Estimating functional time series by moving average model fitting*

Alexander Aue† Johannes Klepsch‡§

January 6, 2017

Abstract

Functional time series have become an integral part of both functional data and time series analysis. Important contributions to methodology, theory and application for the prediction of future trajectories and the estimation of functional time series parameters have been made in the recent past. This paper continues this line of research by proposing a first principled approach to estimate invertible functional time series by fitting functional moving average processes. The idea is to estimate the coefficient operators in a functional linear filter. To do this a functional Innovations Algorithm is utilized as a starting point to estimate the corresponding moving average operators via suitable projections into principal directions. In order to establish consistency of the proposed estimators, asymptotic theory is developed for increasing subspaces of these principal directions. For practical purposes, several strategies to select the number of principal directions to include in the estimation procedure as well as the choice of order of the functional moving average process are discussed. Their empirical performance is evaluated through simulations and an application to vehicle traffic data.

Keywords: Dimension reduction; Estimation, Functional data analysis; Functional linear process; Functional time series, Hilbert spaces; Innovations Algorithm, Moving average process

MSC 2010: Primary: 62M10, 62M15, 62M20; Secondary: 62H25, 60G25

1 Introduction

With the advent of complex data came the need for methods to address novel statistical challenges. Among the new methodologies, functional data analysis provides a particular set of tools for tackling questions related to observations conveniently viewed as entire curves rather than individual data points. The current state of the field may be reviewed in one of the comprehensive monographs written by Bosq [4], Ramsay and Silverman [23], Horváth and Kokoszka [11], and Hsing and Eubank [12]. Many of the applications discussed there

*This research was partially supported by NSF grants DMS 1305858 and DMS 1407530

†Department of Statistics, University of California, Davis, CA 95616, USA, email: aaue@ucdavis.edu

‡Center for Mathematical Sciences, Technische Universität München, 85748 Garching, Boltzmannstraße 3, Germany, email: j.klepsch@tum.de

§Corresponding author

point to an intrinsic time series nature of the underlying curves. This has led to an upsurge in contributions to the functional time series literature. The many recent works in this area include papers on time-domain methods such as Hörmann and Kokoszka [10], who introduced a framework to describe weakly stationary functional time series, and Aue et al. [3] and Klepsch and Klüppelberg [13], who developed functional prediction methodology; as well as frequency domain methods such as Panaretos and Tavakoli [22], who utilized functional cumulants to justify their functional Fourier analysis, Hörmann et al. [9], who defined the concept of dynamic functional principal components, and Aue and van Delft [1], who designed stationarity tests based on functional periodogram properties.

This paper is concerned with functional moving average (FMA) processes as a building block to estimate potentially more complicated functional time series. Together with the functional autoregressive (FAR) processes, the FMA processes comprise one of the basic functional time series model classes. They are used, for example, as a building block in the L^p-m-approximability concept of Hörmann and Kokoszka [10], which is based on the idea that a sufficiently close approximation with truncated linear processes may adequately capture more complex dynamics, based on a causal infinite MA representation. It should be noted that, while there is a significant number of papers on the use of both FMA and FAR processes, the same is not the case for the more flexible functional autoregressive moving average (FARMA) processes. This is due to the technical difficulties that arise from transitioning from the multivariate to the functional level. One advantage that FMA processes enjoy over other members of the FARMA class is that their projections remain multivariate MA processes (of potentially lower order). This is one of the reasons that makes them attractive for further study.

Here interest is in estimating the dynamics of an invertible functional linear process through fitting FMA models. The operators in the FMA representation, a functional linear filter, are estimated using a functional Innovations Algorithm. This counterpart of the well-known univariate and multivariate Innovations Algorithms was recently introduced by Klepsch and Klüppelberg [13], where its properties were analyzed on a population level. These results are extended to the sample case and used as a first step in the estimation. The proposed procedure uses projections to a number of principal directions, estimated through functional principal components analysis (see, for example, Ramsay and Silverman [23]). To ensure appropriate large-sample properties of the proposed estimators, the dimensionality of the principle directions space is allowed to grow slowly with the sample size. In this framework, the consistency of the estimators of the functional linear filter is the main theoretical contribution. It is presented in Section 3.

The theoretical results are accompanied by selection procedures to guide the selection of the order of the approximating FMA process and the dimension of the subspace of principal directions. To choose the dimension of the subspace a sequential test procedure is proposed. Order selection based on AICC, Box–Ljung and FPE type criteria are suggested. Details of the proposed model selection procedures are given in

Section 4. Their practical performance is highlighted in Section 5, where results of a simulation study are reported, and Section 6, where an application to real-world data on vehicle traffic data is discussed.

To summarize, this paper is organized as follows. Section 2 briefly reviews basic notions of Hilbert-space valued random variables before introducing the setting and the main assumptions. The proposed estimation methodology for functional time series is detailed in Section 3. Section 4 discusses in some depth the practical selection of the dimension of the projection space and the order of the approximating FMA process. These suggestions are tested in a Monte Carlo simulation study and an application to traffic data in Sections 5 and 6, respectively. Section 7 concludes and proofs of the main results can be found in Section 8.

2 Setting

Functional data is often conducted in $H = L^2[0,1]$, the Hilbert-space of square-integrable functions, with canonical norm $\|x\| = \langle x, x\rangle^{1/2}$ induced by the inner product $\langle x, y\rangle = \int_0^1 x(s)y(s)ds$ for $x, y \in H$. For an introduction to Hilbert spaces from a functional analytic perspective, the reader is referred to Chapters 3.2 and 3.6 in Simon [24]. All random functions considered in this paper are defined on a probability space $(\Omega, \mathcal{A}, \mathcal{P})$ and are assumed to be $\mathcal{A}$-$\mathcal{B}_H$-measurable, where $\mathcal{B}_H$ denotes the Borel σ-algebra of subsets of H. Note that the space of square integrable random functions $L_H^2 = L^2(\Omega, \mathcal{A}, \mathcal{P})$ is a Hilbert space with inner product $\mathrm{E}[\langle X, Y\rangle] = \mathrm{E}[\int_0^1 X(s)Y(s)ds]$ for $X, Y \in L_H^2$. Similary, denote by $L_H^p = L^p(\Omega, \mathcal{A}, \mathcal{P})$ the space of H-valued functions such that $\nu_p(X) = (\mathrm{E}[\|X\|^p])^{1/p} < \infty$. Let $\mathbb{Z}$, $\mathbb{N}$ and $\mathbb{N}_0$ denote the set of integers, positive integers and non-negative integers, respectively.

Interest in this paper is in fitting techniques for functional time series $(X_j \colon j \in \mathbb{Z})$ taking values in L_H^2. To describe a wide variety of temporal dynamics, the framework is established for functional linear processes $(X_j \colon j \in \mathbb{Z})$ defined through the series expansion

$$X_j = \sum_{\ell=0}^{\infty} \psi_\ell \varepsilon_{j-\ell}, \qquad j \in \mathbb{Z}, \tag{2.1}$$

where $(\psi_\ell \colon \ell \in \mathbb{N}_0)$ is a sequence in $\mathcal{L}$, the space of bounded linear operators acting on H, equipped with the standard norm $\|A\|_{\mathcal{L}} = \sup_{\|x\|\leq 1} \|Ax\|$, and $(\varepsilon_j \colon j \in \mathbb{Z})$ is assumed to be an independent and identically distributed sequence in L_H^2. Additional summability conditions are imposed on the sequence of coefficient operators $(\psi_\ell \colon \ell \in \mathbb{N}_0)$ if it is necessary to control the rate of decay of the temporal dependence. Whenever the terminology "functional linear process" is used in this paper it is understood to be in the sense of (2.1). Note that, as for univariate and multivariate time series models, every stationary causal functional autoregressive moving average (FARMA) process is a functional linear process (see Spangenberg [25], Theorem 2.3). Special cases include functional autoregressive processes of order p, FAR(p), which have been thoroughly investigated

in the literature, and the *functional moving average process of order* q, FMA(q), which is given by the equation

$$X_j = \sum_{\ell=1}^{q} \theta_\ell \varepsilon_{j-\ell} + \varepsilon_j, \qquad j \in \mathbb{Z}, \tag{2.2}$$

with $\theta_1, \ldots, \theta_q \in \mathcal{L}$.

While the functional linear process in (2.1) is the prototypical causal time series, in the context of prediction, the concept of invertibility naturally enters; see Chapter 5.5 of Brockwell and Davis [6], and Nsiri and Roy [21]. For a functional time series $(X_j \colon j \in \mathbb{Z})$ to be *invertible*, it is required that

$$X_j = \sum_{\ell=1}^{\infty} \pi_\ell X_{j-\ell} + \varepsilon_j, \qquad j \in \mathbb{Z}, \tag{2.3}$$

for $(\pi_\ell \colon \ell \in \mathbb{N})$ in $\mathcal{L}$ such that $\sum_{\ell=1}^{\infty} \|\pi_\ell\|_{\mathcal{L}} < \infty$; see Merlevède [18]. A sufficient condition for invertibility of a functional linear process, which is assumed throughout, is given in Theorem 7.2 of Bosq [4].

The definition of a functional linear process in (2.1) provides a convenient framework for the formulation of large-sample results and their verification. In order to analyze time series characteristics in practice, however, most statistical methods require a more in-depth understanding of the underlying dependence structure. This is typically achieved through the use of autocovariances which determine the second-order structure. Observe first that any random variable in L_H^p with $p \geq 1$ possesses a unique *mean function* in H, which allows for a pointwise definition; see Bosq [4]. For what follows, it is assumed without loss of generality that $\mu = 0$, the zero function. If $X \in L_H^p$ with $p \geq 2$ such that $\mathrm{E}[X] = 0$, then the *covariance operator* of X exists and is given by

$$C_X(y) = \mathrm{E}[\langle X, y \rangle X], \qquad y \in H.$$

If $X, Y \in L_H^p$ with $p \geq 2$ such that $\mathrm{E}[X] = \mathrm{E}[Y] = 0$, then the *cross covariance operator* of X and Y exists and is given by

$$C_{X,Y}(y) = C_{Y,X}^*(y) = \mathrm{E}[\langle X, y \rangle Y], \qquad y \in H.$$

where $C_{Y,X}^*$ denotes the adjoint of $C_{Y,X}$, noting that the adjoint A^* of an operator A is defined by the equality $\langle Ax, y \rangle = \langle x, A^* y \rangle$ for $x, y \in H$. The operators C_X and $C_{Y,X}$ belong to $\mathcal{N}$, the class of *nuclear operators*, whose elements A have a representation $A = \sum_{j=1}^{\infty} \lambda_j \langle e_j, \cdot \rangle f_j$ with $\sum_{j=1}^{\infty} |\lambda_j| < \infty$ for two orthonormal bases (ONB) $(e_j)_{j \in \mathbb{N}}$ and $(f_j)_{j \in \mathbb{N}}$ of H. In that case $\|A\|_{\mathcal{N}} = \sum_{j=1}^{\infty} |\lambda_j| < \infty$; see Section 1.5 of Bosq [4]. Furthermore, C_X is self-adjoint ($C_X = C_X^*$) and non-negative definite with spectral representation

$$C_X(y) = \sum_{i=1}^{\infty} \lambda_i \langle y, \nu_i \rangle \nu_i, \qquad y \in H,$$

where $(\nu_i \colon i \in \mathbb{N})$ is an ONB of H and $(\lambda_i \colon i \in \mathbb{N})$ is a sequence of positive real numbers such that $\sum_{i=1}^{\infty} \lambda_i < \infty$. When considering spectral representations, it is standard to assume that the $(\lambda_i \colon i \in \mathbb{N})$ are ordered decreasingly and that there are no ties between consecutive λ_i.

For ease of notation, introduce the operator $x \otimes y(\cdot) = \langle x, \cdot \rangle y$ for $x, y \in H$. Then, $C_X = \mathrm{E}[X \otimes X]$ and $C_{X,Y} = \mathrm{E}[X \otimes Y]$. Moreover, for a stationary process $(X_j \colon j \in \mathbb{Z})$, the *lag-$h$ covariance operator* can be written as

$$C_{X;h} = \mathrm{E}[X_0 \otimes X_h], \qquad h \in \mathbb{Z}. \tag{2.4}$$

The quantities in (2.4) are the basic building block in the functional Innovations Algorithm and the associated estimation strategy to be discussed in the next section.

3 Estimation methodology

3.1 Linear prediction in function spaces

Briefly recall the concept of linear prediction in Hilbert spaces as defined in Section 1.6 of Bosq [4]. Let $(X_j \colon j \in \mathbb{Z})$ be an invertible, functional linear process. Let $\bar{L}_{n,k}$ be the $\mathcal{L}$-closed subspace (LCS) generated by the stretch of functions $X_{n-k}, \ldots, X_n$. LCS here is to be understood in the sense of Fortet [7] that is $\bar{L}_{n,k}$ is the smallest subspace of H containing $X_{n-k}, \ldots, X_n$, closed with respect to operators in $\mathcal{L}$. Then, the best linear predictor of X_{n+1} given $\{X_n, X_{n-1}, \ldots, X_{n-k}\}$ at the population level is given by

$$\tilde{X}^f_{n+1,k} = P_{\bar{L}_{n,k}}(X_{n+1}), \tag{3.1}$$

where the superscript f in the predictor notation indicates the fully functional nature of the predictor and $P_{\bar{L}_{n,k}}$ denotes projection on $\bar{L}_{n,k}$. Note that there are major differences to the multivariate prediction case. Due to the infinite dimensionality of function spaces, $\tilde{X}^f_{n+1,k}$ in (3.1) is not guaranteed to have a representation in terms of its past values and operators in $\mathcal{L}$, see for instance Proposition 2.2 in Bosq [5] and the discussion in Section 3 of Klepsch and Klüppelberg [13]. A typical remedy in FDA is to resort to projections into principal directions and then to let the dimension d of the projection subspace grow to infinity. At the subspace-level, multivariate methods may be applied to compute the predictors; for example the multivariate Innovations Algorithm; see Lewis and Reinsel [17] and Mitchell and Brockwell [20]. This, however, has to be done with care, especially if sample versions of the predictors in (3.1) are considered. Even at the population level, the rate at which d tends to infinity has to be calibrated scrupulously to ensure that the inversions of matrices occurring, for example, in the multivariate Innovations Algorithm are meaningful and well defined (see Theorem 5.3 of Klepsch and Klüppelberg [13]).

Therefore, the following alternative to the functional best linear predictor defined in (3.1) is proposed. Recall that $(\nu_j \colon j \in \mathbb{N})$ are the eigenfunctions of the covariance operator C_X. Let $\mathcal{V}_d = \overline{\mathrm{sp}}\{\nu_1, \ldots, \nu_d\}$ be

the subspace generated by the first d principal directions and let $P_{\mathcal{V}_d}$ be the projection operator projecting from H onto $\mathcal{V}_d$. Let furthermore $(d_i \colon i \in \mathbb{N})$ be an increasing sequence of positive integers and define

$$X_{d_i,j} = P_{\mathcal{V}_{d_i}} X_j, \qquad j \in \mathbb{Z},\ i \in \mathbb{N}. \tag{3.2}$$

Note that (3.2) allows for the added flexibility of projecting different X_j into different subspaces $\mathcal{V}_i$. Then, X_{n+1} can be projected into the LCS generated by $X_{d_k,n}, X_{d_{k-1},n-1}, \ldots, X_{d_1,n-k}$, which is denoted by $\bar{F}_{n,k}$. Consequently, write

$$\tilde{X}_{n+1,k} = P_{\bar{F}_{n,k}}(X_{n+1}) \tag{3.3}$$

for the best linear predictor of X_{n+1} given $\bar{F}_{n,k}$. This predictor could be computed by regressing X_{n+1} onto $X_{d_k,n}, X_{d_{k-1},n-1}, \ldots, X_{d_1,n-k}$, but interest is here in the equivalent representation of $\tilde{X}_{n+1,k}$ in terms of one-step ahead prediction residuals given by

$$\tilde{X}_{n+1,k} = \sum_{i=1}^{k} \theta_{k,i}(X_{d_{k+1-i},n+1-i} - \tilde{X}_{n+1-i,k-i}), \tag{3.4}$$

where $\tilde{X}_{n-k,0} = 0$. On a population level, it was shown in Klepsch and Klüppelberg [13] that the coefficients $\theta_{k,i}$ with $k, i \in \mathbb{N}$ can be computed with the following algorithm.

Algorithm 3.1 (**Functional Innovations Algorithm**). *Let $(X_j \colon j \in \mathbb{Z})$ be a stationary functional linear process with covariance operator C_X possessing eigenpairs $(\lambda_i, \nu_i \colon i \in \mathbb{N})$ with $\lambda_i > 0$ for all $i \in \mathbb{N}$. The best linear predictor $\tilde{X}_{n+1,k}$ of X_{n+1} based on $\bar{F}_{n,k}$ defined in* (3.4) *can be computed by the recursions*

$$\begin{aligned}
\tilde{X}_{n-k,0} &= 0 \qquad \textit{and} \qquad V_1 = P_{\mathcal{V}_{d_1}} C_X P_{\mathcal{V}_{d_1}}, \\
\tilde{X}_{n+1,k} &= \sum_{i=1}^{k} \theta_{k,i}(X_{d_{k+1-i},n+1-i} - \tilde{X}_{n+1-i,k-i}), \\
\theta_{k,k-i} &= \left(P_{\mathcal{V}_{d_{k+1}}} C_{X;k-i} P_{\mathcal{V}_{d_{i+1}}} - \sum_{j=0}^{i-1} \theta_{k,k-j} V_j \theta^*_{i,i-j} \right) V_i^{-1}, \qquad i = 1, \ldots, n-1, \qquad (3.5) \\
V_k &= C_{X_{d_{k+1}} - \tilde{X}_{n+1,k}} = C_{X_{d_{k+1}}} - \sum_{i=0}^{k-1} \theta_{k,k-i} V_i \theta^*_{k,k-i}. \qquad (3.6)
\end{aligned}$$

Note that $\theta_{k,k-i}$ and V_i are operators in $\mathcal{L}$ for all $i = 1, \ldots, k$.

The first main goal is now to show how a finite sample version of this algorithm can be used to estimate the operators in (2.2), as these FMA processes will be used to approximate the more complex processes appearing in Definition 8.1. Note that Hörmann and Kokoszka [10] give assumptions under which $\sqrt{n}$-consistent estimators can be obtained for the lag-h autocovariance operator $C_{X;h}$, for $h \in \mathbb{Z}$. However, in (3.5), estimators are required for the more complicated quantities $P_{\mathcal{V}_{d_{k+1}}} C_{X;k-i} P_{\mathcal{V}_{d_{i+1}}}$, for $k,\ i \in \mathbb{N}$. If, for $i \in \mathbb{N}$, the projection

subspace $\mathcal{V}_{d_i}$ is known, consistent estimators of $P_{\mathcal{V}_{d_{k+1}}} C_{X;k-i} P_{\mathcal{V}_{d_{i+1}}}$ can be obtained by estimating $C_{X;k-i}$ and projecting the operator on the desired subspace. This case will be dealt with in Section 3.2. In practice, however, the subspaces $\mathcal{V}_{d_i}$, $i \in \mathbb{N}$, need to be estimated. This is a further difficulty that will be addressed separately in an additional step as part of Section 3.3.

Now, introduce additional notation. For $k \in \mathbb{N}$, denote by $(X_j(k)\colon j \in \mathbb{Z})$ the functional process taking values in H^k such that

$$X_j(k) = (X_j, X_{j-1}, \ldots, X_{j-k+1})^\top,$$

where $^\top$ signifies transposition. Let

$$\Gamma_k = C_{X(k)} \qquad \text{and} \qquad \Gamma_{1,k} = C_{X_{n+1},X_n(k)} = \mathrm{E}\big[X_{n+1} \otimes X_n(k)\big].$$

Based on a realization $X_1, \ldots, X_n$ of $(X_j\colon j \in \mathbb{Z})$, estimators of the above operators are given by

$$\hat{\Gamma}_k = \frac{1}{N-k} \sum_{j=k}^{N-1} X_j(k) \otimes X_j(k) \qquad \text{and} \qquad \hat{\Gamma}_{1,k} = \frac{1}{N-k} \sum_{j=k}^{N-1} X_{j+1} \otimes X_j(k). \tag{3.7}$$

The following theorem establishes the $\sqrt{n}$-consistency of the estimator $\hat{\Gamma}_k$ of Γ_k defined in (3.7).

Theorem 3.1. *If $(X_j\colon j \in \mathbb{Z})$ is a functional linear process defined in* (2.1) *such that the coefficient operators $(\psi_\ell\colon \ell \in \mathbb{N}_0)$ satisfy the summability condition $\sum_{m=1}^{\infty} \sum_{\ell=m}^{\infty} \|\psi_\ell\|_{\mathcal{L}} < \infty$ and with independent, identically distributed innovations $(\varepsilon_j\colon j \in \mathbb{Z})$ such that $\mathrm{E}[\|\varepsilon_0\|^4] < \infty$, then*

$$(N-k)\,\mathrm{E}\big[\|\hat{\Gamma}_k - \Gamma_k\|_{\mathrm{N}}^2\big] \leq k\, U_X,$$

where U_X is a constant that does not depend on n.

The proof of Theorem 3.1 is given in Section 8. There, an explicit expression for the constant U_X is derived that depends on moments of the underlying functional linear process and on the rate of decay of the temporal dependence implied by the summability condition on the coefficient operators $(\psi_\ell\colon \ell \in \mathbb{N}_0)$.

3.2 Known projection subspaces

In this section, conditions are established that ensure consistency of estimators of a functional linear process under the assumption that the projection subspaces $\mathcal{V}_{d_i}$ are known in advance. In this case as well as in the unknown subspace case, the following the general strategy is pursued; see Mitchell and Brockwell [20]. Start by providing consistency results for the estimators regression estimators of $\beta_{k,1}, \ldots, \beta_{k,k}$ in the linear model formulation

$$\tilde{X}_{n+1,k} = \beta_{k,1} X_{d_k,n} + \beta_{k,2} X_{d_{k-1},n-1} + \cdots + \beta_{k,k} X_{d_1,n-k+1}$$

of (3.3). To obtain the consistency of the estimators $\theta_{k,1}, \ldots, \theta_{k,k}$ exploit then that regression operators and Innovations Algorithm coefficient operators are, for $k \in \mathbb{N}$, linked through the recursions

$$\theta_{k,i} = \sum_{j=1}^{i} \beta_{k,j} \theta_{k-j,i-j}, \qquad i = 1, \ldots, k. \tag{3.8}$$

Define furthermore $P_{(k)} = \operatorname{diag}(P_{\mathcal{V}_{d_k}}, \ldots, P_{\mathcal{V}_{d_1}})$, the operator from H^k to H^k whose ith diagonal entry is given by the projection operator onto $\mathcal{V}_{d_i}$. One verifies that $P_{(k)} X_n(k) = (X_{d_k,n}, X_{d_{k-1},n-1}, \ldots, X_{d_1,n-k})^\top$, $C_{P_{(k)}X(k)} = P_{(k)} \Gamma_k P_{(k)} = \Gamma_{k,d}$ and $C_{X,P_{(k)}X(k)} = P_{(k)} \Gamma_{1,k} = \Gamma_{1,k,d}$. With this notation, it can be shown that $B(k) = (\beta_{k,1}, \ldots, \beta_{k,k})$ satisfies the population Yule–Walker equations

$$B(k) = \Gamma_{1,k,d} \, \Gamma_{k,d}^{-1},$$

of which sample versions are needed. In the known subspace case, estimators of $\Gamma_{1,k,d}$ and $\Gamma_{k,d}$ are given by

$$\hat{\Gamma}_{k,d} = P_{(k)} \hat{\Gamma}_k P_{(k)} \qquad \text{and} \qquad \hat{\Gamma}_{1,k,d} = \hat{\Gamma}_{1,k} P_{(k)}, \tag{3.9}$$

where $\hat{\Gamma}_k$ and $\hat{\Gamma}_{1,k}$ are as in (3.7). With this notation, $B(k)$ is estimated by the sample Yule–Walker equations

$$\hat{B}(k) = \hat{\Gamma}_{1,k,d} \hat{\Gamma}_{k,d}^{-1}. \tag{3.10}$$

Furthermore, the operators $\theta_{k,i}$ in (3.4) are estimated by $\hat{\theta}_{k,i}$, resulting from Algorithm 3.1 applied to the estimated covariance operators with $\mathcal{V}_{d_i}$ known. In order to derive asymptotic properties of $\hat{\beta}_{k,i}$ and $\hat{\theta}_{k,i}$ as both k and n tend to infinity, the following assumptions are imposed. Let α_{d_k} denote the infimum of the eigenvalues of all spectral density operators of $(X_{d_k,j} : j \in \mathbb{Z})$.

Assumption 3.1. *As $n \to \infty$, let $k = k_n \to \infty$ and $d_k \to \infty$ such that*

(i) *$(X_j : j \in \mathbb{Z})$ is as in Theorem 3.1 and invertible.*

(ii) *$k^{1/2}(n-k)^{-1/2} \alpha_{d_k}^{-2} \to 0$ as $n \to \infty$.*

(iii) *$k^{1/2} \alpha_{d_k}^{-1} \left(\sum_{\ell > k} \|\pi_\ell\|_{\mathcal{L}} + \sum_{\ell=1}^{k} \|\pi_\ell\|_{\mathcal{L}} \sum_{i > d_{k+1-\ell}} \lambda_i \right) \to 0$ as $n \to \infty$.*

Invertibility imposed in part *(i)* of Assumption 3.1 is a standard requirement in the context of prediction and is also necessary for the univariate Innovations Algorithm to be consistent. Assumption *(ii)* describes the restrictions on the relationship between k, d_k and n. The corresponding multivariate assumption in Mitchell and Brockwell [20] is $k^3/n \to 0$ as $n \to \infty$. Assumption *(iii)* is already required in the population version of the functional Innovations Algorithm in Klepsch and Klüppelberg [13]. It ensures that the best linear predictor based on the last k observations converges to the conditional expectation for $k \to \infty$. The corresponding multivariate condition in Brockwell and Mitchell [20] is $k^{1/2} \sum_{\ell > k} \|\pi_\ell\| \to 0$ as $n \to \infty$, where $(\pi_\ell : \ell \in \mathbb{N})$ here denote the matrices in the invertible representation of a multivariate linear process.

The main result concerning the asymptotic behavior of the estimators $\hat{\beta}_{k,i}$ and $\hat{\theta}_{k,i}$ is given next.

Theorem 3.2. *Let $\mathcal{V}_{d_i}$ be known for all $i \in \mathbb{N}$ and let Assumption 3.1 be satisfied. Then, for all $x \in H$ and all $i \in \mathbb{N}$ as $n \to \infty$,*

(i) $\|(\hat{\beta}_{k,i} - \pi_i)(x)\| \xrightarrow{p} 0$,

(ii) $\|(\hat{\theta}_{k,i} - \psi_i)(x)\| \xrightarrow{p} 0$.

If the operators $(\psi_\ell \colon \ell \in \mathbb{N})$ and $(\pi_\ell \colon \ell \in \mathbb{N})$ in the respective causal and invertible representations are assumed Hilbert–Schmidt, then the convergence in (i) and (ii) is uniform.

The proof of Theorem 3.2 is given in Section 8. The theorem establishes the pointwise convergence of the estimators needed in order to get a sample proxy for the functional linear filter $(\pi_\ell \colon \ell \in \mathbb{N})$. This filter encodes the second-order dependence in the functional linear process and can therefore be used for estimating the underlying dynamics for the case of known projection subspaces.

3.3 Unknown projection subspaces

The goal of this section is to remove the assumption of known $\mathcal{V}_{d_i}$. Consequently, the standard estimators for the eigenfunctions $(\nu_i \colon i \in \mathbb{N})$ of the covariance operator C_X are used, obtained as the sample eigenfunctions $\hat{\nu}_j$ of $\hat{C}_X$. Therefore, for $i \in \mathbb{N}$, the estimators of $\mathcal{V}_{d_i}$ and $P_{\mathcal{V}_{d_i}}$ are

$$\hat{\mathcal{V}}_{d_i} = \overline{\mathrm{sp}}\{\hat{\nu}_1, \hat{\nu}_2, \ldots, \hat{\nu}_{d_i}\} \qquad \text{and} \qquad \hat{P}_{\mathcal{V}_{d_i}} = P_{\hat{\mathcal{V}}_{d_i}}. \tag{3.11}$$

For $i \in \mathbb{N}$, let $\hat{\nu}_i' = c_i \hat{\nu}_i$, where $c_i = \mathrm{sign}(\langle \hat{\nu}_i, \nu_i \rangle)$. Then, Theorem 3.1 in Hörmann and Kokoszka [10] implies the consistency of $\hat{\nu}_i'$ for $\hat{\nu}_i$, with the quality of approximation depending on the spectral gaps of the eigenvalues $(\lambda_i \colon i \in \mathbb{N})$ of C_X. With this result in mind, define

$$\hat{\hat{\Gamma}}_{k,d} = \hat{P}_{(k)} \hat{\Gamma}_k \hat{P}_{(k)} \qquad \text{and} \qquad \hat{\hat{\Gamma}}_{1,k,d} = \hat{\Gamma}_{1,k} \hat{P}_{(k)}. \tag{3.12}$$

Now, if the projection subspace $\mathcal{V}_{d_i}$ is not known, the operators appearing in (3.8) and can be estimated by solving the estimated Yule–Walker equations

$$\hat{\hat{B}}(k) = \hat{\hat{\Gamma}}_{1,k,d} \hat{\hat{\Gamma}}_{k,d}^{-1}. \tag{3.13}$$

The coefficient operators in Algorithm 3.1 obtained from estimated covariance operators and estimated projection space $\hat{P}_{\mathcal{V}_{d_i}}$ are denoted by $\hat{\hat{\theta}}_{k,i}$. In order to derive results concerning their asymptotic behavior, an additional assumption concerning the decay of the spectral gaps of C_X is needed. Let $\delta_1 = \lambda_1 - \lambda_2$ and $\delta_j = \min\{\lambda_{j-1} - \lambda_j, \lambda_j - \lambda_{j+1}\}$ for $j \geq 2$.

Assumption 3.2. *As $n \to \infty$, $k = k_n \to \infty$ and $d_k \to \infty$ such that*

(iv) $k^{3/2} \alpha_{d_k}^{-2} \, n^{-1} (\sum_{\ell=1}^{d_k} \delta_\ell^{-2})^{1/2} \to 0$.

This type of assumption dealing with the spectral gaps is typically encountered when dealing with the estimation of eigenelements of functional linear processes (see, for example, Bosq [4], Theorem 8.7). We are now ready to derive the asymptotic result of the estimators in the general case that A_{d_i} is not known.

Theorem 3.3. *Let Assumptions 3.1 and 3.2 be satisfied. Then, for all $x \in H$ and $i \in \mathbb{N}$ as $n \to \infty$,*

(i) $\|(\hat{\hat{\beta}}_{k,i} - \pi_i)(x)\| \xrightarrow{p} 0$,

(ii) $\|(\hat{\hat{\theta}}_{k,i} - \psi_i)(x)\| \xrightarrow{p} 0$.

If the operators $(\psi_\ell \colon \ell \in \mathbb{N})$ and $(\pi_\ell \colon \ell \in \mathbb{N})$ are Hilbert–Schmidt, then the convergence is uniform.

The proof of Theorem 3.3 is given in Section 8. The theoretical results quantify the large-sample behavior of the estimates of the linear filter operators in the causal and invertible representations of the strictly stationary functional time series $(X_j \colon j \in \mathbb{Z})$. How to guide the application of the proposed method in finite samples is addressed in the next section.

4 Selection of principal directions and FMA order

Model selection is a difficult problem when working with functional time series. Contributions to the literature have been made in the context of functional autoregressive models by Kokoszka and Reimherr [15], who devised a sequential test to decide on the FAR order, and Aue et al. [3], who introduced an FPE-type criterion. To the best of our knowledge, there are no contributions in the context of model selection in functional moving average models. This section introduces several procedures. A method for the selection of the subspace dimension is introduced in Section 4.1, followed by a method for the FMA order selection in Section 4.2. A criterion for the simultaneous selection is in Section 4.3.

4.1 Selection of principal directions

The most well-known method for the selection of d in functional data analysis is based on total variance explained, TVE, where d is chosen such that the first d eigenfunctions of the covariance operator explain a predetermined amount P of the variability; see, for example, Horváth and Kokoszka [11]. In order to apply the TVE criterion in the functional time series context, one has to ensure that no essential parts of the dependence structure in the data are omitted after the projection into principal directions. This is achieved as follows. First choose an initial d^* with the TVE criterion such with a fraction P of variation in the data is explained. This should be done conservatively. Then apply the portmanteau test of Gabrys and Kokoszka [8] to check whether the non-projected part $(I_H - P_{\mathcal{V}_{d^*}})X_1, \ldots, (I_H - P_{\mathcal{V}_{d^*}})X_n$ of the observed functions $X_1, \ldots, X_n$ can be considered independent. Modifying their test to the current situation, yields the statistic

$$Q_n^{d^*} = n \sum_{h=1}^{\bar{h}} \sum_{\ell,\ell'=d^*+1}^{d^*+p} f_h(\ell, \ell') b_h(\ell, \ell'), \tag{4.1}$$

where $f_h(\ell, \ell')$ and $b_h(\ell, \ell')$ denote the (ℓ, ℓ')th entries of $C^{-1}_{\mathbf{X}^*;0} C_{\mathbf{X}^*;h}$ and $C_{\mathbf{X}^*;h} C^{-1}_{\mathbf{X}^*;0}$, respectively, and $(\mathbf{X}^*_j \colon j \in \mathbb{Z})$ is the p-dimensional vector process consisting of the $d+1$st to $d+p$th eigendirections of the covariance operator C_X. Following Gabrys and Kokoszka [8], it follows under the assumption of independence of the non-projected series that $Q_N^{d^*} \to \chi^2_{p^2\bar{h}}$ in distribution. If the assumption of independence is rejected, set $d^* = d^* + 1$. Repeat the test until the independence hypothesis cannot be rejected and choose $d = d^*$ to estimate the functional linear filters. This leads to the following algorithm.

Algorithm 4.1 (**Test for independence**). *Perform the following steps.*

(1) For given observed functional time series data $X_1, \ldots, X_n$, estimate the eigenpairs $(\hat{\lambda}_1, \hat{\nu}_1), \ldots, (\hat{\lambda}_n, \hat{\nu}_n)$ of the covariance operator C_X. Select d^* such that

$$\mathrm{TVE}(d^*) = \frac{\sum_{i=1}^{d^*} \hat{\lambda}_i}{\sum_{i=1}^{n} \hat{\lambda}_i} \geq P$$

for some prespecified $P \in (0, 1)$.

(2) While $Q_n^{d^*} > q_{\chi^2_{p^2\bar{h}}, \alpha}$, set $d^* = d^* + 1$.

(3) If $Q_n^{d^*} \leq q_{\chi^2_{p^2\bar{h}}, \alpha}$ stop and apply Algorithm 3.1 with $d_i = d^*$, for all $i \leq k$.

Note that the Algorithm 4.1 does not specify the choices of P, p, H and α. Recommendations on their selection are given in Section 5. Multiple testing could potentially be an issue, but intensive simulation studies have shown that, since d^* is initialized with the TVE criterion, usually no more than one or two iterations and tests are required for practical purposes. Therefore the confidence level is not adjusted, even though it would be feasible to incorporate this additional step into the algorithm.

4.2 Selection of FMA order

For a fixed d, multivariate model selection procedures can be applied to choose q. In fact, it is shown in Theorem 4.7 of Klepsch and Klüppelberg [13] that the projection of an FMA(q) process on a finite-dimensional space is a VMA(q^*) with $q^* \leq q$. Assuming that the finite-dimensional space is chosen such that no information on the dependence structure of the process is lost, $q = q^*$. Then, the FMA order q may be chosen by performing model selection on the d-dimensional vector model given by the first d principal directions of $(X_j \colon j \in \mathbb{Z})$. Methods for selecting the order of VMA models are described, for example, in Chapter 11.5 of Brockwell and Davis [6], and Chapter 3.2 of Tsai [26].

The latter book provides arguments for the identification of the VMA order via cross correlation matrices. This Ljung–Box (LB) method for testing the null hypothesis $H_0 \colon C_{\mathbf{X};\underline{h}} = C_{\mathbf{X};\underline{h}+1} = \cdots = C_{\mathbf{X};\overline{h}} = 0$ versus the alternative that $C_{\mathbf{X};h} \neq 0$ for a lag h between $\underline{h}$ and $\overline{h}$ is based on the statistic

$$Q_{\underline{h},\overline{h}} = n^2 \sum_{h=\underline{h}}^{\overline{h}} \frac{1}{n-h} \mathrm{tr}(\hat{C}^\top_{\mathbf{X};h} \hat{C}^{-1}_{\mathbf{X};0} \hat{C}_{\mathbf{X};h} C^{-1}_{\mathbf{X};0}). \tag{4.2}$$

Under regularity conditions $Q_{\underline{h},\overline{h}}$ is asymptotically distributed as a $\chi^2_{d^2(\overline{h}-\underline{h}+1)}$ random variable if the multivariate procss $(\mathbf{X}_j\colon j \in \mathbb{Z})$ on the first d principal directions follows a VMA(q) model and $\underline{h} > q$. For practical implementation, one computes iteratively $Q_{1,\overline{h}}, Q_{2,\overline{h}}, \ldots$ and selects the order q as the largest $\underline{h}$ such that $Q_{\underline{h},\overline{h}}$ is significant, but $Q_{\underline{h}+h,\overline{h}}$ is insignificant for all $h > 0$.

Alternatively, the well-known AICC criterion could be utilized. Algorithm 3.1 allows for the computationally efficient maximization of the likelihood function through the use of its innovation form; see Chapter 11.5 of Brockwell and Davis [6]. The AICC criterion is then given by

$$\mathrm{AICC}(q) = -2\ln L(\Theta_1, \ldots, \Theta_q, \Sigma) + \frac{2nd(qd^2+1)}{nd - qd^2 - 2}, \tag{4.3}$$

where $\Theta_1, \ldots, \Theta_q$ are the fitted VMA coefficient matrices and Σ its fitted covariance matrix. The minimizer of (4.3) is selected as order of the FMA process. Both methods are compared in Section 5.

4.3 Functional FPE criterion

In this section a criterion that allows to choose d and q simultaneously is introduced. A similar criterion was established in Aue et al. [3], based on a decomposition of the functional mean squared prediction error. Note that, due to the orthogonality of the eigenfunctions $(\nu_i\colon i \in \mathbb{N})$ and the fact that $\hat{X}_{n+1,k}$ lives in $\mathcal{V}_d$,

$$\mathrm{E}\big[\|X_{n+1} - \hat{X}_{n+1,k}\|^2\big] = \mathrm{E}\big[\|P_{\mathcal{V}_d}(X_{n+1} - \hat{X}_{n+1,k})\|^2\big] + \mathrm{E}\big[\|(I_H - P_{\mathcal{V}_d})X_{n+1}\|^2\big]. \tag{4.4}$$

The second summand in (4.4) satisfies $\mathrm{E}[\|(I_H - P_{\mathcal{V}_d})X_{n+1}\|^2] = \mathrm{E}[\|\sum_{i>d}\langle X_{n+1}, \nu_i\rangle\nu_i\|^2] = \sum_{i>d}\lambda_i$. The first summand in (4.4) is, due to the isometric isomorphy between $\mathcal{V}_d$ and $\mathbb{R}^d$ equal to the mean squared prediction error of the vector model fit on the d dimensional principal subspace. It can be shown using the results of Lai and Lee [16] that it is of order $\mathrm{tr}(C_{\mathbf{Z}}) + qd\,\mathrm{tr}(C_{\mathbf{Z}})/n$, where $C_{\mathbf{Z}}$ denotes the covariance matrix of the innovations of the vector process. Using the matrix version $\mathbf{V}_n$ of the operator V_n given through Algorithm 3.1 as a consistent estimator for $C_{\mathbf{Z}}$, the functional FPE criterion

$$\mathrm{fFPE}(d,q) = \frac{n + q\,d}{n}\,\mathrm{tr}(\mathbf{V}_n) + \sum_{i>d}\hat{\lambda}_i \tag{4.5}$$

is obtained. It can be minimized over both d and q to select the dimension of the principal subspace and the order of the FMA process jointly. As is noted in Aue et al. [3], where a similar criterion is proposed for the selection of the order of an FAR(p) model, the fFPE method is fully data driven: no further selection of tuning parameters is required.

5 Simulation evidence

5.1 Simulation setting

In this section, results from Monte Carlo simulations are reported. The simulation setting was as follows. Using the first D Fourier basis functions $f_1, \ldots, f_D$, the D-dimensional subspace $G^D = \overline{\mathrm{sp}}\{f_1, \ldots, f_D\}$ of H was generated following the setup in Aue et al. [3], then the isometric isomorphy between $\mathbb{R}^D$ and G^D is utilized to represent elements in G^D by D-dimensional vectors and operators acting on G^D by $D \times D$ matrices. Therefore $N + q$ D-dimensional random vectors as innovations for an FMA(q) model and q $D \times D$ matrices as operators were generated. Two different settings were of interest: processes possessing covariance operators with slowly and quickly decaying eigenvalues. Those cases were represented by selecting two sets of standard deviations for the innovation process, namely

$$\sigma_{\text{slow}} = (i^{-1} \colon i = 1, \ldots, D) \qquad \text{and} \qquad \sigma_{\text{fast}} = (2^{-i} \colon i = 1, \ldots, D). \tag{5.1}$$

With this, innovations

$$\varepsilon_j = \sum_{i=1}^{D} c_{j,i} f_i, \qquad j = 1 - q, \ldots, n,$$

were simulated, where $c_{j,i}$ are independent normal random variables with mean 0 and standard deviation $\sigma_{\cdot,i}$, the $\cdot$ being replaced by either slow or fast, depending on the setting. The parameter operators $\tilde{\theta}_\ell$, for $\ell = 1, \ldots, q$, were chosen at random by generating $D \times D$ matrices, whose entries $\langle \tilde{\theta}_\ell f_i, f_{i'} \rangle$ were independent zero mean normal random variables with variance $\sigma_{\cdot,i}\sigma_{\cdot,i'}$. The matrices were then rescaled to have spectral norm 1. Combining the forgoing, the FMA(q) process

$$X_j = \sum_{\ell=1}^{q} \theta_\ell \varepsilon_{j-\ell} + \varepsilon_j, \qquad j = 1, \ldots, n \tag{5.2}$$

were simulated, where $\theta_\ell = \kappa_\ell \tilde{\theta}_\ell$ with κ_ℓ being chosen to ensure invertibility of the FMA process. In the following section, the performance of the proposed estimator is evaluated, and compared and contrasted to other methods available in the literature for the special case of FMA(1) processes, in a variety of situations.

5.2 Estimation of FMA(1) processes

In this section, the performance of the proposed method is compared to two approaches introduced in Turbillon et al. [27] for the special case of FMA(1) processes. These methods are based on the following idea. Denote by C_ε the covariance operator of $(\varepsilon_n \colon n \in \mathbb{Z})$. Observe that since $C_{X;1} = \theta_1 C_\varepsilon$ and $C_X = C_\varepsilon + \theta_1 C_\varepsilon \theta_1^*$, it follows that $\theta_1 C_X = \theta_1 C_\varepsilon + \theta_1^2 C_\varepsilon \theta_1^* = C_{X;1} + \theta_1^2 C_{X;1}^*$, and especially

$$\theta_1^2 C_{X;1}^* - \theta_1 C_X + C_{X;1} = 0. \tag{5.3}$$

The estimators in Turbillon et al. [27] are based on solving the quadratic equation in (5.3) for θ_1. The first of these only works under the restrictive assumption that θ_1 and C_ε commute. Then, solving (5.3) is equivalent to solving univariate equations generated by individually projecting (5.3) onto the eigenfunctions of C_X. The second approach is inspired by the Riesz–Nagy method. It relies on regarding (5.3) as a fixed-point equation and therefore establishing a fixed-point iteration. Since solutions may not exist in H, suitable projections have to be applied. Consistency of both estimators is established in Turbillon et al. [27].

To compare the performance of the methods, FMA(1) time series were simulated as described in Section 5.1. As measure of comparison the estimation error $\|\theta_1 - \hat{\theta}_1\|_{\mathcal{L}}$ was used after computing $\hat{\theta}_1$ with the three competing procedures. Rather than selecting the dimension of the subspace via Algorithm 4.1, the estimation error is computed for $d = 1, \ldots, 5$. The results are summarized in Table 5.1, where estimation errors were averaged over 1000 repetitions for each specification, using sample sizes $n = 100, 500$ and $1{,}000$.

		$n = 100$			$n = 500$			$n = 1000$		
	d	Proj	Iter	Inn	Proj	Iter	Inn	Proj	Iter	Inn
σ_{fast}	1	0.539	0.530	0.514	0.527	0.521	0.513	0.518	0.513	0.508
	2	0.528	0.433	**0.355**	0.508	0.391	0.287	0.500	0.386	0.277
	3	0.533	0.534	0.448	0.512	0.467	**0.235**	0.503	0.460	**0.197**
	4	0.534	0.650	0.582	0.513	0.573	0.276	0.504	0.567	0.216
	5	0.534	0.736	0.646	0.513	0.673	0.311	0.504	0.662	0.239
σ_{slow}	1	0.610	0.602	0.588	0.579	0.574	0.566	0.575	0.573	0.569
	2	0.614	0.527	**0.513**	0.581	0.487	0.434	0.577	0.483	0.422
	3	0.618	0.552	0.610	0.583	0.504	**0.389**	0.578	0.500	0.362
	4	0.620	0.591	0.861	0.584	0.531	0.402	0.579	0.522	**0.344**
	5	0.620	0.630	1.277	0.584	0.556	0.448	0.579	0.548	0.358

Table 5.1: Estimation error $\|\theta_1 - \hat{\theta}_1\|_{\mathcal{L}}$, with $\theta_1 = \kappa_1 \tilde{\theta}_1$ and $\kappa_1 = 0.8$, with $\hat{\theta}_1$ computed with the projection method (Proj) and the iterative method (Iter) of [27], and the proposed method based on the functional Innovations Algorithm (Inn). The smallest estimation error is highlighted in bold for each case.

For all three sample sizes, the operator kernel estimated with the proposed algorithm is closest to the real kernel. As can be expected, the optimal dimension increases with the sample size, especially for the case where the eigenvalues decay slowly. The projection method does not perform well, which is also to be expected, because the condition of commuting θ_1 and C_ε is violated. One can see that the choice of d is crucial: especially for small sample sizes for the proposed method, the estimation error explodes for large d. In order to get an intuition for the shape of the estimators, the kernels of the estimators resulting from the different estimation methods, using $n = 500$ and $\kappa_1 = 0.8$, are plotted in Figure 5.1. It can again be seen that the projection method yields results that are significantly different from both the truth and the other two methods who produce estimated operator kernels, whose shapes look roughly similar to the truth.

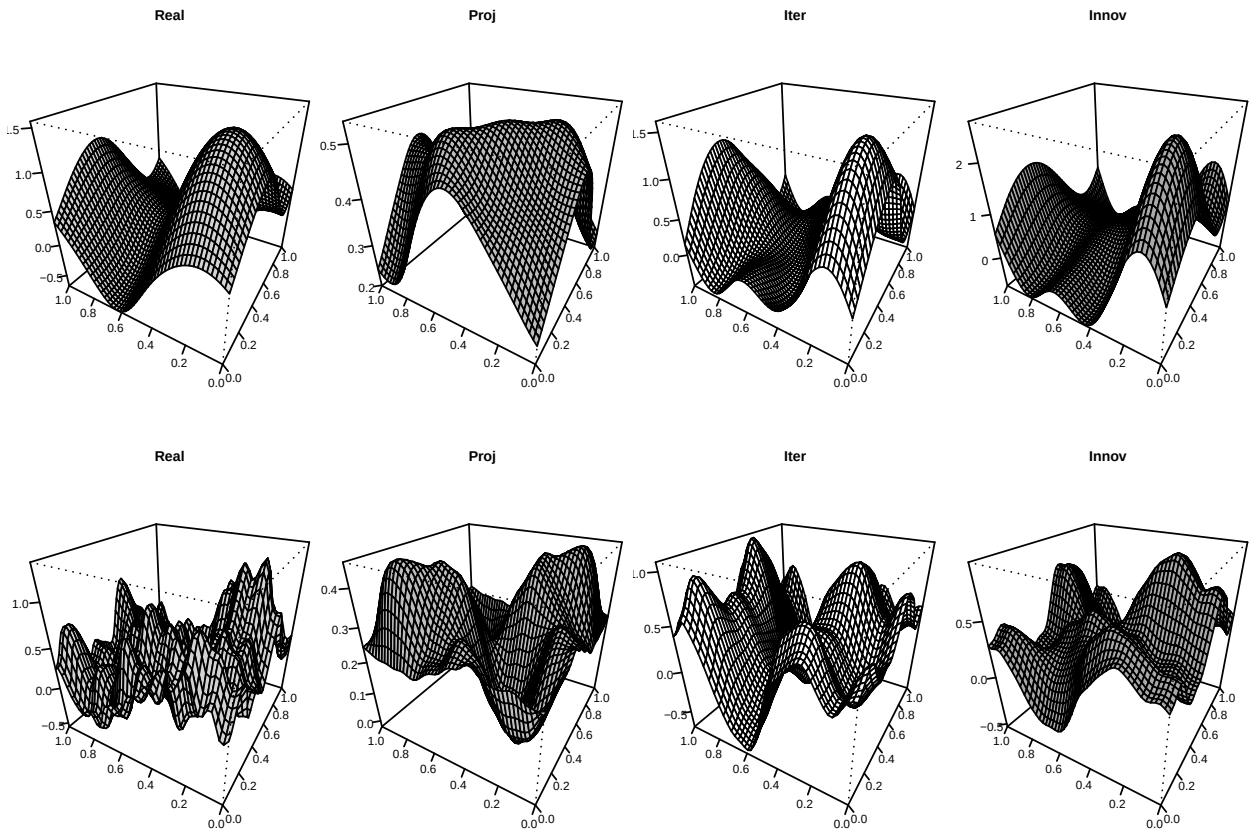

Figure 5.1: Estimated operator kernel of simulated FMA(1) process with $\kappa_1 = 0.8$, $d = 3$ and σ_{fast} (first row) and σ_{slow} (second row), using $n = 500$ sampled functions. Labeling of procedures is as in Table 5.1.

5.3 Model selection

In this section, the performance of the different model selection methods introduced in Section 4 is demonstrated. To do so, FMA(1) processes with weights $\kappa_1 = 0.4$ and 0.8 were simulated as in the previous section. In addition, two different FMA(3) processes were simulated according to the setting described in Section 5.1, namely

- Model 1: $\kappa_1 = 0.8$, $\kappa_2 = 0.6$, and $\kappa_3 = 0.4$.
- Model 2: $\kappa_1 = 0$, $\kappa_2 = 0$, and $\kappa_3 = 0.8$.

For sample sizes $n = 100$, 500 and $1{,}000$, $1{,}000$ processes of both Model 1 and 2 were simulated using σ_{slow} and σ_{fast}. The estimation process was done as follows. First, the dimension d of the principal projection subspace was chosen using Algorithm 4.1 with TVE such that $P = 0.8$. With this selection of d, the LB and AICC criteria described in Section 4.2 were applied to choose q. Second, the fFPE criterion was used for a simultaneous selection of d and q. The results are summarized in Figures 5.2 and 5.3.

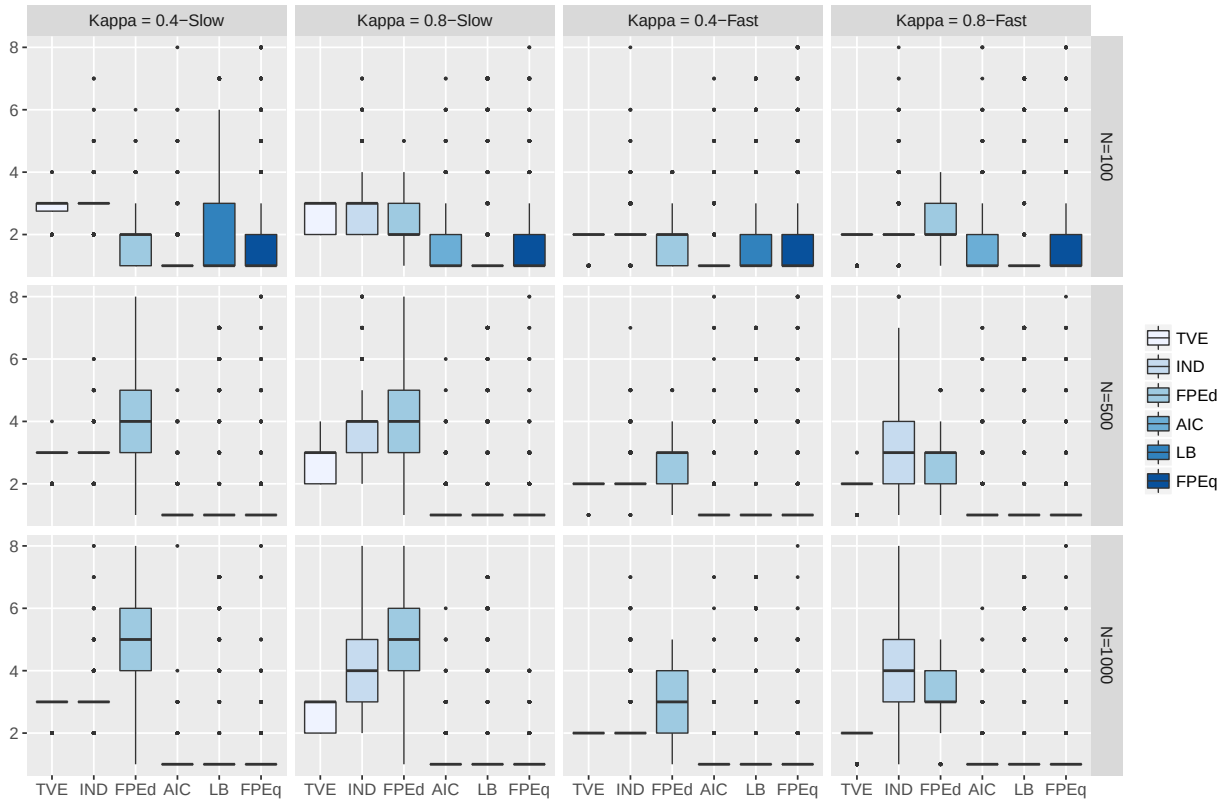

Figure 5.2: Model selection for different MA(1) processes. The left three plots in each small figure give the d chosen by total variation explained with $P = 0.8$ (TVE), Algorithm 4.1 (IND) and the functional FPE criterion (FPEd). The right three plots in each small figure give the selected order q by AICC, LB and fFPE.

Figures 5.2 and 5.3 allow for a number of interesting observations. For both the FMA(1) and the FMA(3) example, the model order is estimated well. In all cases, especially for sample sizes larger than 100, all three selection methods (AIC, LB, FPEq) for the choice of q yield the correct model order (1 or 3). The Ljung–Box (LB) method seems to have the most stable results. The methods for the choice of d are more heterogeneous. The TVE method yields the most stable results among different sample sizes. For σ_{fast}, it almost always selects $d = 2$ and for σ_{slow} the choice varies between $d = 2$ and $d = 3$. However, the TVE method seems to underestimate d. Often there appears to be dependence left in the data, as one can see from the selection of d by Algorithm 4.1. Especially in the FMA(3) case and Model 1, this algorithm yields some large choices for d of about 7 or 8. The choice of FPEd seems to increase with increasing sample size: this is to be expected as for increasing sample size the variance of the estimators decreases and the resulting predictors get more precise, even for high-dimensional models. This is valid especially for σ_{slow} where a larger d is needed to explain the dynamics of the functional process. A similar trade-off is occasionally observed for Algorithm 4.1.

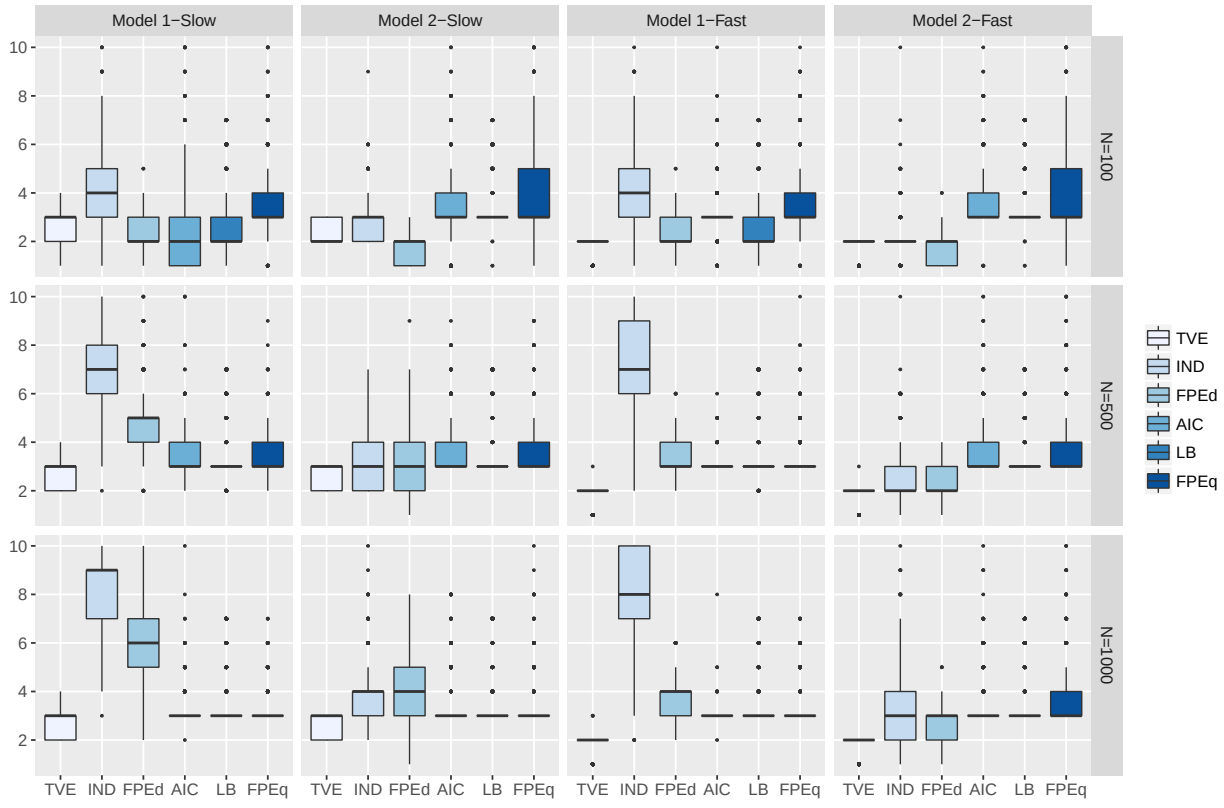

Figure 5.3: Model selection for different MA(3) processes. Labeling of procedures is as in Figure 5.2.

6 Application to traffic data

In this section, the proposed estimation method is applied to vehicle traffic data provided by the Autobahndirektion Südbayern. The dataset consists of measurements at a fixed point on a highway (A92) in Southern Bavaria, Germany. Recorded is the average velocity per minute from 1/1/2014 00:00 to 30/06/2014 23:59 on three lanes. After taking care of missing values and outliers, the velocity per minute was averaged over the three lanes, weighted by the number of vehicles per lane. This leads to 1440 preprocessed and cleaned data points per day, which were transformed into functional data using the first 30 Fourier basis functions with the `R` package `fda`. The result is a functional time series $(X_j \colon j = 1, \dots, n = 119)$, which is deemed stationary and exhibits temporal dependence, as evidenced by Klepsch et al. [14].

The goal then is to approximate the temporal dynamics in this stationary functional time series with an FMA fit. Observe that the plots of the spectral norms $\|\hat{C}_{\mathbf{X};h}\hat{C}_{\mathbf{X};0}^{-1}\|_{\mathcal{L}}$ for $h = 0, \dots, 5$ in Figure 6.1 display a pattern typical for MA models of low order. Here $\mathbf{X}$ stands for the multivariate auxiliary model of dimension d obtained from projection into the corresponding principal subspace. Consequently, the methodology introduced in Section 3 and 4 was applied to the data. First, the covariance operator $C_{X;0}$ and its first 15

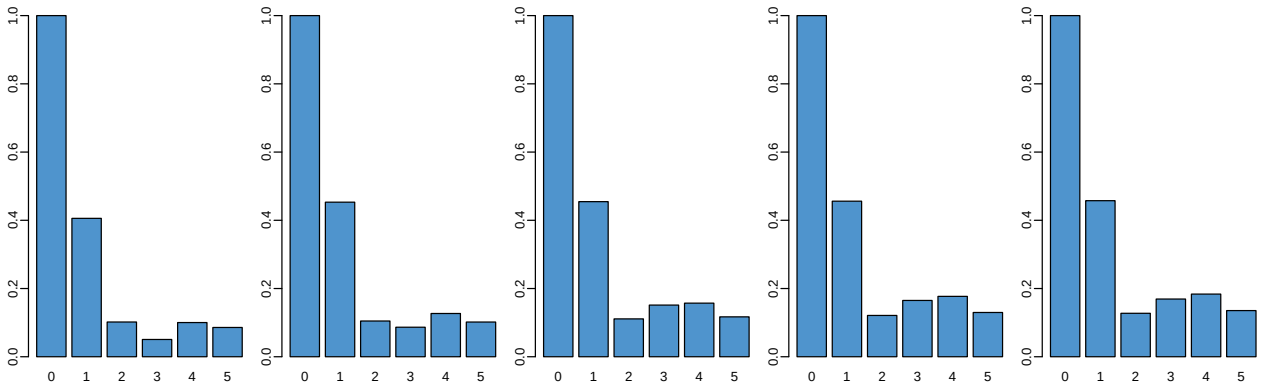

Figure 6.1: Spectral norm of estimated cross-correlation matrices for lags $h = 1, \ldots, 5$ of the vector model based on principal subspaces of dimension $d = 1$ to $d = 5$ (from left to right).

eigenelements $(\lambda_1, \nu_1), \ldots, (\lambda_{15}, \nu_{15})$ were estimated to construct the vector process $(\hat{\mathbf{X}}_j \colon j = 1, \ldots, n)$, where $\hat{\mathbf{X}}_j = (\langle X_j, \hat{\nu}_1 \rangle, \ldots, \langle X_j, \hat{\nu}_{15} \rangle)^\top$. Then, the methods described in Sections 4 were applied to choose the appropriate dimension d and model order q.

The first four sample eigenfunctions explained 81% of the variability, hence the TVE criterion with $P = 0.8$ gave $d^* = 4$ to initialize Algorithm 4.1. The hypothesis of independence of the left-out score vector process $(\hat{\mathbf{X}}_j[4{:}15] \colon j = 1, \ldots, n)$ was rejected with p-value 0.03. Here $\mathbf{X}_j[i{:}i']$ is used as notation for the vector comprised of coordinates $i, \ldots, i'$, with $i \leq i'$, of the original 15-dimensional vector $\hat{\mathbf{X}}_j$. In the next step of Algorithm 4.1, d^* is increased to 5. A second independence test was run on $(\hat{\mathbf{X}}_j[5{:}15] \colon j = 1, \ldots, n)$ and did not result in a rejection; the corresponding p-value was 0.25.

This analysis led to using $d = 5$ as dimension of the principal subspace to conduct model selection with the methods of Section 4.2. Since TVE indicated $d = 4$, the selection procedures were applied also with this choice. In both cases, the AICC criterion in (4.3) and LB criterion in (4.2) opted for $q = 1$, in accordance with the spectral norms observed in Figure 6.1. Simultaneously choosing d and q with the fFPE criterion of Section 4.3 yields $d = 3$ and $q = 1$.

After the model selection step, the operator of the chosen FMA(1) process was estimated using Algorithm 3.1. Similarly the methods introduced in Section 5.2 were applied. Figure 6.2 displays the kernels of the estimated integral operator for all methods, selecting for $d = 3$ and $d = 5$. The plots indicate that, on this particular data set, all three methods produce estimated operators that lead to kernels of roughly similar shape. The similarity is also reflected in the covariance of the estimated innovations. For $d = 3$, the trace of the covariance matrix is 43.14, 45.4 and 44.41 for the Innovations Algorithm, iterative method and projective method, respectively. For $d = 4$, the trace of the covariance of the estimated innovations is 48.19, 46.00 and 45.74 for the different methods in the same order.

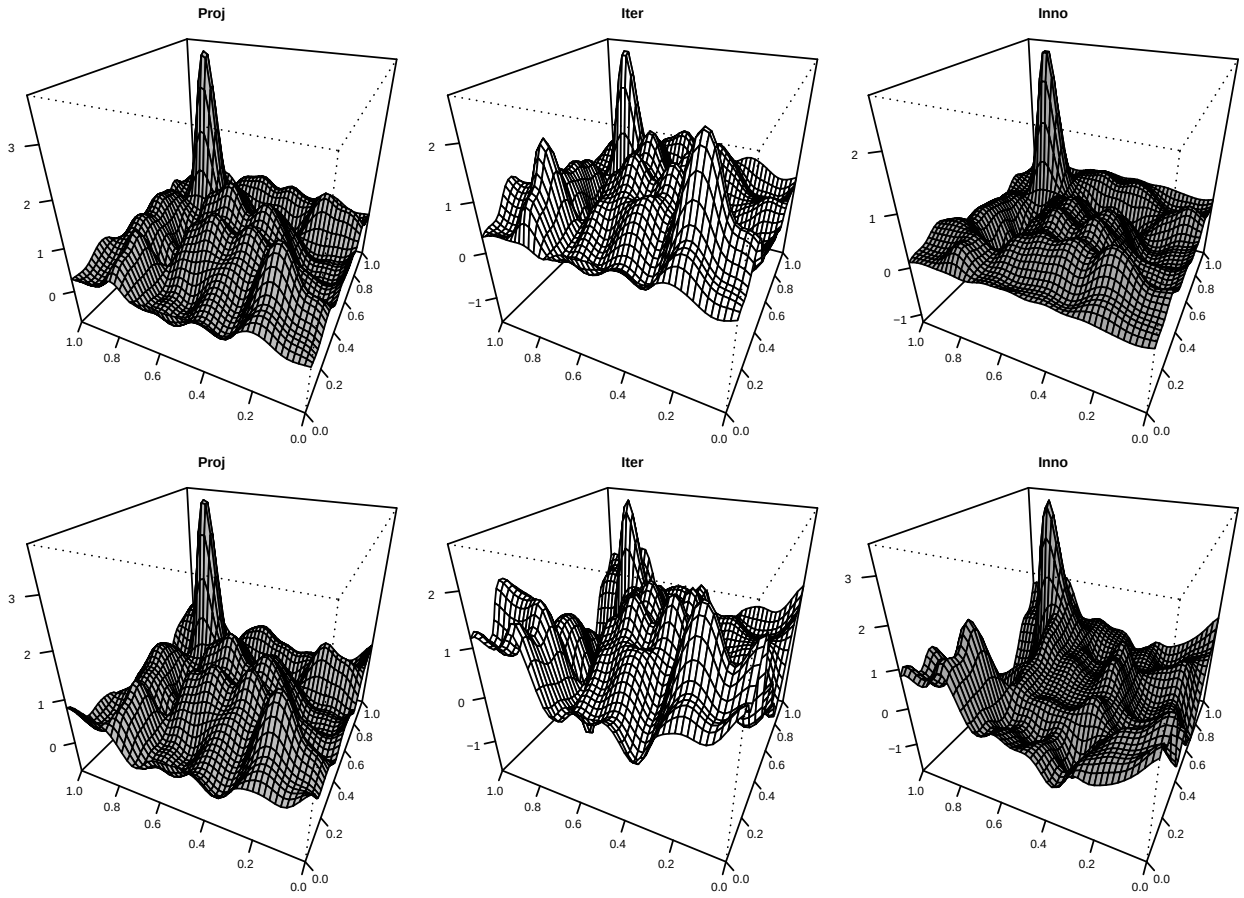

Figure 6.2: Estimated FMA(1) kernel with the three methods for $d = 3$ (first row) and $d = 4$ (second row)

7 Conclusions

This paper is the first to introduce a complete methodology to estimate any stationary, causal and invertible functional time series. This is achieved by approximating the functional linear filters in the causal representation with functional moving average processes obtained from an application of the functional Innovations Algorithm. The consistency of the estimators is verified as the main theoretical contribution. The proof relies on the fact that d-dimensional projections of FMA(q) processes are isomorph to d dimensional VMA(q^*) models, with $q^* \leq q$. Introducing appropriate sequences of increasing subspaces of H, consistency can be established in the two cases of known and unknown principal projection subspaces. This line of reasoning follows multivariate techniques given in Lewis and Reinsel [17] and Mitchell and Brockwell [20].

The theoretical underpinnings are accompanied by model selection procedures facilitating the practical implementation of the proposed method. An independence test is introduced to select the dimension of the principal projection subspace, which can be used as a starting point for the suggested order selection procedures based on AICC and Ljung–Box criteria. Additionally, an fFPE criterion is established that jointly selects

dimension d and order q. Illustrative results from a simulation study and the analysis of traffic velocity data show that the practical performance of the proposed method is satisfactory and at least competitive with other methods available in the literature for the case of FMA(1) processes.

Future research could focus on an extension of the methodology to FARMA processes in order to increase parsimony in the estimation. It should be noted, however, that this not a straightforward task as identifying the dynamics of the projection of an FARMA(p, q) to a finite-dimensional space is a non-resolved problem. In addition, the proposed methodology could be applied to offer an alternative route to estimate the spectral density operator, a principal object in the study of functional time series in the frequency domain; see Aue and van Delft [1], Hörmann et al. [9] and Panaretos and Tavakoli [22].

Acknowledgement: We thank the Autobahndirektion Südbayern and especially J. Grötsch for their support and for providing the traffic data. J. Klepsch furthermore acknowledges financial support from the Munich Center for Technology and Society based project ASHAD.

8 Proofs

The notion of L^p-m-approximability is utilized for the proofs. A version of this notion was used for multivariate time series in Aue et al. [2] and then translated to the functional domain by Hörmann and Kokoszka [10]. The definition is as follows.

Definition 8.1. Let $p \geq 1$. A sequence $(X_j \colon j \in \mathbb{Z})$ with values in L_H^p is called *L^p-m-approximable* if

$$X_j = f(\varepsilon_j, \varepsilon_{j-1}, \ldots), \qquad j \in \mathbb{Z},$$

can be represented as a functional Bernoulli shift with a sequence of independent, identically distributed random elements $(\varepsilon_j \colon j \in \mathbb{Z})$ taking values in the measurable space S, potentially different from H, and a measurable function $f \colon S^\infty \to H$ such that

$$\sum_{m=0}^{\infty} \left(\mathrm{E}[\|X_j - X_j^{(m)}\|^p]\right)^{1/p} < \infty,$$

where $X_j^{(m)} = f(\varepsilon_j, \ldots, \varepsilon_{j-m+1}, \varepsilon_{j-m}^{(j)}, \varepsilon_{j-m-1}^{(j)}, \ldots)$ with $(\varepsilon_j^{(i)} \colon j \in \mathbb{Z})$, $i \in \mathbb{N}_0$, being independent copies of $(\varepsilon_j \colon j \in \mathbb{Z})$.

Conditions can be established for most of the common linear and nonlinear functional time series models to be L^p-m-approximable. In particular, the functional linear processes $(X_j \colon j \in \mathbb{Z})$ defined in (2.1) are naturally included if the summability condition $\sum_{m=1}^{\infty} \sum_{\ell=m}^{\infty} \|\psi_\ell\|_{\mathcal{L}} < \infty$ is met (see Proposition 2.1 in Hörmann and Kokoszka [10]).

Proof of Theorem 3.1. Using that $(X_j\colon j\in\mathbb{Z})$ is L^4-m-approximable, write

$$\begin{aligned} X_j(k) &= (f(\varepsilon_j,\varepsilon_{j-1},\dots),\dots,f(\varepsilon_{j-k+1},\varepsilon_{j-k},\dots))^\top \\ &= g(\varepsilon_j,\varepsilon_{j-1},\dots), \end{aligned}$$

where $g\colon H^\infty\to H^k$ is defined accordingly. For $k,m\in\mathbb{N}$ and $j\in\mathbb{Z}$, define

$$\begin{aligned} X_j^{(m)}(k) &= \big(f(\varepsilon_j,\dots,\varepsilon_{j-m+1},\varepsilon_{j-m}^{(j)},\varepsilon_{j-m-1}^{(j)},\dots),\dots,f(\varepsilon_{j-k+1},\dots,\varepsilon_{j-m+1},\varepsilon_{j-m}^{(j)},\varepsilon_{j-m-1}^{(j)},\dots)\big)^\top \\ &= g(\varepsilon_j,\varepsilon_{j-1},\dots,\varepsilon_{j-m+1},\varepsilon_{j-m}^{(j)},\varepsilon_{j-m-1}^{(j)},\dots). \end{aligned}$$

Now, by definition of the norm in H^k,

$$\begin{aligned} \sum_{m=k}^\infty \big(\mathrm{E}\big[\|X_m(k)-X_m^{(m)}(k)\|^4\big]\big)^{1/4} &= \sum_{m=k}^\infty\bigg(\sum_{i=0}^{k-1}\mathrm{E}\big[\|X_{m-i}-X_{m-i}^{(m-i)}\|^4\big]\bigg)^{1/4} \\ &\le \sum_{m=k}^\infty\bigg(\sum_{i=0}^{k-1}\mathrm{E}\big[\|X_{m-i}-X_{m-i}^{(m-k)}\|^4\big]\bigg)^{1/4} \\ &= \sum_{m=k}^\infty\big(k\mathrm{E}\big[\|X_{m-k}-X_{m-k}^{(m-k)}\|^4\big]\big)^{1/4} \\ &= k^{1/4}\sum_{m=0}^\infty\big(\mathrm{E}\big[\|X_m-X_m^{(m)}\|^4\big]\big)^{1/4}, \end{aligned} \tag{8.1}$$

where the first inequality is implied by Assumption 3.1, since $\mathrm{E}[\|X_j-X_j^{(m-i)}\|^2]\le\mathrm{E}[\|X_j-X_j^{(m)}\|^2]$ for all $i\ge 0$, and the last inequality, since $\mathrm{E}[\|X_1-X_1^{(m-k)}\|^2]=\mathrm{E}[\|X_j-X_j^{(m-k)}\|^2]$ by stationarity. But the right-hand side of (8.1) is finite because $(X_j\colon j\in\mathbb{Z})$ is L^4-m-approximable by assumption. This shows that $(X_j(k)\colon j\in\mathbb{Z})$ is also L^4-m approximable.

To prove the consistency of the estimator $\hat{C}_{X(k)}$, note that the foregoing implies, by Theorem 3.1 in Hörmann and Kokoszka [10], that the bound

$$n\,\mathrm{E}\big[\|\hat{C}_{X(k)}-C_{X(k)}\|_{\mathcal{N}}^2\big]\le U_{X(k)},$$

holds, where $U_{X(k)}=\mathrm{E}[\|X_1(k)\|^4]+4\sqrt{2}(\mathrm{E}[\|X_1(k)\|^4])^{3/4}\sum_{m=0}^\infty(\mathrm{E}[\|X_m(k)-X_m^{(m)}(k)\|^4])^{1/4}$ is a constant that does not depend on n. Since $\mathrm{E}[\|X_1(k)\|^4]=k\mathrm{E}[\|X_1\|^4]$, (8.1) yields that $U_{X(k)}=kU_X$, which is the assertion. □

Corollary 8.1. *The operators $\hat{\beta}_{k,i}$ from* (3.10) *and $\hat{\theta}_{k,i}$ from* (3.4) *related through*

$$\hat{\theta}_{k,i}=\sum_{j=1}^i\hat{\beta}_{k,j}\hat{\theta}_{k-j,i-j},\qquad i=1,\dots,k,\ k\in\mathbb{N}. \tag{8.2}$$

Proof. The proof is based on the finite-sample versions of the regression formulation of (3.1) and the innovations formulation given in (3.4). Details are omitted to conserve space. □

Proof of Theorem 3.2. *(i)* It is first shown that, for all $x \in H^k$,

$$\|(\hat{B}(k) - \Pi(k))(x)\| \xrightarrow{p} 0 \qquad (n \to \infty),$$

where $\Pi(k) = (\pi_1, \ldots, \pi_k)^\top$ is the vector of the first k operators in the invertibility representation of the functional time series $(X_j \colon j \in \mathbb{Z})$. Define the process $(e_{j,k} \colon j \in \mathbb{Z})$ by letting

$$e_{j,k} = X_j - \sum_{\ell=1}^{k} \pi_\ell X_{j-\ell} \tag{8.3}$$

and let I_{H^k} be the identity operator on H^k. Note that

$$\begin{aligned}\hat{B}(k) - \Pi(k) &= \hat{\Gamma}_{1,k,d}\hat{\Gamma}_{k,d}^{-1} - \Pi(k)\hat{\Gamma}_{k,d}\hat{\Gamma}_{k,d}^{-1} + \Pi(k)(I_{H^k} - P_{(k)}) \\ &= \big(\Gamma_{1,k,d} - \Pi(k)\hat{\Gamma}_{k,d}\big)\hat{\Gamma}_{k,d}^{-1} + \Pi(k)(I_{H^k} - P_{(k)}).\end{aligned}$$

Plugging in the estimators defined in (3.9) and subsequently using (8.3), it follows that

$$\begin{aligned}\hat{B}(k) - \Pi(k) &= \left(\frac{1}{n-k}\sum_{j=k}^{n-1}\big((P_{(k)}X_{j,k} \otimes X_{j+1}) - (P_{(k)}X_{j,k} \otimes \Pi(k)X_{j,k})\big)\right)\hat{\Gamma}_{k,d}^{-1} + \Pi(k)(I_{H^k} - P_{(k)}) \\ &= \left(\frac{1}{n-k}\sum_{j=k}^{n-1}\big(P_{(k)}X_{j,k} \otimes (X_{j+1} - \Pi(k)X_{j,k})\big)\right)\hat{\Gamma}_{k,d}^{-1} + \Pi(k)(I_{H^k} - P_{(k)}) \\ &= \left(\frac{1}{n-k}\sum_{j=k}^{n-1}\big(P_{(k)}X_{j,k} \otimes e_{j+1,k}\big)\right)\hat{\Gamma}_{k,d}^{-1} + \Pi(k)(I_{H^k} - P_{(k)}).\end{aligned}$$

Two applications of the triangle inequality imply that, for all $x \in H^k$,

$$\begin{aligned}\|(\hat{B}(k) - \Pi(k))(x)\| &\le \left\|\left(\frac{1}{n-k}\sum_{j=k}^{n-1}\big(P_{(k)}X_j(k) \otimes e_{j+1,k}\big)\right)\hat{\Gamma}_{k,d}^{-1}(x)\right\| + \|\Pi(k)(I_{H^k} - P_{(k)})(x)\| \\ &\le \left\|\left(\frac{1}{n-k}\sum_{j=k}^{n-1}\big(P_{(k)}X_j(k) \otimes (e_{j+1,k} - \varepsilon_{j+1})\big)\right)\hat{\Gamma}_{k,d}^{-1}\right\|_{\mathcal{L}} \\ &\quad + \left\|\left(\frac{1}{n-k}\sum_{j=k}^{n-1}\big(P_{(k)}X_j(k) \otimes \varepsilon_{j+1}\big)\right)\hat{\Gamma}_{k,d}^{-1}\right\|_{\mathcal{L}} + \|\Pi(k)(I_{H^k} - P_{(k)})(x)\| \\ &\le \big(\|U_{1n}\|_{\mathcal{L}} + \|U_{2n}\|_{\mathcal{L}}\big)\|\hat{\Gamma}_{k,d}^{-1}\|_{\mathcal{L}} + \|\Pi(k)(I_{H^k} - P_{(k)})(x)\|,\end{aligned} \tag{8.4}$$

where U_{1n} and U_{2n} have the obvious definitions. Arguments similar to those used in Proposition 6.4 of Klepsch and Klüppelberg [13] yield that the second term on the right-hand side of (8.4) can be made arbitrarily small by increasing k. To be more precise, for $\delta > 0$, there is $k_\delta \in \mathbb{N}$ such that

$$\|\Pi(k)(I_{H^k} - P_{(k)})(x)\| < \delta \tag{8.5}$$

for all $k \ge k_\delta$ and all $x \in H^k$.

To estimate the first term on the right-hand side of (8.4), focus first on $\|\hat{\Gamma}_{k,d}^{-1}\|_{\mathcal{L}}$. Using the triangular inequality, $\|\hat{\Gamma}_{k,d}^{-1}\|_{\mathcal{L}} \leq \|\hat{\Gamma}_{k,d}^{-1} - \Gamma_{k,d}^{-1}\|_{\mathcal{L}} + \|\Gamma_{k,d}^{-1}\|_{\mathcal{L}}$. Theorem 1.2 in Mitchell [19] and Lemma 6.1 in Klepsch and Klüppelberg [13] give the bound

$$\|\Gamma_{k,d}^{-1}\|_{\mathcal{L}} \leq \alpha_{d_k}^{-1}, \tag{8.6}$$

where α_{d_k} is the infimum of the eigenvalues of all spectral density operators of $(X_{d_k,j} \colon j \in \mathbb{Z})$. Furthermore, using the triangle inequality and then again Lemma 6.1 of Klepsch and Klüppelberg [13],

$$\begin{aligned}\|\hat{\Gamma}_{k,d}^{-1} - \Gamma_{k,d}^{-1}\|_{\mathcal{L}} &= \|\hat{\Gamma}_{k,d}^{-1}(\hat{\Gamma}_{d,k} - \Gamma_{d,k})\Gamma_{k,d}^{-1}\|_{\mathcal{L}} \\ &\leq \big(\|\hat{\Gamma}_{k,d}^{-1} - \Gamma_{k,d}^{-1}\|_{\mathcal{L}} + \|\Gamma_{k,d}^{-1}\|_{\mathcal{L}}\big)\|\hat{\Gamma}_{d,k} - \Gamma_{d,k}\|_{\mathcal{L}}\alpha_{d_k}^{-1}. \end{aligned} \tag{8.7}$$

Hence, following arguments in the proof of Theorem 1 in Lewis and Reinsel [17],

$$0 \leq \frac{\|\hat{\Gamma}_{k,d}^{-1} - \Gamma_{k,d}^{-1}\|_{\mathcal{L}}}{\alpha_{d_k}^{-1}(\|\hat{\Gamma}_{k,d}^{-1} - \Gamma_{k,d}^{-1}\|_{\mathcal{L}} + \alpha_{d_k}^{-1})} \leq \|\hat{\Gamma}_{d,k} - \Gamma_{d,k}\|_{\mathcal{L}},$$

by (8.7). This yields

$$\|\hat{\Gamma}_{d,k}^{-1} - \Gamma_{d,k}^{-1}\|_{\mathcal{L}} \leq \frac{\|\hat{\Gamma}_{d,k} - \Gamma_{d,k}\|_{\mathcal{L}}\alpha_{d_k}^{-2}}{1 - \|\hat{\Gamma}_{d,k} - \Gamma_{d,k}\|_{\mathcal{L}}\alpha_{d_k}^{-1}}. \tag{8.8}$$

Note that, since $P_{(k)}P_k = P_{(k)}$, $\|\Gamma_{k,d}\|_{\mathcal{L}} = \|P_{(k)}P_k\Gamma_k P_k P_{(k)}\|_{\mathcal{L}} \leq \|P_k\Gamma_k P_k\|_{\mathcal{L}}$. Also, by Theorem 3.1, for some positive finite constant M_1, $\mathrm{E}[\|P_k\hat{\Gamma}_k P_k - P_k\Gamma_k P_k\|^2] \leq M_1 k/(n-k)$. Therefore,

$$\|\hat{\Gamma}_{d,k} - \Gamma_{d,k}\| = O_p\bigg(\sqrt{\frac{k}{n-k}}\bigg). \tag{8.9}$$

Hence, the second part of Assumption 3.1 and (8.8) lead first to $\|\hat{\Gamma}_{d,k}^{-1} - \Gamma_{d,k}^{-1}\|_{\mathcal{L}} \xrightarrow{p} 0$ and, consequently, combining the above arguments,

$$\|\hat{\Gamma}_{k,d}^{-1}\|_{\mathcal{L}} = O_p(\alpha_{d_k}^{-1}). \tag{8.10}$$

Next consider U_{1n} in (8.4). With the triangular and Cauchy–Schwarz inequalities, calculate

$$\begin{aligned}\mathrm{E}[\|U_{1n}\|] &= \mathrm{E}\bigg[\bigg\|\frac{1}{n-k}\sum_{j=k}^{n-1} P_{(k)}X_j(k) \otimes (e_{j+1,k} - \varepsilon_{j+1})\bigg\|_{\mathcal{L}}\bigg] \\ &\leq \frac{1}{n-k}\sum_{j=k}^{N-1} \mathrm{E}\bigg[\bigg\|P_{(k)}X_j(k) \otimes (e_{j+1,k} - \varepsilon_{j+1})\bigg\|_{\mathcal{L}}\bigg] \\ &\leq \frac{1}{n-k}\sum_{j=k}^{N-1} \big(\mathrm{E}[\|P_{(k)}X_j(k)\|^2]\big)^{1/2}\big(\mathrm{E}[\|e_{j+1,k} - \varepsilon_{j+1}\|^2]\big)^{1/2}. \end{aligned}$$

The stationarity of $(X_j\colon j\in\mathbb{Z})$ and the fact that $X_j\in L_H^2$ imply that, for a positive finite constant M_2,

$$\begin{aligned}
\mathrm{E}[\|U_{1n}\|_{\mathcal{L}}] &\le \big(\mathrm{E}[\|P_{(k)}X_j(k)\|^2]\big)^{1/2}\big(\mathrm{E}[\|e_{j+1,k}-\varepsilon_{j+1}\|^2]\big)^{1/2}\\
&\le \sqrt{k}\big(\mathrm{E}[\|P_{\mathcal{V}_{d_k}}X_0\|^2]\big)^{1/2}\bigg(\mathrm{E}\bigg[\bigg\|\sum_{\ell>k}\pi_\ell X_{1-\ell}+\sum_{\ell=1}^{k}\pi_\ell(I_H-P_{\mathcal{V}_{d_{k+1-\ell}}})X_{1-\ell}\bigg\|^2\bigg]\bigg)^{1/2}\\
&\le \sqrt{k}\bigg(2\mathrm{E}\bigg[\|\sum_{\ell>k}\pi_\ell X_{j+1-\ell}\|^2\bigg]+2\mathrm{E}\bigg[\bigg\|\sum_{\ell=1}^{k}\pi_\ell(I_H-P_{\mathcal{V}_{d_{k+1-\ell}}})X_{1-\ell}\bigg\|^2\bigg]\bigg)^{1/2}\\
&= M_2\sqrt{k(J_1+J_2)}\\
&\le M_2\sqrt{k}(\sqrt{J_1}+\sqrt{J_2}),
\end{aligned}\tag{8.11}$$

where J_1 and J_2 have the obvious definition. Since for $X\in L_H^2$, $\mathrm{E}[\|X\|^2]=\|C_X\|_{\mathcal{N}}$, the term J_1 can be bounded as follows. Observe that

$$\begin{aligned}
J_1 &= \bigg\|\mathrm{E}\bigg[\sum_{\ell>k}\pi_\ell X_{1-\ell}\otimes\sum_{\ell'>k}\pi_{\ell'}X_{1-\ell'}\bigg]\bigg\|_{\mathcal{N}}\\
&= \bigg\|\sum_{\ell,\ell'>k}\pi_\ell C_{X;\ell-\ell'}\pi_{\ell'}^*\bigg\|_{\mathcal{N}}\\
&\le \sum_{\ell,\ell'>k}\|\pi_\ell\|_{\mathcal{L}}\|\pi_{\ell'}\|_{\mathcal{L}}\|C_{X;\ell-\ell'}\|_{\mathcal{N}}.
\end{aligned}$$

Now $C_{X;\ell-\ell'}\in\mathcal{N}$ for all $\ell,\ell'\in\mathbb{Z}$, hence $\|C_{X;\ell-\ell'}\|_{\mathcal{N}}\le M_3$ and $J_1\le M_3(\sum_{\ell>k}\|\pi_\ell\|_{\mathcal{L}})^2$. Concerning J_2, note first that, since $\mathrm{E}[\|X\|^2]=\|C_X\|_{\mathcal{N}}$,

$$J_2=\bigg\|\mathrm{E}\bigg[\sum_{\ell=1}^{k}\pi_\ell(I_H-P_{\mathcal{V}_{d_{k+1-\ell}}})X_{1-\ell}\otimes\sum_{\ell'=1}^{n}\pi_{\ell'}(I_H-P_{\mathcal{V}_{d_{k+1-\ell'}}})X_{1-\ell'}\bigg]\bigg\|_{\mathcal{N}}.$$

Using the triangle inequality together with properties of the nuclear operator norm and the definition of $C_{X;h}$ in display (2.4) leads to

$$\begin{aligned}
J_2 &\le \sum_{\ell,\ell'=1}^{k}\|\pi_\ell\|_{\mathcal{L}}\|\pi'_\ell\|_{\mathcal{L}}\big\|\mathrm{E}\big[(I_H-P_{\mathcal{V}_{d_{k+1-\ell}}})X_{1-\ell}\otimes(I_H-P_{\mathcal{V}_{d_{k+1-\ell'}}})X_{1-\ell'}\big]\big\|_{\mathcal{N}}\\
&= \sum_{\ell,\ell'=1}^{k}\|\pi_\ell\|_{\mathcal{L}}\|\pi'_\ell\|_{\mathcal{L}}\big\|(I_H-P_{\mathcal{V}_{d_{k+1-\ell}}})C_{X;\ell-\ell'}(I_H-P_{\mathcal{V}_{d_{k+1-\ell'}}})\big\|_{\mathcal{N}}\\
&= \sum_{\ell,\ell'=1}^{k}\|\pi_\ell\|_{\mathcal{L}}\|\pi'_\ell\|_{\mathcal{L}}K(\ell,\ell').
\end{aligned}\tag{8.12}$$

By the definition of $\mathcal{V}_d$ in (3.2) and since $(I_H-P_{\mathcal{V}_{d_i}})=\sum_{r>d_i}\nu_r\otimes\nu_r$, it follows that

$$K(\ell,\ell')=\bigg\|\sum_{s>d_{k+1-\ell'}}\sum_{r>d_{k+1-\ell}}\langle C_{X;\ell-\ell'}(\nu_r),\nu_s\rangle\nu_r\otimes\nu_s\bigg\|_{\mathcal{N}}$$

$$
\begin{aligned}
&\leq \Big\| \sum_{s>d_{k+1-\ell'}} \sum_{r>d_{k+1-\ell}} \sqrt{\lambda_r \lambda_s} \nu_r \otimes \nu_s \Big\|_{\mathcal{N}} \\
&= \sum_{i=1}^{\infty} \Big\langle \sum_{s>d_{k+1-\ell'}} \sum_{r>d_{k+1-\ell}} \sqrt{\lambda_r \lambda_s} \nu_r \otimes \nu_s(\nu_i), \nu_i \Big\rangle \\
&\leq \sum_{i>d_{k+1-\ell}} \lambda_i, \qquad (8.13)
\end{aligned}
$$

where Lemma 6.2 in Klepsch and Klüppelberg [13] was applied to give $\langle C_{X;\ell-\ell'} \nu_r, \nu_s \rangle \leq \sqrt{\lambda_r \lambda_s}$. Plugging (8.13) into (8.12), and recalling that $\sum_{\ell=1}^{\infty} \|\pi_\ell\|_{\mathcal{L}} = M_4 < \infty$, gives that

$$
J_2 \leq M_4 \sum_{\ell=1}^{k} \|\pi_\ell\|_{\mathcal{L}} \sum_{i>d_{k+1-\ell}} \lambda_i. \qquad (8.14)
$$

Inserting the bounds for J_1 and J_2 into (8.11), for some $M < \infty$,

$$
\begin{aligned}
\mathrm{E}[\|U_{1n}\|] &\leq \sqrt{k} M_2 (M_3 \sqrt{J_1} + \sqrt{J_2}) \\
&\leq \sqrt{k} M_2 \Big(M_3 \sum_{\ell>k} \|\pi_\ell\|_{\mathcal{L}} + M_4 \sum_{\ell=1}^{k} \|\pi_\ell\|_{\mathcal{L}} \sum_{i>d_{k+1-\ell}} \lambda_i \Big) \\
&\leq \sqrt{k} M \Big(\sum_{\ell>k} \|\pi_\ell\|_{\mathcal{L}} + \Big(\sum_{\ell=1}^{k} \|\pi_\ell\|_{\mathcal{L}} \sum_{i>d_{k+1-\ell}} \lambda_i) \Big). \qquad (8.15)
\end{aligned}
$$

Concerning U_{2n} in (8.4), use the linearity of the scalar product, the independence of the innovations $(\varepsilon_j \colon j \in \mathbb{Z})$ and the stationarity of the functional time series $(X_j \colon j \in \mathbb{Z})$ to calculate

$$
\begin{aligned}
E[\|U_{2n}\|^2] &\leq \left(\frac{1}{n-k} \right)^2 \sum_{j=k}^{n-1} \mathrm{E}\big[\|P_{(k)} X_j(k)\|^2\big] \mathrm{E}\big[\|\varepsilon_{j+1}\|^2\big] \\
&\leq \frac{1}{n-k} \mathrm{E}\big[\|P_{(k)} X_0(k)\|^2\big] \mathrm{E}\big[\|\varepsilon_0\|^2\big] \\
&\leq \frac{k}{n-k} \mathrm{E}\big[\|X_0\|^2\big] \mathrm{E}\big[\|\varepsilon_0\|^2\big].
\end{aligned}
$$

Since both $(X_j \colon j \in \mathbb{Z})$ and $(\varepsilon_j \colon j \in \mathbb{Z})$ are in L_H^2, (8.10) implies that

$$
\|U_{2n}\|_{\mathcal{L}} \|\hat{\Gamma}_{k,d}^{-1}\|_{\mathcal{L}} = O_p\Big(\frac{1}{\alpha_{d_k}} \sqrt{\frac{k}{n-k}} \Big).
$$

Furthermore, (8.10) and (8.15) show that

$$
\|U_{1n}\|_{\mathcal{L}} \|\hat{\Gamma}_{k,d}^{-1}\|_{\mathcal{L}} = O_p\Big(\frac{\sqrt{k}}{\alpha_{d_k}} \Big(\sum_{\ell>k} \|\pi_\ell\|_{\mathcal{L}} + \sum_{\ell=1}^{k} \|\pi_\ell\|_{\mathcal{L}} \sum_{i>d_{k+1-\ell}} \lambda_i \Big) \Big).
$$

Thus Assumption 3.1, (8.4) and (8.5) assert that, for all $x \in H^k$, $\|\hat{B}_k - \Pi(k)(x)\| \xrightarrow{p} 0$, which proves the first statement of the theorem.

(ii) First note that, for all $x \in H^k$, $\|(\hat{\beta}_{k,i} - \beta_{k,i})(x)\| \leq \|(\hat{\beta}_{k,i} - \pi_i)(x)\| + \|(\pi_i - \beta_{k,i})(x)\| \xrightarrow{p} 0$ as $n \to \infty$. Now $\theta_{k,1} = \beta_{k,1}$ and by Corollary 8.1 $\hat{\theta}_{k,1} = \hat{\beta}_{k,1}$. Since furthermore $\sum_{j=1}^{k} \pi_j \psi_{k-j} = \psi_k$ (see, for instance, the proof of Theorem 5.3 in Klepsch and Klüppelberg [13]), $\psi_1 = \pi_1$. Therefore,

$$\|(\hat{\theta}_{k,1} - \psi_1)(x)\| = \|(\hat{\beta}_{k,1} - \pi_1)(x)\| \xrightarrow{p} 0$$

as $n \to \infty$. This proves the statement for $i = 1$. Proceed by assuming the statement of the theorem is true for $i = 1, \ldots, N \in \mathbb{N}$, and then use induction on N. Indeed, for $i = N + 1$, the triangle inequality yields, for all $x \in H$,

$$\begin{aligned}\|(\hat{\theta}_{k,N+1} - \psi_{N+1})(x)\| &= \left\| \left(\sum_{j=1}^{N+1} \hat{\beta}_{k,j} \hat{\theta}_{k-j,N+1-j} - \pi_j \psi_{N+1-j} \right)(x) \right\| \\ &\leq \sum_{j=1}^{N+1} \|(\hat{\beta}_{k,j} - \pi_j)\hat{\theta}_{k-j,N+1-j}(x)\| + \|\pi_j(\hat{\theta}_{k-j,N+1-j} - \psi_{N+1-j})(x)\|.\end{aligned}$$

Now, for $n \to \infty$, the first summand converges in probability to 0 by part *(i)*, while the second summand converges to 0 in probability by induction. Therefore the statement is proven. □

Proof of Theorem 3.3. *(i)* The proof is based again on showing that, for all $x \in H^k$, $\|(\hat{\hat{B}}(k) - \Pi(k))(x)\| \xrightarrow{p} 0$ as $n \to \infty$, where $\hat{\hat{B}}(k) = (\hat{\hat{\beta}}_{k,1}, \ldots, \hat{\hat{\beta}}_{k,k})$. To this end, first note that

$$\|(\hat{\hat{B}}(k) - \Pi(k))(x)\| \leq \|(\hat{\hat{B}}(k) - \hat{B}(k))(x)\| + \|(\hat{B}(k) - \Pi(k))(x)\|. \tag{8.16}$$

Under Assumptions 3.1, the second term of the right-hand side converges to 0 in probability for all $x \in H^k$ by part *(i)* of Theorem 3.2. The first term of the right-hand side of (8.16) can be investigated uniformly over H^k. Using the plug-in estimators defined as in (3.13), we get for $k \in \mathbb{N}$

$$\begin{aligned}\|\hat{\hat{B}}(k) - \hat{B}(k)\|_{\mathcal{L}} &= \|\hat{\hat{\Gamma}}_{1,k,d}\hat{\hat{\Gamma}}_{k,d}^{-1} - \hat{\Gamma}_{1,k,d}\hat{\Gamma}_{k,d}^{-1}\|_{\mathcal{L}} \\ &\leq \|\big(\hat{\hat{\Gamma}}_{1,k,d} - \hat{\Gamma}_{1,k,d}\big)\hat{\hat{\Gamma}}_{k,d}^{-1}\|_{\mathcal{L}} + \|\hat{\Gamma}_{1,k,d}\big(\hat{\Gamma}_{k,d}^{-1} - \hat{\hat{\Gamma}}_{k,d}^{-1}\big)\|_{\mathcal{L}}.\end{aligned} \tag{8.17}$$

Following the same intuition as in the proof of Theorem 3.2, start by investigating the term $\|(\hat{\Gamma}_{k,d} - \hat{\hat{\Gamma}}_{k,d})\|_{\mathcal{L}}$. Applying triangle inequality, linearity of the inner product and the inequalities $\|P_{(k)}X_j(k)\| \leq \|X_j(k)\|$ and $\|\hat{P}_{(k)}X_j(k)\| \leq \|X_j(k)\|$, it follows that

$$\begin{aligned}\|(\hat{\Gamma}_{k,d} - \hat{\hat{\Gamma}}_{k,d})\|_{\mathcal{L}} &= \left\| \frac{1}{n-k} \sum_{j=k}^{n-1} \big(P_{(k)}X_j(k) \otimes P_{(k)}X_j(k) - \hat{P}_{(k)}X_j(k) \otimes \hat{P}_{(k)}X_j(k) \big) \right\|_{\mathcal{L}} \\ &\leq \frac{2}{n-k} \sum_{j=k}^{n-1} \big\|X_j(k)\big\| \big\|P_{(k)}X_j(k) - \hat{P}_{(k)}X_j(k)\big\|.\end{aligned} \tag{8.18}$$

Note that, from the definitions of $X_j(k)$, $P_{(k)}$ and $\hat{P}_{(k)}$,

$$P_{(k)}X_j(k) = \Big(\sum_{i=1}^{d_k} \langle X_j, \nu_i \rangle \nu_i, \ldots, \sum_{i=1}^{d_1} \langle X_{j-k}, \nu_i \rangle \nu_i \Big)^\top,$$
$$\hat{P}_{(k)}X_j(k) = \Big(\sum_{i=1}^{d_k} \langle X_j, \hat{\nu}_i \rangle \hat{\nu}_i, \ldots, \sum_{i=1}^{d_1} \langle X_{j-k}, \hat{\nu}_i \rangle \hat{\nu}_i \Big)^\top.$$

These relations show that

$$\begin{aligned}
\big\| P_{(k)}X_j(k) - \hat{P}_{(k)}X_j(k) \big\| &= \Big\| \Big(\sum_{i=1}^{d_k} \langle X_j, \hat{\nu}_i \rangle \hat{\nu}_i - \langle X_j, \nu_i \rangle \nu_i, \ldots, \sum_{i=1}^{d_1} \langle X_{j-k}, \hat{\nu}_i \rangle \hat{\nu}_i - \langle X_{j-k}, \nu_i \rangle \nu_i \Big)^\top \Big\| \\
&= \Big\| \Big(\sum_{i=1}^{d_k} \langle X_j, \hat{\nu}_i - \nu_i \rangle \hat{\nu}_i, \ldots, \sum_{i=1}^{d_1} \langle X_{j-k}, \hat{\nu}_i - \nu_i \rangle \hat{\nu}_i \Big)^\top \Big\| \\
&\quad + \Big\| \Big(\sum_{i=1}^{d_k} \langle X_j, \nu_i \rangle (\nu_i - \hat{\nu}_i), \ldots, \sum_{i=1}^{d_1} \langle X_{j-k}, \nu_i \rangle (\nu_i - \nu_i) \Big)^\top \Big\|.
\end{aligned}$$

Observe that, for $x = (x_1, \ldots, x_k) \in H^k$, $\|x\| = (\sum_{i=1}^k \|x_i\|^2)^{1/2}$, Then, applications of the Cauchy–Schwarz inequality and the orthonormality of $(\nu_i \colon i \in \mathbb{N})$ and $(\hat{\nu}_i \colon i \in \mathbb{N})$ lead to

$$\begin{aligned}
\big\| P_{(k)}X_j(k) - \hat{P}_{(k)}X_j(k) \big\| &\le \Big(\sum_{i=0}^{k-1} \Big\| \sum_{i=1}^{d_i} \langle X_{j-i}, \hat{\nu}_l - \nu_l \rangle \hat{\nu}_l \Big\|^2 \Big)^{1/2} + \Big(\sum_{i=0}^{k-1} \Big\| \sum_{l=1}^{d_i} \langle X_{j-i}, \nu_l \rangle (\nu_l - \hat{\nu}_l) \Big\|^2 \Big)^{1/2} \\
&\le \Big(\sum_{i=0}^{k-1} \sum_{l=1}^{d_i} \|X_{j-i}\|^2 \|\hat{\nu}_l - \nu_l\|^2 \Big)^{1/2} + \Big(\sum_{i=0}^{k-1} \sum_{l=1}^{d_i} \|X_{j-i}\|^2 \|\nu_l - \hat{\nu}_l\|^2 \Big)^{1/2} \\
&\le 2 \Big(\sum_{i=0}^{k-1} \sum_{l=1}^{d_k} \|X_{j-i}\|^2 \|\hat{\nu}_l - \nu_l\|^2 \Big)^{1/2} \\
&\le 2 \|X_j(k)\| \Big(\sum_{l=1}^{d_k} \|\hat{\nu}_l - \nu_l\|^2 \Big)^{1/2}.
\end{aligned}$$

Plugging this relation back into (8.18), it follows that

$$\|\hat{\Gamma}_{k,d} - \hat{\hat{\Gamma}}_{k,d}\|_{\mathcal{L}} \le 4 \Big(\sum_{l=1}^{d_k} \|\hat{\nu}_l - \nu_l\|^2 \Big)^{1/2} \frac{2}{n-k} \sum_{j=k}^{n-1} \|X_j(k)\|^2.$$

Since $(X_j \colon j \in \mathbb{Z})$ is L^4-m approximable, Theorems 3.1 and 3.2 in Hörmann and Kokoszka [10] imply that, for some finite positive constant C_1, $N\mathrm{E}[\|\hat{\nu}_l - \nu_l\|^2] \le C_1/\delta_l$, where δ_l is the l-th spectral gap. Hence,

$$\sum_{l=1}^{d_k} \|\hat{\nu}_l - \nu_l\|^2 \le \frac{C_1}{N} \sum_{l=1}^{d_k} \frac{1}{\alpha_l^2}.$$

Furthermore, note that

$$\frac{2}{n-k} \sum_{j=k}^{n-1} \mathrm{E}\big[\|X_j(k)\|^2\big] \le 2 \sum_{i=0}^{k-1} \mathrm{E}\big[\|X_{k-i}\|^2\big] = 2k \|C_X\|_{\mathbb{N}}.$$

Therefore, collecting the previous results yields the rate

$$\|\hat{\Gamma}_{k,d} - \hat{\hat{\Gamma}}_{k,d}\|_{\mathcal{L}} = O_p\Big(\frac{k}{n}\Big(\sum_{l=1}^{d_k}\frac{1}{\alpha_l^2}\Big)^{1/2}\Big). \tag{8.19}$$

Next, investigate $\|\hat{\hat{\Gamma}}_{k,d}^{-1}\|_{\mathcal{L}}$. Similarly as in the corresponding part of the proof of Theorem 3.2, it follows that $\|\hat{\hat{\Gamma}}_{k,d}^{-1}\|_{\mathcal{L}} \leq \|\hat{\hat{\Gamma}}_{k,d}^{-1} - \hat{\Gamma}_{k,d}^{-1}\|_{\mathcal{L}} + \|\hat{\Gamma}_{k,d}^{-1}\|_{\mathcal{L}}$. By (8.10), $\|\hat{\Gamma}_{k,d}^{-1}\|_{\mathcal{L}} = O_p(\alpha_{d_k}^{-1})$. Furthermore, the same arguments as in (8.7) and (8.8) imply that

$$\|\hat{\hat{\Gamma}}_{k,d}^{-1} - \hat{\Gamma}_{k,d}^{-1}\|_{\mathcal{L}} \leq \frac{\|\hat{\hat{\Gamma}}_{d,k} - \hat{\Gamma}_{d,k}\|_{\mathcal{L}}\|\hat{\Gamma}_{k,d}^{-1}\|_{\mathcal{L}}^2}{1 - \|\hat{\hat{\Gamma}}_{d,k} - \hat{\Gamma}_{d,k}\|_{\mathcal{L}}\|\hat{\Gamma}_{k,d}^{-1}\|_{\mathcal{L}}}. \tag{8.20}$$

Hence, by (8.10) and (8.19),

$$\|\hat{\hat{\Gamma}}_{d,k} - \hat{\Gamma}_{d,k}\|_{\mathcal{L}}\|\hat{\Gamma}_{k,d}^{-1}\|_{\mathcal{L}}^2 = O_p\Big(\frac{k}{n\alpha_{d_k}^2}\Big(\sum_{l=1}^{d_k}\frac{1}{\alpha_l^2}\Big)^{1/2}\Big).$$

Therefore, by Assumption 3.2 as $n \to \infty$, $\|\hat{\hat{\Gamma}}_{k,d}^{-1} - \hat{\Gamma}_{k,d}^{-1}\|_{\mathcal{L}} \xrightarrow{p} 0$. Taken the previous calculations together, this gives the rate

$$\|\hat{\hat{\Gamma}}_{k,d}^{-1}\|_{\mathcal{L}} = O_p\Big(\frac{1}{\alpha_{d_k}}\Big). \tag{8.21}$$

Going back to (8.17) and noticing that $\|\hat{\hat{\Gamma}}_{1,k,d} - \hat{\Gamma}_{1,k,d}\|_{\mathcal{L}} \leq \|(I_H, 0, \ldots, 0)(\hat{\hat{\Gamma}}_{k,d} - \hat{\Gamma}_{k,d})\|_{\mathcal{L}}$, the first summand in this display can be bounded by

$$\begin{aligned}
\|\big(\hat{\hat{\Gamma}}_{1,k,d} - \hat{\Gamma}_{1,k,d}\big)\hat{\hat{\Gamma}}_{k,d}^{-1}\|_{\mathcal{L}} &\leq \|\hat{\hat{\Gamma}}_{1,k,d} - \hat{\Gamma}_{1,k,d}\|_{\mathcal{L}}\|\hat{\hat{\Gamma}}_{k,d}^{-1}\|_{\mathcal{L}} \\
&\leq \|(I_H, 0, \ldots, 0)(\hat{\hat{\Gamma}}_{k,d} - \hat{\Gamma}_{k,d})\|_{\mathcal{L}}\|\hat{\hat{\Gamma}}_{k,d}^{-1}\|_{\mathcal{L}} \\
&= O_p\Big(\frac{k}{n\alpha_{d_k}}\Big(\sum_{l=1}^{d_k}\frac{1}{\alpha_l^2}\Big)^{1/2}\Big),
\end{aligned} \tag{8.22}$$

where the rate in (8.19) was used in the last step. For the second summand in (8.17), use the plug-in estimator for $\hat{\Gamma}_{1,k,d}$ to obtain, for all $k < n$,

$$\|\hat{\Gamma}_{1,k,d}\big(\hat{\Gamma}_{k,d}^{-1} - \hat{\hat{\Gamma}}_{k,d}^{-1}\big)\|_{\mathcal{L}} \leq \Big\|\frac{1}{n-k}\sum_{j=k}^{n-1} P_{(k)}X_j(k) \otimes X_{j+1}\Big\|_{\mathcal{L}}\big\|\hat{\Gamma}_{k,d}^{-1} - \hat{\hat{\Gamma}}_{k,d}^{-1}\|_{\mathcal{L}}.$$

Since

$$\begin{aligned}
\mathrm{E}\Big[\Big\|\frac{1}{n-k}\sum_{j=k}^{n-1} P_{(k)}X_j(k) \otimes X_{j+1}\Big\|_{\mathcal{L}}\Big] &\leq \frac{1}{n-k}\sum_{j=k}^{n-1}\mathrm{E}\big[\|P_{(k)}X_j(k) \otimes X_{j+1}\|_{\mathcal{L}}\big] \\
&\leq \frac{1}{n-k}\sum_{j=k}^{n-1}\big(\mathrm{E}[\|P_{(k)}X_j(k)\|^2]\big)^{1/2}\big(\mathrm{E}[\|X_{j+1}\|^2]\big)^{1/2}
\end{aligned}$$

$$= \Big(\sum_{l=0}^{k-1} \mathrm{E}[\|X_{j-l}\|^2] \Big)^{1/2} \|C_X\|_{\mathcal{N}}^{1/2}$$
$$= \sqrt{k} \|C_X\|_{\mathcal{N}},$$

the result in (8.20) implies that

$$\big\| \hat{\Gamma}_{1,k,d} \big(\hat{\Gamma}_{k,d}^{-1} - \hat{\hat{\Gamma}}_{k,d}^{-1} \big) \big\|_{\mathcal{L}} = O_p \Big(\frac{k^{3/2}}{n \alpha_{d_k}^2} \Big(\sum_{l=1}^{d_k} \frac{1}{\alpha_l^2} \Big)^{1/2} \Big). \tag{8.23}$$

Applying Assumption 3.2 to this rate and collecting the results in (8.16), (8.17), (8.22) and (8.23), shows that, for all $x \in H^k$ as $n \to \infty$, $\|(\hat{\hat{B}}(k) - \Pi(k))(x)\| \xrightarrow{p} 0$. This is the claim.

(ii) Similar to the proof of part *(ii)* of Theorem 3.6. □

References

[1] A. Aue and A. Van Delft. Testing for stationarity of functional time series in the frequency domain. *Preprint*, 2017.

[2] A. Aue, S. Hörmann, L. Horváth, and M. Reimherr. Detecting changes in the covariance structure of multivariate time series. *The Annals of Statistics*, 37:4046–4087, 2009.

[3] A. Aue, D. Dubart Norinho, and S. Hörmann. On the prediction of stationary functional time series. *Journal of the American Statistical Association*, 110:378–392, 2015.

[4] D. Bosq. *Linear Processes in Function Spaces: Theory and Applications*. Springer, New York, 2000.

[5] D. Bosq. Computing the best linear predictor in a Hilbert space. Applications to general ARMAH processes. *Journal of Multivariate Analysis*, 124:436–450, 2014.

[6] P.J. Brockwell and R.A. Davis. *Time Series: Theory and Methods (2nd Ed.)*. Springer, New York, 1991.

[7] R. Fortet. *Vecteurs, fonctions et distributions aàtoires dans les espaces de Hilbert*. Hermes, Paris, 1995.

[8] R. Gabrys and P. Kokoszka. Portmanteau test of independence for functional observations. *Journal of the American Statistical Association*, 102(480):1338–1348, 2007.

[9] S. Hörmann, L. Kidzinski, and M. Hallin. Dynamic functional principal components. *Journal of the Royal Statistical Society: Series B*, 77:319–348, 2015.

[10] S. Hörmann and P. Kokoszka. Weakly dependent functional data. *The Annals of Statistics*, 38:1845–1884, 2010.

[11] L. Horvàth and P. Kokoszka. *Inference for Functional Data with Applications*. Springer, New York, 2012.

[12] T. Hsing and R. Eubank. *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. Wiley, West Sussex, UK, 2015.

[13] J. Klepsch and C. Klüppelberg. An Innovations Algorithm for the prediction of functional linear processes. *eprint arXiv:1607.05874*, 2016.

[14] J. Klepsch, C. Klüppelberg, and T. Wei. Prediction of functional ARMA processes with an application to traffic data. *Econometrics and Statistics*, 1:128–149, 2016.

[15] P. Kokoszka and M. Reimherr. Determining the order of the functional autoregressive model. *Journal of Time Series Analysis*, 34:116–129, 2013.

[16] T.L. Lai and C.P. Lee. Information and prediction criteria for model selection in stochastic regression and ARMA models. *Statistica Sinica*, 7:285–309, 1997.

[17] R. Lewis and G.C. Reinsel. Prediction of Multivariate Time Series by Autoregressive Model Fitting. *Journal of Multivariate Analysis*, 16:393–411, 1985.

[18] F. Merlevède. Sur l'inversibilité des processus linéaires à valeurs dans un espace de Hilbert. *Comptes rendus de l'Académie des Sciences, Série I*, 321:477–480, 1995.

[19] H. Mitchell. *Topics in Multiple Time Series*. PhD thesis, Royal Melbourne Institute of Technology, 1996.

[20] H. Mitchell and P.J. Brockwell. Estimation of the coefficients of a multivariate linear filter using the Innovations Algorithm. *Journal of Time Series Analysis*, 18:157–179, 1997.

[21] S. Nsiri and R. Roy. On the invertibility of multivariate linear processes. *Journal of Time Series Analysis*, 14:305–316, 1993.

[22] V. Panaretos and S. Tavakoli. Fourier analysis of stationary time series in function space. *The Annals of Statistics*, 41:568–603, 2012.

[23] J.O. Ramsay and B.W. Silverman. *Functional Data Analysis (2nd ed.)*. Springer Series in Statistics, 2005.

[24] B. Simon. *Operator Theory — A Comprehensive Course in Analysis, Part 4*. AMS, 2015.

[25] F. Spangenberg. Strictly stationary solutions of ARMA equations in Banach spaces. *Journal of Multivariate Analysis*, 121:127–138, 2013.

[26] R.S. Tsai. *Multivariate Time Series Analysis*. Wiley, Hoboken, 2014.

[27] C. Turbillon, D. Bosq, J.M. Marion, and B. Pumo. Parameter estimation of moving averages in Hilbert spaces. *Comptes rendus de l'Académie des Sciences, Série I*, 346:347–350, 2008.