as

$$\frac{1}{k_N^2}\sum_{\ell,s=1}^{L}\sum_{k,\tau=1}^{L}\sum_{i,j=1}^{k_N}\mathbb{E}\left[\theta_{N(i)+1,\ell}\theta_{N(j)+1,k}\delta_s\delta_\tau\left\|\widehat{C}_N-C\right\|_{\mathcal{S}}^2\right]=A_1+A_2-2A_3,$$

where

$$\begin{aligned}
A_1 &:= \frac{1}{k_N^2N^2}\sum_{\ell,s=1}^{L}\sum_{k,\tau=1}^{L}\sum_{i,j=1}^{k_N}\sum_{n,m=1}^{N}\sum_{h_1,h_2=1}^{\infty}\delta_s\delta_\tau\mathbb{E}\left[\theta_{N(i)+1,\ell}\theta_{N(j)+1,k}\theta_{n,h_1}\theta_{n,h_2}\theta_{m,h_1}\theta_{m,h_2}\right],\\
A_2 &:= \frac{1}{k_N^2N^2}\sum_{\ell,s=1}^{L}\sum_{k,\tau=1}^{L}\sum_{i,j=1}^{k_N}\sum_{n,m=1}^{N}\sum_{h_1=1}^{\infty}\delta_s\delta_\tau\lambda_{h_1}^2\mathbb{E}\left[\theta_{N(i)+1,\ell}\theta_{N(j)+1,k}\right],\\
A_3 &:= \frac{1}{k_N^2N^2}\sum_{\ell,s=1}^{L}\sum_{k,\tau=1}^{L}\sum_{i,j=1}^{k_N}\sum_{n,m=1}^{N}\sum_{h_1=1}^{\infty}\delta_s\delta_\tau\lambda_{h_1}\mathbb{E}\left[\theta_{N(i)+1,\ell}\theta_{N(j)+1,k}\theta_{n,h_1}^2\right].
\end{aligned}$$

The analysis of the terms above now proceeds by considering the relationship between higher order moments and joint cumulants as defined in (A.1) and noting that the random variables $\theta_{\cdot,h}=\langle X_\cdot,\psi_h\rangle$ have zero mean by construction and are independent across h by assumption.

We start with term A_2. The relevant case for us to consider is $\ell=k$ as otherwise $A_2=0$ by the above arguments. Distinguishing the cases where $\ell\neq h_1$ and $\ell=h_1$ then yields

$$\begin{aligned}
A_2 &= \frac{1}{k_N^2N^2}\sum_{\ell,s=1}^{L}\sum_{\tau=1}^{L}\sum_{i,j=1}^{k_N}\sum_{n,m=1}^{N}\sum_{h_1\neq\ell=1}^{\infty}\delta_s\delta_\tau\lambda_{h_1}^2\kappa_{\ell,\ell}(0,|N(i)-N(j)|)\\
&+ \frac{1}{k_N^2N^2}\sum_{\ell,s=1}^{L}\sum_{\tau=1}^{L}\sum_{i,j=1}^{k_N}\sum_{n,m=1}^{N}\delta_s\delta_\tau\lambda_\ell^2\kappa_{\ell,\ell}(0,|N(i)-N(j)|)\\
&=: A_{2,1}+A_{2,2}. \qquad (A.17)
\end{aligned}$$

Now consider the term A_3 and again note that it suffices to consider only the case $\ell=k$. Again distinguishing the cases where $\ell\neq h_1$ and $\ell=h_1$ we have by (A.1) that

$$\begin{aligned}
A_3 &= \tfrac{1}{k_N^2N^2}\textstyle\sum_{l=1}^{L}\sum\sum_{i,j=1}^{k_N}\sum\sum_{n,m=1}^{N}\sum_{h_1=L+1}^{\infty}\delta_l^2\lambda_{h_1}^2\kappa_l(0,|N(i)-N(j)|)\\
&+\tfrac{1}{k_N^2N^2}\textstyle\sum_{l=1}^{L}\sum\sum_{i,j=1}^{k_N}\sum\sum_{n,m=1}^{N}\delta_l^2\lambda_l\kappa_{\ell,\ell}(0,|N(i)-N(j)|,|N(i)+1-n|,|N(i)+1-n|)\\
&+\tfrac{2}{k_N^2N^2}\textstyle\sum_{l=1}^{L}\sum\sum_{i,j=1}^{k_N}\sum\sum_{n,m=1}^{N}\delta_l^2\lambda_l\kappa_l(0,|N(i)+1-n|)\kappa_l(0,|N(j)+1-n|)\\
&+\tfrac{1}{k_N^2N^2}\textstyle\sum_{l=1}^{L}\sum\sum_{i,j=1}^{k_N}\sum\sum_{n,m=1}^{N}\delta_l^2\lambda_l\kappa_l(0,|N(i)-N(j)|)\kappa_l(0,0)\\
&=: A_{3,1}+A_{3,2}+A_{3,3}+A_{3,4}. \qquad (A.18)
\end{aligned}$$

Note that the term A_3 enters the object of interest twice with a negative sign, such that all terms of which A_2 is comprised are canceled in view of $A_{2,1} = A_{3,1}$ and $A_{2,2} = A_{3,4}$ and since $\kappa_l(0,0) = \lambda_l$.

We now tun to term A_1 and first decompose into the cases where $h_1 \neq h_2$ and $h_1 = h_2$. The second case is furthermore decomposed into cases where $l = k$ and $l \neq k$. This yields

$$\begin{aligned} A_1 \quad &= \tfrac{1}{k_N^2 N^2} \sum\sum_{l,k=1}^{L} \sum\sum_{i,j=1}^{k_N} \sum\sum_{n,m=1}^{N} \sum\sum_{h_1 \neq h_2}^{\infty} \delta_l \delta_k \mathbb{E}\left[\theta_{N(i)+1,l}\theta_{N(j)+1,k}\theta_{n,h_1}\theta_{n,h_2}\theta_{m,h_1}\theta_{m,h_2}\right] \\ &+ \tfrac{1}{k_N^2 N^2} \sum\sum_{l \neq k}^{L} \sum\sum_{i,j=1}^{k_N} \sum\sum_{n,m=1}^{N} \sum_{h_1=1}^{\infty} \delta_l \delta_k \mathbb{E}\left[\theta_{N(i)+1,l}\theta_{N(j)+1,k}\theta_{n,h_1}^2\theta_{m,h_1}^2\right] \\ &+ \tfrac{1}{k_N^2 N^2} \sum_{l=1}^{L} \sum\sum_{i,j=1}^{k_N} \sum\sum_{n,m=1}^{N} \sum_{h_1=1}^{\infty} \delta_l^2 \mathbb{E}\left[\theta_{N(i)+1,l}\theta_{N(j)+1,k}\theta_{n,h_1}^2\theta_{m,h_1}^2\right] \\ &=: A_{1,1} + A_{1,2} + A_{1,3}. \end{aligned} \tag{A.19}$$

Now note that $A_{1,2} = 0$ by the same arguments as above. For term $A_{1,3}$, we decompose into the cases where $l \neq h_1$ and $l = h_1$ which yields

$$\begin{aligned} A_{1,3} \quad &= \tfrac{1}{k_N^2 N^2} \sum_{l=1}^{L} \sum\sum_{i,j=1}^{k_N} \sum\sum_{n,m=1}^{N} \delta_l^2 \mathbb{E}\left[\theta_{N(i)+1,l}\theta_{N(j)+1,l}\theta_{n,l}^2\theta_{m,l}^2\right] \\ &+ \tfrac{1}{k_N^2 N^2} \sum_{l=1}^{L} \sum\sum_{i,j=1}^{k_N} \sum\sum_{n,m=1}^{N} \sum_{h_1=L+1}^{\infty} \delta_l^2 \mathbb{E}\left[\theta_{N(i)+1,l}\theta_{N(j)+1,l}\right] \mathbb{E}\left[\theta_{n,h_1}^2\theta_{m,h_1}^2\right] \end{aligned} \tag{A.20}$$

We consider first the first term of (A.20). By (A.1) and writing, with some abuse of notation, $\kappa^{(p)}$ for the p-th order cumulant, we have

$$\begin{aligned} &\mathbb{E}\left[\theta_{N(i)+1,l}\theta_{N(j)+1,l}\theta_{n,l}^2\theta_{m,l}^2\right] \\ &= \kappa_l^{(6)} + 15\kappa_l^{(4)}\kappa_l^{(2)} + 10\kappa_l^{(3)}\kappa_l^{(3)} + 15\kappa_l^{(2)}\kappa_l^{(2)}\kappa_l^{(2)}. \end{aligned}$$

There are 15 instances of $\kappa_l^{(2)}$ which are of the form

$$\begin{gathered} 1 \times \kappa_l(0,|N(i)-N(j)|) \\ 2 \times \kappa_l(0,|N(i)+1-n|) \\ 2 \times \kappa_l(0,|N(i)+1-m|) \\ 2 \times \kappa_l(|N(i)-N(j)|,|N(i)+1-n|) \\ 2 \times \kappa_l(|N(i)-N(j)|,|N(i)+1-m|) \\ 4 \times \kappa_l(|N(i)+1-n|,|N(i)+1-m|) \\ 1 \times \kappa_l(|N(i)+1-n|,|N(i)+1-n|) \\ 1 \times \kappa_l(|N(i)+1-m|,|N(i)+1-m|) \end{gathered}$$

Now note that there are precisely four instances where $\kappa_l^{(2)}$ is such that the first term in

(A.20) takes the form

$$\frac{1}{k_N^2 N^2} \sum_{l=1}^{L} \sum \sum_{i,j=1}^{k_N} \sum \sum_{n,m=1}^{N} \delta_l^2 \lambda_l \kappa_{l(0,|N(i)+1-n|)} \kappa_{l(0,|N(j)+1-n|)}$$

and precisely one instance where $\kappa_l^{(2)}$ is such that the first term in (A.20) takes the form

$$\frac{1}{k_N^2 N^2} \sum_{l=1}^{L} \sum \sum_{i,j=1}^{k_N} \sum \sum_{n,m=1}^{N} \delta_l^2 \lambda_l^2 \kappa_{l(0,|N(i)-N(j)|)}$$

which are canceled by $A_{3,3}$ and $A_{3,4}$, respectively, since these terms enters twice with a negative sign. By similar arguments, we have two instances in which $\kappa_l^{(4)}$ is such that the first term in (A.20) takes the form

$$\frac{1}{k_N^2 N^2} \sum_{l=1}^{L} \sum \sum_{i,j=1}^{k_N} \sum \sum_{n,m=1}^{N} \delta_l^2 \lambda_l \kappa_{l(0,|N(i)-N(j)|,|N(i)+1-n|,|N(i)+1-n|)}$$

which are canceled by $A_{3,2}$, again since that term enters twice with a negative sign. The remaining terms of the first term in (A.20) do not provide the dominant rate of convergence such that we skip the further analysis and consider next the second term in (A.20). By (A.1) we have

$$\mathbb{E}\left[\theta_{n,h_1}^2 \theta_{m,h_1}^2\right]$$
$$= \kappa_{h_1(0,0,|n-m|,|n-m|)} + \kappa_{h_1(0,0)}\ \kappa_{h_1(|n-m|,|n-m|)} + 2\kappa_{h_1(0,|n-m|)}\ \kappa_{h_1(0,|n-m|)}$$

such that we obtain for the second term of (A.20)

$$\begin{aligned}
&\tfrac{1}{k_N^2 N^2} \textstyle\sum_{l=1}^{L} \sum \sum_{i,j=1}^{k_N} \sum \sum_{n,m=1}^{N} \sum_{h_1=L+1}^{\infty} \delta_l^2 \mathbb{E}\left[\theta_{N(i)+1,l} \theta_{N(j)+1,l}\right] \mathbb{E}\left[\theta_{n,h_1}^2 \theta_{m,h_1}^2\right] \\
&= \tfrac{1}{k_N^2 N^2} \textstyle\sum_{l=1}^{L} \sum \sum_{i,j=1}^{k_N} \sum \sum_{n,m=1}^{N} \sum_{h_1=L+1}^{\infty} \delta_l^2 \lambda_{h_1}^2 \mathbb{E}\left[\theta_{N(i)+1,l} \theta_{N(j)+1,l}\right] \\
&+ \tfrac{1}{k_N^2 N^2} \textstyle\sum_{l=1}^{L} \sum \sum_{i,j=1}^{k_N} \sum \sum_{n,m=1}^{N} \sum_{h_1=L+1}^{\infty} \delta_l^2 \mathbb{E}\left[\theta_{N(i)+1,l} \theta_{N(j)+1,l}\right] \kappa_{h_1(0,0,|n-m|,|n-m|)} \\
&+ 2\tfrac{1}{k_N^2 N^2} \textstyle\sum_{l=1}^{L} \sum \sum_{i,j=1}^{k_N} \sum \sum_{n,m=1}^{N} \sum_{h_1=L+1}^{\infty} \delta_l^2 \mathbb{E}\left[\theta_{N(i)+1,l} \theta_{N(j)+1,l}\right] \kappa_{h_1(0,|n-m|)^2}.
\end{aligned}$$

Observe now that the first term in the above display is canceled by $A_{3,1}$ as it enters twice with a negative sign. As a consequence, the terms A_2, A_3 and parts of A_1 cancel each other out. The dominant rate of convergence is now obtained by considering the third term in the above display for which we have

$$
\begin{aligned}
&\tfrac{2}{k_N^2 N^2} \textstyle\sum_{l=1}^{L} \sum\sum_{i,j=1}^{k_N} \sum\sum_{n,m=1}^{N} \sum_{h_1=L+1}^{\infty} \delta_l^2 \mathbb{E}\left[\theta_{N(i)+1,l}\theta_{N(j)+1,l}\right] \kappa_{h_1}(0,|n-m|)^2 \\
&\quad = 2\left(\tfrac{1}{k_N N} \textstyle\sum_{l=1}^{L} \delta_l^2 \sum\sum_{i,j=1}^{k_N} \mathbb{E}\left[\theta_{N(i)+1,l}\theta_{N(j)+1,l}\right]\right) \times \\
&\qquad \left(\tfrac{1}{k_N N} \textstyle\sum_{h_1=L+1}^{\infty} \sum\sum_{n,m=1}^{N} \kappa_{h_1}(0,|n-m|)^2\right). \qquad \text{(A.21)}
\end{aligned}
$$

For the first term in brackets in (A.21) we have, for some constant $C > 0$,

$$
\begin{aligned}
(\ldots) \quad &\leq \tfrac{1}{k_N N} \textstyle\sum_{l=1}^{L} \delta_l^2 \sum_{i=1}^{k_N} \mathbb{E}\left[\theta_{N(i)+1,l}^2\right] + \tfrac{1}{k_N N} \sum_{l=1}^{L} \delta_l^2 \sum\sum_{i\neq j}^{k_N} \left|\mathbb{E}\left[\theta_{N(i)+1,l}\theta_{N(j)+1,l}\right]\right| \\
&\leq \tfrac{1}{k_N N} \textstyle\sum_{l=1}^{L} \delta_l^2 \sum_{i=1}^{k_N} \lambda_l + \tfrac{2}{k_N N} \sum_{m=1}^{k_N-1} \sum_{i=m+1}^{k_N} \sum_{l=1}^{L} \delta_l^2 B_{m,l} \\
&\leq \tfrac{1}{N} \textstyle\sum_{l=1}^{L} \delta_l^2 \lambda_l + \tfrac{C}{k_N N} \sum_{m=1}^{k_N-1} (k_N - m) m^{-\beta} \sum_{l=1}^{L} \delta_l^2 \lambda_l \\
&= O\left(\frac{k_N^{1-\widetilde{\beta}} L^{3+\alpha}}{N}\right),
\end{aligned}
$$

where the last equality follows from Assumption 1. For the second term in brackets in (A.21) we have by similar arguments for some constants $C, C^* > 0$,

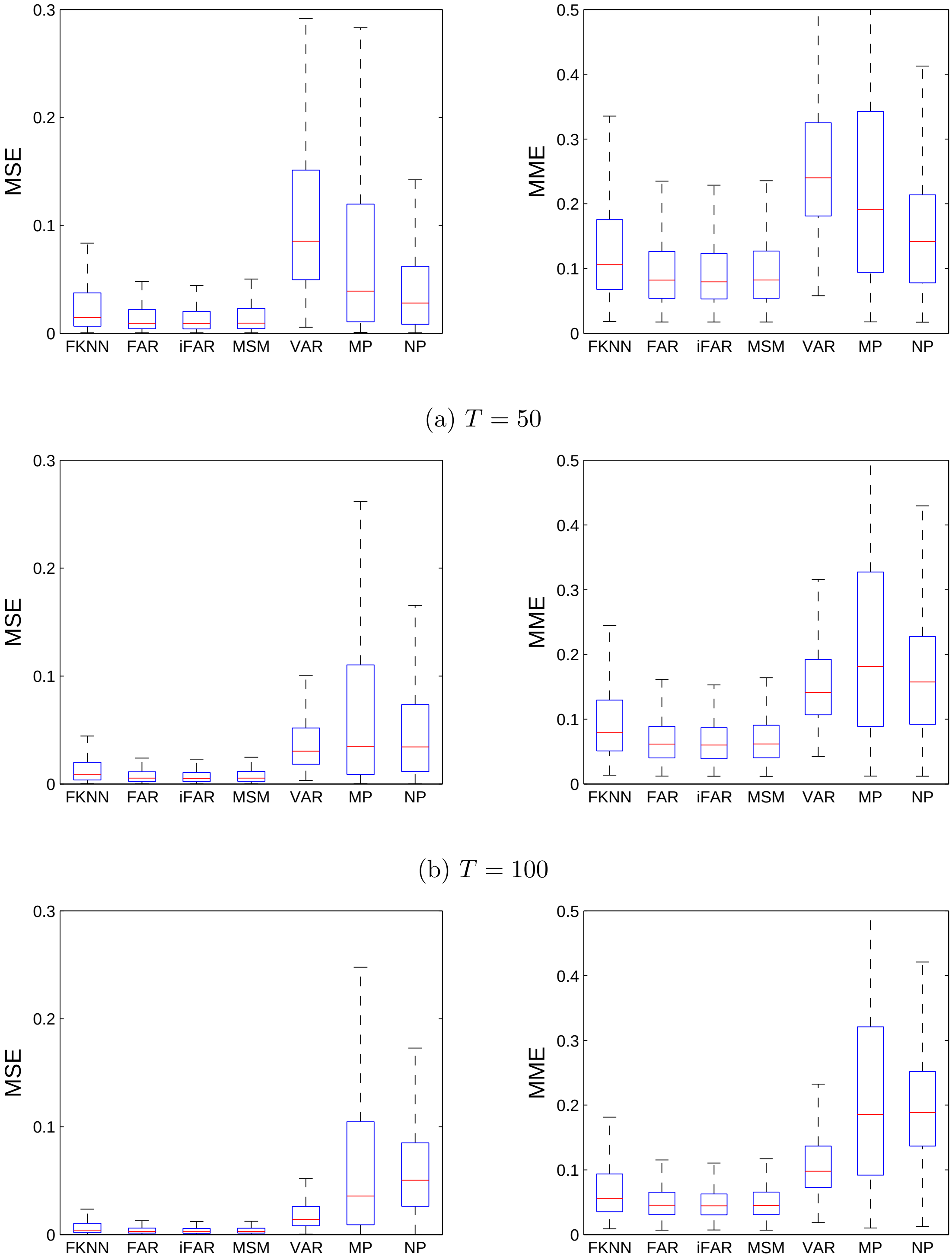
$$
\begin{aligned}
(\ldots) \quad &\leq \frac{1}{k_N N} \sum_{h_1=1}^{\infty} \sum\sum_{n,m=1}^{N} \mathbb{E}\left[\theta_{n,h_1}\theta_{m,h_1}\right]^2 \\
&\leq \frac{1}{k_N N} \sum_{h_1=1}^{\infty} \sum_{n=1}^{N} \mathbb{E}\left[\theta_{n,h_1}^2\right]^2 + \frac{1}{k_N N} \sum_{h_1=1}^{\infty} \sum\sum_{n\neq m}^{N} \left|\mathbb{E}\left[\theta_{n,h_1}\theta_{m,h_1}\right]\right|^2 \\
&\leq \frac{1}{k_N} \sum_{h_1=1}^{\infty} \lambda_{h_1}^2 + \frac{2}{k_N N} \sum_{m=1}^{N-1} \sum_{i=1}^{N} \sum_{h_1=1}^{\infty} B_{m,h_1}^2 \\
&\leq \frac{C}{k_N} + \frac{C^*}{k_N N} \sum_{m=1}^{N-1} \sum_{i=1}^{N} m^{-2\beta} \sum_{h_1=1}^{\infty} \lambda_{h_1}^2 = O\left(\frac{N^{1-2\beta^*}}{k_N}\right).
\end{aligned}
$$

where $\beta^* = \min\{\beta, 1/2\}$. Combining these results we obtain the following rate of convergence

$$
O\left(\frac{L^{3+\alpha}}{k_N^{\beta^*} N^{2\beta^{**}}}\right).
$$

Note that we omit the analysis of term $A_{1,1}$ for brevity as it follows by the same arguments presented above and yields the same rate of convergence.

C Appendix: Figures

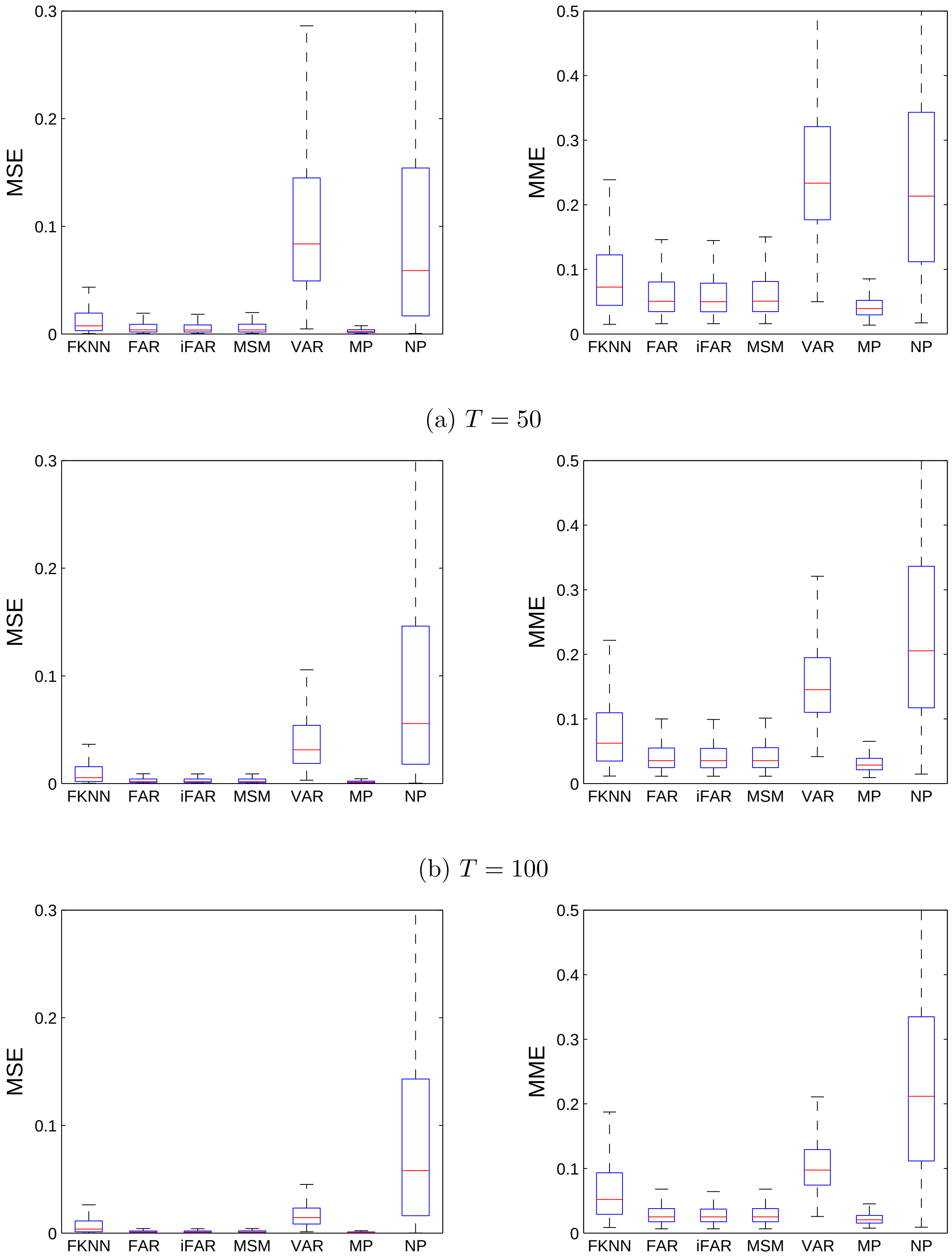


(a) $T = 50$

(b) $T = 100$

(c) $T = 200$

Figure 1. Boxplots of the prediction errors MSE (left panel) and MME (right panel) when DGP has kernel $\rho(t, s) = Ce^{\frac{-(t^2+s^2)}{2}}$.

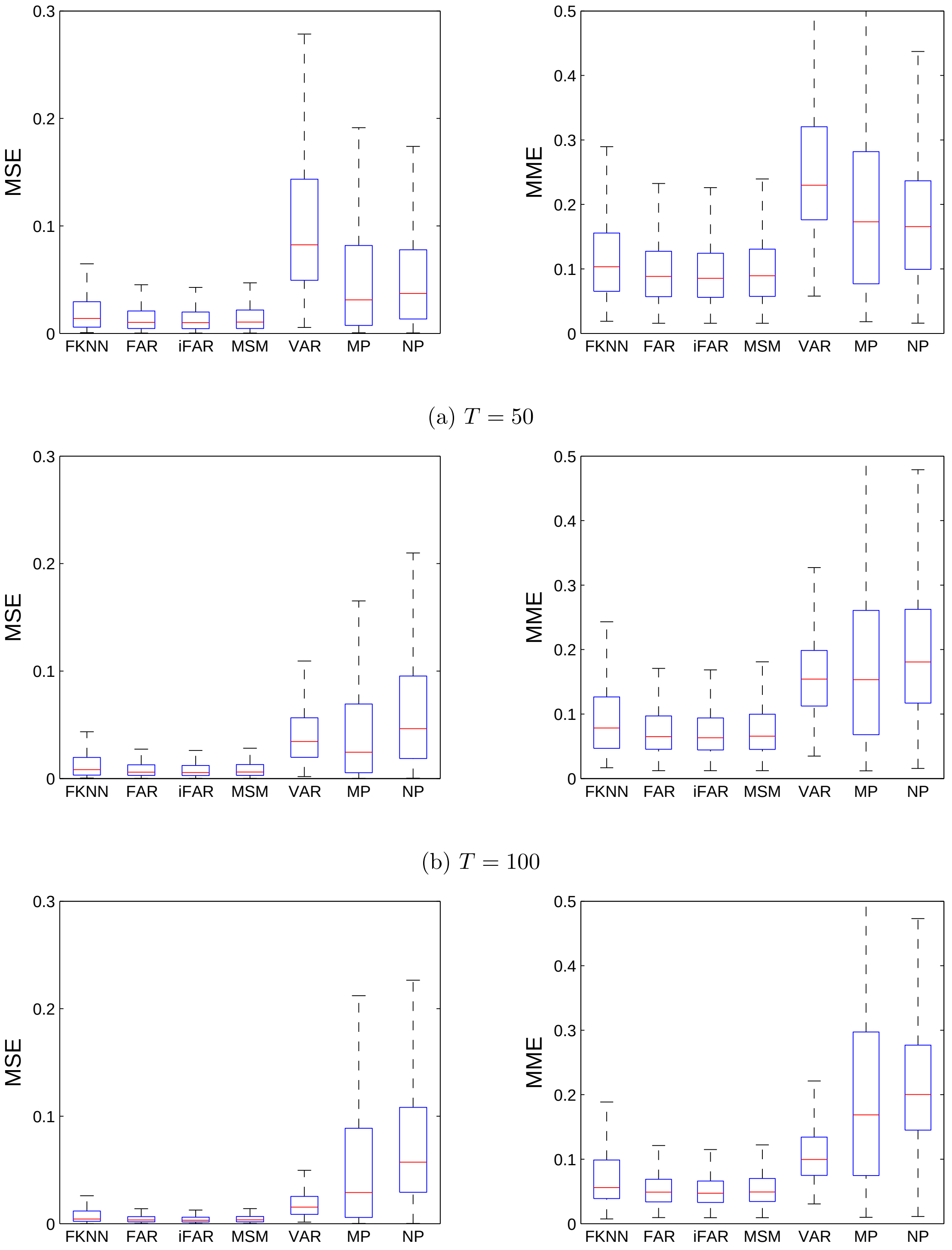


(a) $T = 50$

(b) $T = 100$

(c) $T = 200$

Figure 2. Boxplots of the prediction errors MSE (left panel) and MME (right panel) when DGP has kernel $\rho(t, s) = C$.



(a) $T = 50$

(b) $T = 100$

(c) $T = 200$

Figure 3. Boxplots of the prediction errors MSE (left panel) and MME (right panel) when DGP has kernel $\rho(t, s) = Ct$.

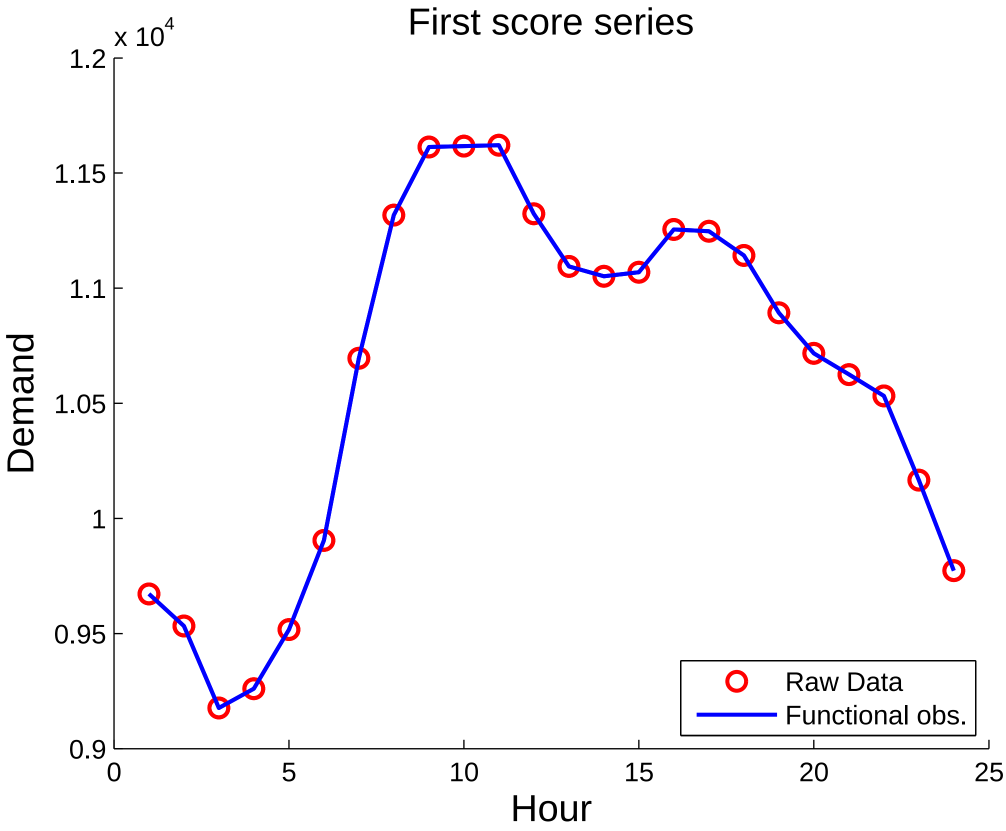
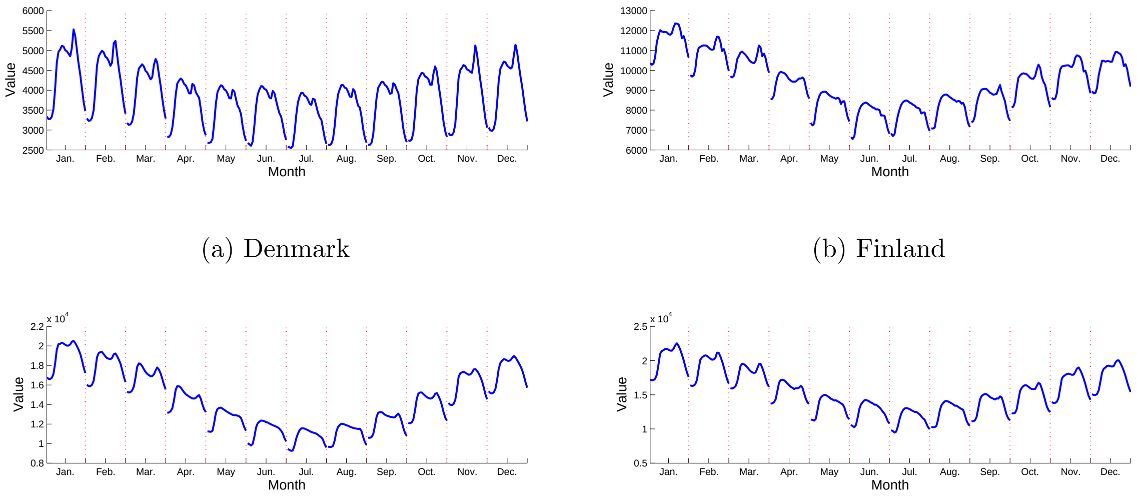


Figure 4. Typical daily discrete observation and reconstructed functional observation for electricity demand in Norway (June 1, 2013).



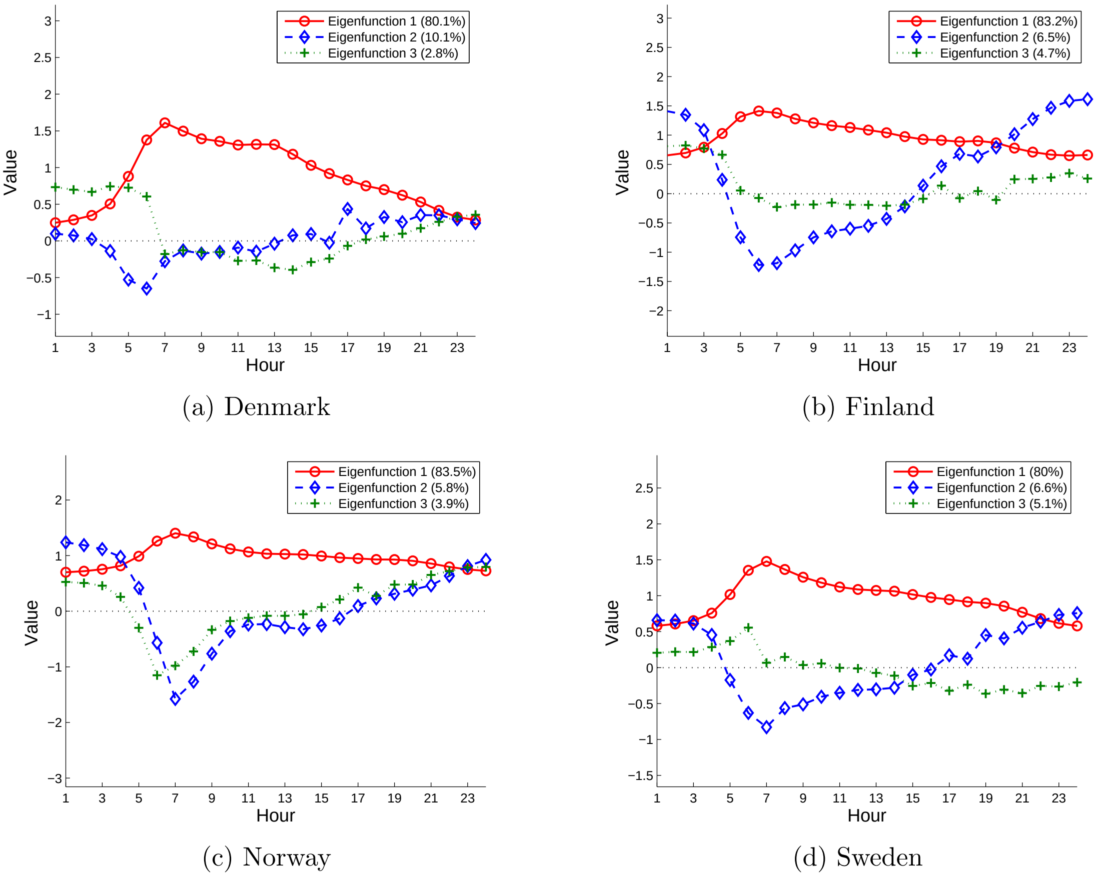
(a) Denmark

(b) Finland

(c) Norway

(d) Sweden

Figure 5. Seasonal monthly averages of the electricity demand in the Nordic countries.



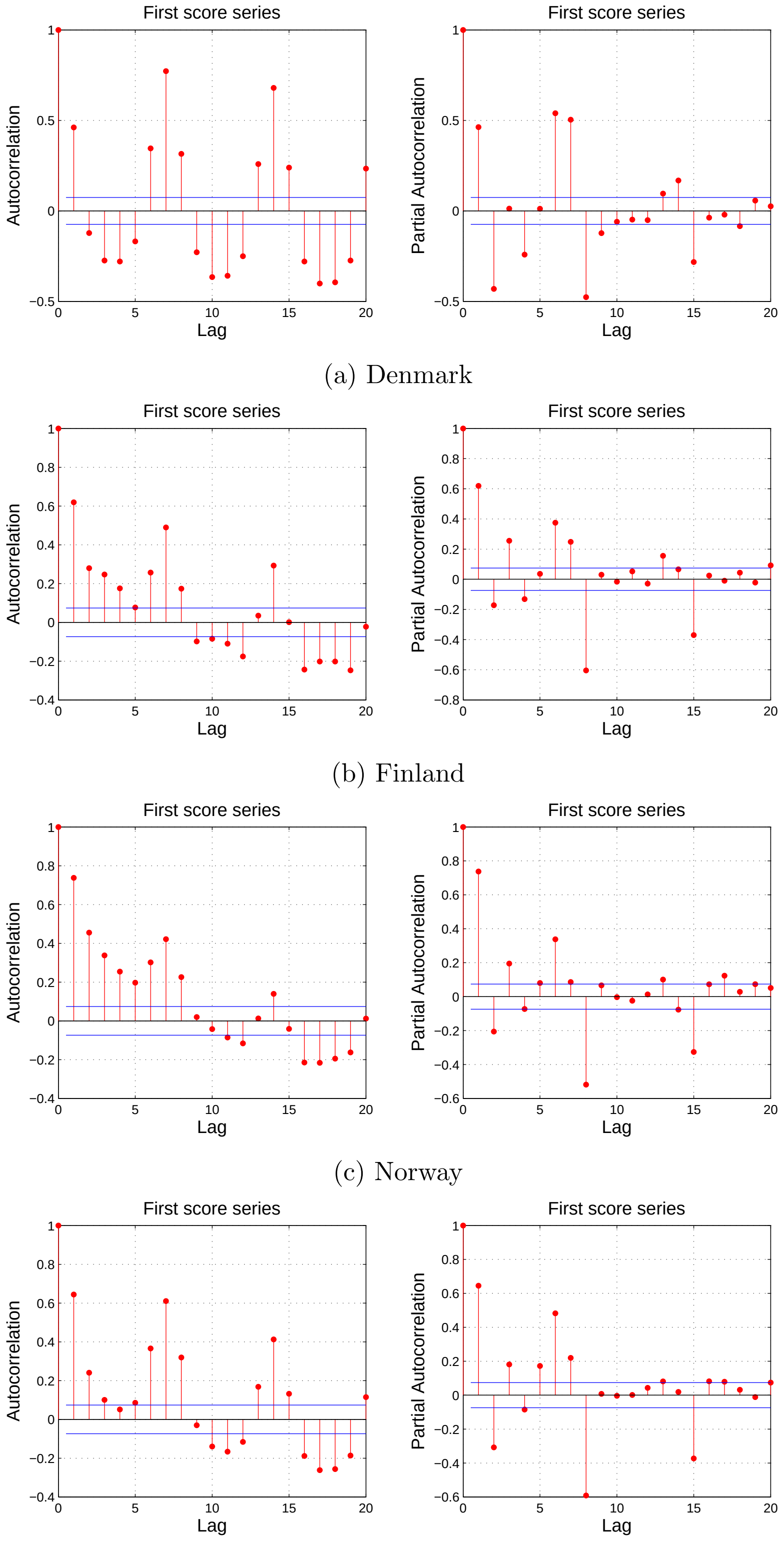
(a) Denmark

(b) Finland

(c) Norway

(d) Sweden

Figure 6. The first three estimated eigenfunctions of the electricity demand in the Nordic countries. The percentages indicate the amount of total variation accounted for by each eigenfunction.

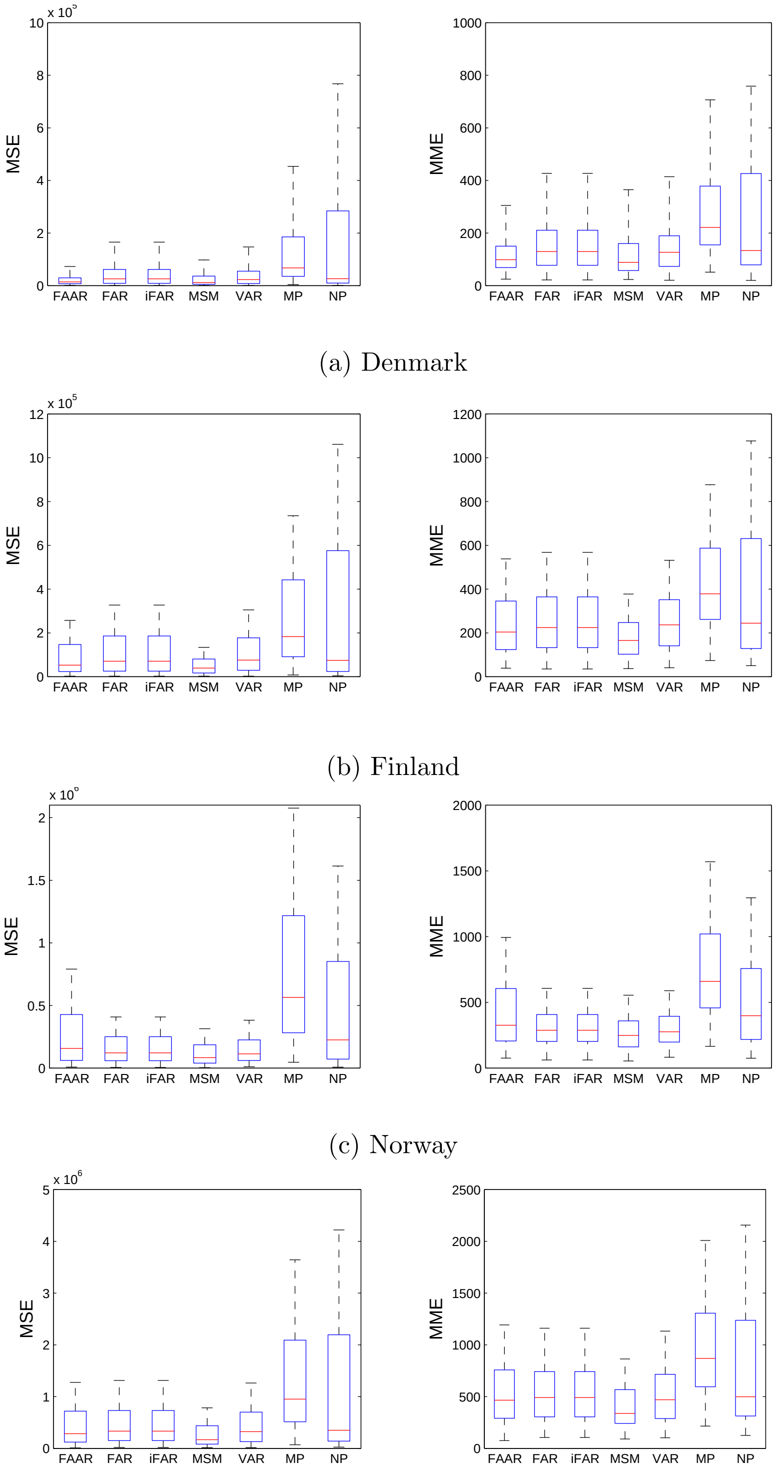

(a) Denmark

(b) Finland

(c) Norway

(d) Sweden

Figure 7. Time dependencies in score series. Left panel: sample autocorrelation of the first empirical FPC score series. Right panel: sample partial autocorrelation function of the first empirical FPC score series.



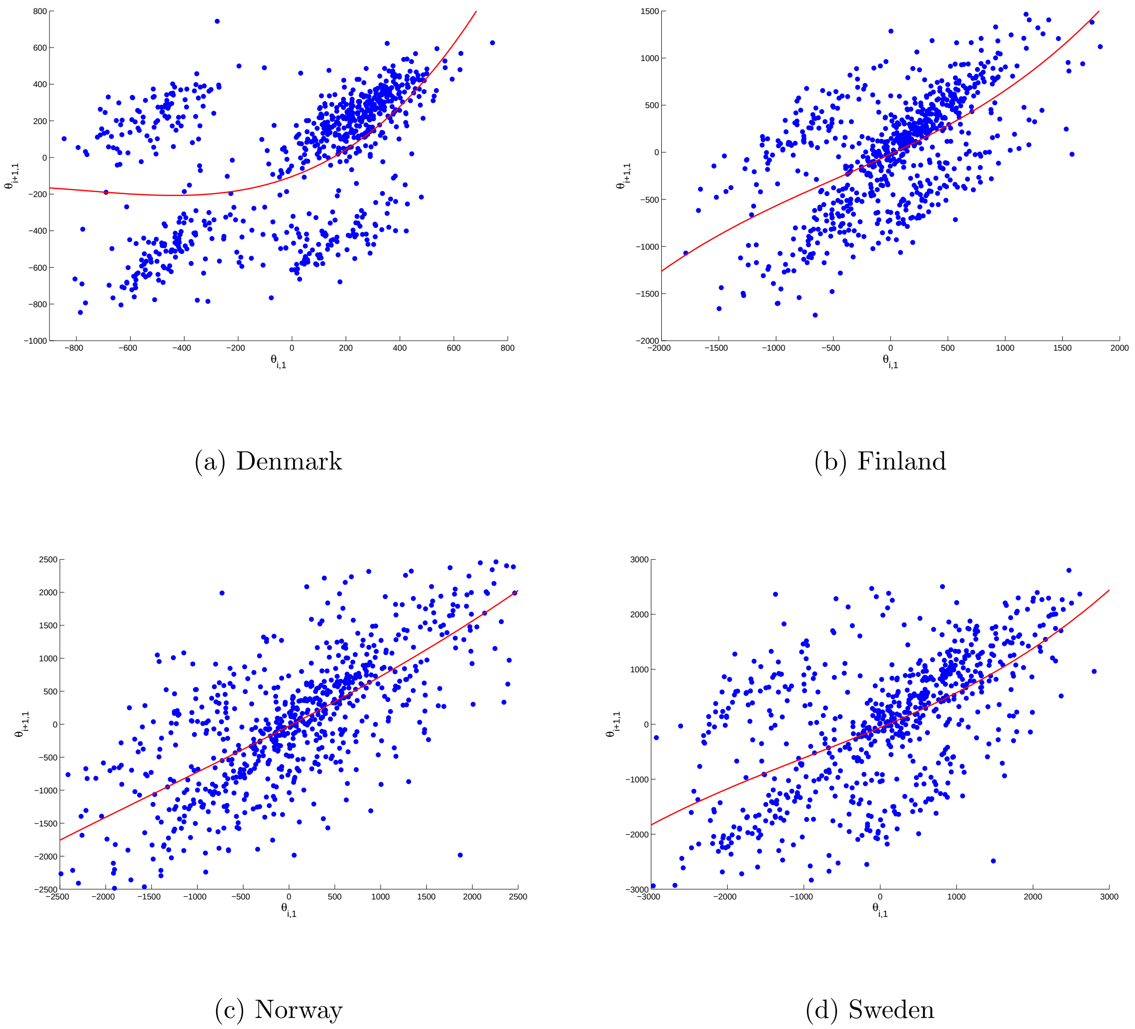
(a) Denmark

(b) Finland

(c) Norway

(d) Sweden

Figure 8. Boxplots of the prediction errors MSE (left panel) and MME (right panel).



(a) Denmark

(b) Finland

(c) Norway

(d) Sweden

Figure 9. Scatter plots of the relationship between for the first FPC score and it lag.

References

Andrews, D. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica 59*, 817–858.

Aue, A., D. D. Norinho, and S. Hörmann (2015). On the prediction of stationary functional time series. *Journal of the American Statistical Association 110*(509), 378–392.

Besse, P. C., H. Cardot, and D. B. Stephenson (2000). Autoregressive forecasting of some functional climatic variations. *Scandinavian Journal of Statistics 27*(4), 673–687.

Bosq, D. (2000). *Linear Processes in Function Spaces.* New York: Springer.

Brillinger, D. R. (2001). *Time Series: Data Analysis and Theory.* Philadelphia: Society for Industrial and Applied Mathematics.

Chen, B. J., M. W. Chang, and C. J. Lin (2004). Load forecasting using support vector machines: A study on eunite competition 2001. *IEEE Transactions on Power Systems 19*, 1821–1830.

Cover, T. and P. Hart (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory 13*, 21–27.

Dauxois, J., A. Pousse, and Y. Romain (1982). Asymptotic theory for the principal component analysis of a vector random function: Some applications to statistical inference. *Journal of Multivariate Analysis 12*(1), 136 – 154.

Davidson, J. (1994). *Stochastic Limit Theory.* Oxford: Oxford University Press.

Demetrescu, M., V. Kuzin, and U. Hassler (2008). Long memory testing in the time domain. *Econometric Theory 24*(1), 176–215.

Didericksen, D., P. Kokoszka, and X. Zhang (2012). Empirical properties of forecasts with the functional autoregressive model. *Computational Statistics 27*(2), 285–298.

Feinberg, E. A. and D. Genthliou (2005). Load forecasting. In J. H. Chow, F. F. Wu, and J. J. Momoh (Eds.), *Applied Mathematics for Restructured Electric Power Systems: Optimization, Control and Computational Intelligence, Power Electronics and Power Systems*, pp. 269–285. New York: Springer.

Gonçalves, S. and L. Kilian (2007). Asymptotic and bootstrap inference for AR(∞) processes with conditional heteroskedasticity. *Econometric Reviews 26*(6), 609–641.

Hall, P. and J. L. Horowitz (2007). Methodology and convergence rates for functional linear regression. *The Annals of Statistics 35*(1), 70–91.

Hamilton, J. D. (1994). *Time Series Analysis.* Princeton: Princeton University Press.

Hippert, H. S., C. E. Pedreira, and R. C. Souza (2001). Neural netowrks for short-term load forecasting: A review and evaluation. *IEEE Transactions on Power Systems 16*, 44–55.

Hörmann, S. and Ł. Kidziński (2015). A note on estimation in hilbertian linear models. *Scandinavian Journal of Statistics 42*(1), 43–62.

Hörmann, S. and P. Kokoszka (2010). Weakly dependend functional data. *The Annals of Statistics 38*, 1845–1884.

Horváth, L. and P. Kokoszka (2012). *Inference for Functional Data with Applications.* New York: Springer.

Horváth, L., P. Kokoszka, and G. Rice (2014). Testing stationarity of functional time series. *Journal of Econometrics 179*(1), 66 – 82.

Kokoszka, P. and M. Reimherr (2013). Determining the order of the functional autoregressive model. *Journal of Time Series Analysis 34*(1), 116–129.

Kokoszka, P. and X. Zhang (2010). Improved estimation of the kernel of the functional autoregressive process. *Technical Report. University of Chicago*.

Kyriakides, E. and M. Polycarpou (2007). Short term electric load forecasting: A tutorial. In C. K. and L. Wang (Eds.), *Trends in Neural Computation, Studies in Computational Intelligence, vol. 35*, pp. 391–418. New York: Springer.

Mas, A. (2007). Weak convergence in the functional autoregressive model. *Journal of Multivariate Analysis 98*(6), 1231 – 1261.

Müller, H.-G. and F. Yao (2008). Functional additive models. *Journal of the American Statistical Association 103*, 1534–1544.

Park, J. Y. and J. Qian (2012). Functional regression of continuous state distributions. *Journal of Econometrics 167*(2), 397 – 412. Fourth Symposium on Econometric Theory and Applications (SETA).

Ramsay, J., G. Hooker, and S. Graves (2009). *Functional Data Analysis with R and MATLAB*. Springer.

Ramsay, J. O. and B. W. Silverman (2005). *Functional Data Analysis* (2nd ed.). New York: Springer.

Stone, C. J. (1977). Consistent nonparametric regression. *The Annals of Statistics 5*, 595–620.

Stute, W. (1984). Asymptotic normality of nearest neighbor regression function estimates. *The Annals of Statistics 12*(3), 917–926.

Yakowitz, S. (1987). Nearest-neighbour methods for time series analysis. *Journal of Time Series Analysis 8*, 235–247.