Zuse Institute Berlin

Takustr. 7
14195 Berlin
Germany

YING CHEN, THORSTEN KOCH, XIAOFEI XU

Regularized partially functional autoregressive model with application to high-resolution natural gas forecasting in Germany

ZIB Report 19-34 (October 2019)

Zuse Institute Berlin
Takustr. 7
14195 Berlin
Germany

Telephone: +49 30-84185-0
Telefax: +49 30-84185-125

E-mail: bibliothek@zib.de
URL: http://www.zib.de

ZIB-Report (Print) ISSN 1438-0064
ZIB-Report (Internet) ISSN 2192-7782

Regularized partially functional autoregressive model with application to high-resolution natural gas forecasting in Germany*

Ying Chen

Department of Mathematics, National University of Singapore,
Risk Management Institute, National University of Singapore, Singapore
Thorsten Koch
Department of Mathematics, Technische Universität Berlin,
Mathematical Optimization Department, Zuse Institute Berlin, Germany
Xiaofei Xu
Department of Statistics & Applied Probability, National University of Singapore, Singapore,
E-mail: xu.xiaofei@u.nus.edu

Abstract

We propose a partially functional autoregressive model with exogenous variables (pFAR) to describe the dynamic evolution of the serially correlated functional data. It provides a unified framework to model both the temporal dependence on multiple lagged functional covariates and the causal relation with ultrahigh-dimensional exogenous scalar covariates. Estimation is conducted under a two-layer sparsity assumption, where only a few groups and elements are supposed to be active, yet without knowing their number and location in advance. We establish asymptotic properties of the estimator and investigate its finite sample performance along with simulation studies. We demonstrate the application of pFAR with the high-resolution natural gas flows in Germany, where the pFAR model provides insightful interpretation as well as good out-of-sample forecast accuracy.

Keywords: Functional data; High dimensionality; Mixed-type covariates; Two-layer sparsity.

*Acknowledgment: The authors gratefully acknowledge the financial support of Singapore Ministry of Education Academic Research Fund Tier 1 and Institute of Data Science at National University of Singapore, and the support of the Research Campus MODAL funded by the German Federal Ministry of Education and Research (BMBF) (fund number 05M14ZAM).

1 Introduction

Functional data or quasi-functional data are available in various fields ranging from medicine to economics and finance. They usually exhibit a serial dependence across time and sometimes are collected together with high-dimensional scalar data. There has been a growing demand on a powerful statistical modeling to utilize this rich information contained in the large scale of mixed-type data to describe and predict the dynamic evolution of the dependent data in an unified framework.

A number of models have been proposed to study the dependence of functional data, such as functional linear models (Ramosay and Dalzell 1991; Yao et al. 2005; Matsui and Konishi 2011; Kong, Staicu and Maity 2016), generalized functional linear models (Müller and Stadtmüller 2005; Gertheiss et al. 2013), functional additive models (Müller and Yao 2008; Müller et al. 2013; Fan et al. 2015) and historical functional linear models (Malfait and Ramsay 2003; Harezlak et al. 2007; Brockhaus et al. 2017), where response can be either functional or scalar and covariates only contain functional variables. As for the mixed-type covariates, i.e., with both functional and scalar variables, the functional partially quantile regression model (Lu et al. 2014), the generalized functional linear model where the response is in the exponential family (Crainiceanu et al. 2009; Goldsmith et al. 2011) and the partially functional linear model (Kong, Xue, Yao and Zhang 2016) have been proposed, where the response variables are assumed to be IID.

In practice, functional data collected over time naturally involves serial dependence. Bosq (1991) pioneered the functional autoregressive (FAR) model and developed the functional Yule-Walker estimator, see also Besse et al. (2000). Bosq (2000) extended the work to the autoregressive process of order p in Hilbert space using cross-covariance operators, which triggered a vivid development of the FAR model, such as the FARX model with exogenous functional covariates (Damon and Guillas 2002; Chen et al. 2018), the convolu-

tional FAR model (Liu et al. 2016), the adaptive FAR model (Chen and Li 2017) and the varying coefficient FAR model (Xu et al. 2017).

Our interest is to incorporate all the available covariates, both functional and scalar, in a unified modeling framework to explore both the serial dependence and the casual relationship between a functional response and the large-dimensional functional and scalar covariates, which is named partially Functional AutoRegressive (pFAR) model. Direct estimation of pFAR easily leads to overfitting due to the curse of dimensionality, making out-of-sample prediction inaccurate. More importantly, an overparametrized model is hard to deliver a clear and insightful interpretation to the essential dependence of data. For example, the demand and supply of natural gas, a key energy resource for Germany and many other countries, are potentially influenced by a huge number of factors, e.g., its own past values recorded as daily curves, the cycles of working routines, the temperatures in different locations of the gas transmission network, the market prices and the supply of renewable energy resources. A high-precision modeling is required for an efficient operation of the gas transmission network, where the knowledge of who are the active covariates is valuable.

Regularized estimation provides a remedy. Instead of assuming all the covariates are relevant, it is to detect key variables that are essentially driving the dynamics of the response variable. Tibshirani (1996) proposed the least absolute shrinkage and selection operator (LASSO) for variable selection under sparse regularity; see also the adaptive LASSO (Zou 2006), the smoothly clipped absolute deviation (SCAD) (Fan and Li 2001) and the least angle regression selection (LARS) (Efron et al. 2004). Grouping structures often arise in problems with large-scale variables where a number of variables naturally forms a group with predetermined characteristics. The group LASSO and group LARS (Yuan and Lin 2006) as well as the group SCAD (Wang, Chen and Li 2007; Huang et al. 2012) select active groups under sparse regularity selections, which, however, do not yield sparsity within a

group. The sparse group LASSO (Simon et al. 2013) and the multivariate sparse group LASSO (Li et al. 2015) detect both the active groups and the active elements within the selected group simultaneously.

In our study, the sparse group LASSO penalty is adopted to detect the active lagged functional and scalar variables. A forward-looking criterion is developed to achieve variable selection at both the group and elementary levels. We establish the asymptotic consistency and sparsity for the pFAR estimator. The finite-sample performance is investigated along with simulations, which shows that both the active elements and groups are accurately selected with robust performance. Moreover, we implement the pFAR model to analyze and predict the day-ahead natural gas flow curves of various types of nodes in the high-pressure gas pipeline network in Germany. The pFAR model detects the essential and interpretable factors from seven lagged gas curves and 85 scalar covariates. It performs good forecasting accuracy in an extensive empirical experiment.

This paper makes contributions in four aspects: 1) Develop a regularized partially functional autoregressive model with mixed type of covariates. Our work differs from Kong, Xue, Yao and Zhang (2016), who considered a partial functional linear model with mixed-type covariates where the scalar response is assumed to be IID and did not study the group structure among exogenous covariates. The pFAR model, on the other hand, has a functional response with serial dependence, and the lagged functional covariates are allowed to be autocorrelated. Our model also adopts certain group structures for scalar covariates. 2) Consider the serial dependence and the impact of large-dimensional mixed-type covariates. Liu et al. (2016), Chen and Li (2017) and Xu et al. (2017) considered the functional autoregressive models with functional response and functional covariates, where the focus is on the estimation of serial dependence and the estimation of constant or time-varying operators. 3) Establish the asymptotic properties of the pFAR estimator. Damon and Guillas (2002) and Chen et al. (2018) proposed the functional autoregressive models

with multiple functional exogenous variables. In our case, given the large-dimensional exogenous scalar variables, we impose regularized estimation under a two-layer sparsity, with asymptotic consistency and sparsity. 4) Provide a powerful interpretation and accurate out-of-sample forecast for the dynamic high-resolution gas flows in Germany.

The rest of the paper is organized as follows. Section 2 details the partial FAR model and the associated regularization and penalty procedures. The asymptotic theories are developed in Section 3. Section 4 investigates the finite-sample performance with Monte Carlo experiments. Section 5 applies the proposed model to the high-resolution gas flow data in Germany. Section 6 summarizes our findings.

2 Regularized pFAR model

In this section, we present the regularized partially Functional AutoRegressive (pFAR) model and the corresponding penalized estimation procedure. In the pFAR framework, the response is the dependent functional variable and the covariates are mixed with large-dimensional scalar variables and lagged values of the functional variable.

2.1 The pFAR model

Let $\{Y_t(\tau)\}_{t=1}^n$ denote a series of n random curves taking values in a continuous domain $\tau \in [0,1]$ in the Hilbert space $\mathcal{H}$. Its p lagged curves are denoted as $Y_{t-1}(\tau), \ldots, Y_{t-p}(\tau)$. Use $\{z_{t\ell}, \ell = 1, \ldots, d\}_{t=1}^n$ to denote d exogenous scalar covariates. Without loss of generality, we assume the functional variables $Y_t(\tau)$ and the scalar variables $z_{t\ell}$ for $\ell = 1, \ldots, d$ are normalized to have zero mean. The partially functional autoregressive model is defined as:

$$Y_t(\tau) = \sum_{j=1}^{p} \int_0^1 \beta_j(\tau - s) Y_{t-j}(s) ds + \sum_{\ell=1}^{d} \gamma_\ell(\tau) z_{t-1,\ell} + \varepsilon_t(\tau) \tag{1}$$

where the serial dependence of the functional response on its own lagged values is measured by $\{\beta_j(\cdot) : j = 1, \ldots, p\}$ which are square-integral regression parameter functions in $\mathcal{H}$. The coefficient function of the scalar covariate $\gamma_\ell(\tau)$ is defined in $\mathcal{H}$, and $\varepsilon_t(\tau)$ is a strong $\mathcal{H}$-white noise with zero mean and finite second moment $E\|\varepsilon_t(\tau)\|^2 < \infty$. The norm $\|\cdot\|$ is induced from the inner product $< \cdot >$ of $\mathcal{H}$. We further assume that the serial dependence controlled by $\{\beta_1(\cdot), \ldots, \beta_p(\cdot)\}$ and the cross-dependence indicated by $\{\gamma_1(\cdot), \ldots, \gamma_d(\cdot)\}$ exhibit certain sparsity. In particular, there are certain group structures among the large number of exogenous variables and only a small number of them are nonzero. However, the number of active groups and active elements are unknown. Bosq (2000) pioneered a Hilbertian autoregressive model of order p, named ARH(p), on functional space where the serial dependence is controlled by the bounded linear AR operators from $\mathcal{H}$ to $\mathcal{H}$. In the pFAR model, the autoregressive term is a special generalization of the ARH(p) model with the AR operators being specified as the regression functions $\{\beta_j(\cdot) : j = 1, \ldots, p\}$. The incorporation of the high-dimensional exogenous variables also differentiates the pFAR model from the ARH(p), as pFAR captures both the serial dependence in the functional time series and the causal relationship of the exogenous factors.

2.2 Fourier expansion and sieves

The functional data is defined in an infinite parameter space. To enable estimation with finite sample, one idea is to decompose the infinite-dimensional process to finite parameter space with information loss controlled, such as B-splines expansion (Liu et al. 2016), functional eigenbasis expansion (Kong, Xue, Yao and Zhang 2016) and Fourier expansion (Chen and Li 2017). Without loss of generality, we consider the Fourier basis expansion motivated by the data's periodicity characteristics and computational tractability. We represent the functional terms, both functional response and functional coefficients, using the trigonometric basis functions: $\Phi_0 = I_{[0,1]}$, $\Phi_{2k}(\tau) = \sqrt{2}cos2\pi k\tau$ and $\Phi_{2k-1}(\tau) = \sqrt{2}sin2\pi k\tau$

in $L^2([0,1])$,

$$Y_t(\tau) = a_{t,0} + \sum_{k=1}^{\infty}[b_{t,k}\Phi_{2k-1}(\tau) + a_{t,k}\Phi_{2k}(\tau)], \quad \beta_j(\tau) = u_{j,0} + \sum_{k=1}^{\infty}[v_{j,k}\Phi_{2k-1}(\tau) + u_{j,k}\Phi_{2k}(\tau)],$$
$$\gamma_\ell(\tau) = h_{\ell,0} + \sum_{k=1}^{\infty}[f_{\ell,k}\Phi_{2k-1}(\tau) + h_{\ell,k}\Phi_{2k}(\tau)], \quad \varepsilon_t(\tau) = \eta_{t,0} + \sum_{k=1}^{\infty}[\omega_{t,k}\Phi_{2k-1}(\tau) + \eta_{t,k}\Phi_{2k}(\tau)],$$

where $a_{t,0}$, $a_{t,k}$, $b_{t,k}$ are the Fourier coefficients for $Y_t(\tau)$; $u_{j,0}$, $u_{j,k}$, $v_{j,k}$ are for $\beta_j(\tau)$; $h_{\ell,0}$, $h_{\ell,k}$, $f_{\ell,k}$ are for $\gamma_\ell(\tau)$, and $\eta_{t,0}$, $\eta_{t,k}$, $\omega_{t,k}$ are for the innovation $\varepsilon_t(\tau)$.

Now the functional variables and functional coefficients are characterized by the discrete Fourier coefficients. There are, however, infinite Fourier coefficients represented by $a_{t,0}$, $a_{t,k}$, $b_{t,k}$ even for the observed functional data $Y_t(\tau)$, as $k = 1, \ldots, \infty$. Estimating infinite coefficients with a finite number of samples is computationally infeasible. Grenander (1981) proposed constructing a sequence of subspaces $\{\Theta_s\}$, called sieves, which are the subspaces of the original infinite-dimensional space Θ. The sieves need to be compact and nondecreasing with $\Theta_s \subseteq \Theta_{s+1} \subseteq \cdots, \subseteq \Theta$, and the union of the subspaces $\bigcup \Theta_s$ is dense in Θ. See Grenander (1981) and Chen (2007) for more theoretical details and implementations of sieve.

Let $\{\Theta_{s_n}\}$ denote the finite-dimensional linear space of the trigonometric polynomials on $[0,1]$ of degree s_n or less, that is

$$\Theta_{s_n} = \Bigg\{ K(\tau) \in L^2 | K(\tau) = \theta_0 1_{[0,1]} + \sum_{k=1}^{s_n} \theta_k \Phi_{2k}(\tau) + \sum_{k=1}^{s_n} \vartheta_k \Phi_{2k-1}(\tau), \\ \sum_{k=1}^{s_n} k^2\theta_k^2 + k^2\vartheta_k^2 \leq c s_n, \theta_0, \theta_k, \vartheta_k \in \mathbb{R}, \tau \in [0,1] \Bigg\}, \tag{2}$$

where $\{\theta_0, \theta_k, \vartheta_k\}$ are the coefficients of Fourier expansion for functional terms, i.e. $Y_t(\tau)$, $\beta_j(\tau)$, $\gamma_\ell(\tau)$, and $\varepsilon_t(\tau)$. And c is a positive constant that makes the constraint satisfied

without sacrifice of the growth rate of s_n. Here s_n is a hyperparameter for sieve. It controls the smoothing degree and balances bias and variance of the approximation projected on the sieve space. We have $s_n \to \infty$ as $n \to \infty$, that is the number of parameters increases with the sample size. Define the number of basis/the dimensionality of Θ_{s_n} as $S = 2 * s_n + 1$.

Denote the approximation of functional terms as $Y_{t,s_n}(\tau)$, $\beta_{j,s_n}(\tau)$, $\gamma_{\ell,s_n}(\tau)$, and $\varepsilon_{t,s_n}(\tau)$ in the sieve space respectively. We carry the approximation of the functional terms over $\{\Theta_{s_n}\}$ into the pFAR model. With the orthogonality property of the complete Fourier basis, we obtain a relationship of the Fourier coefficients that follows a discrete ARX(p) process:

$$
\begin{aligned}
a_{t,0} &= \sum_{j=1}^{p} u_{j,0} a_{t-j,0} + \sum_{\ell=1}^{d} h_{\ell,0} z_{t,\ell} + \eta_{t,0} \\
a_{t,k} &= \frac{1}{\sqrt{2}} \sum_{j=1}^{p} (u_{j,k} a_{t-j,k} - v_{j,k} b_{t-j,k}) + \sum_{\ell=1}^{d} h_{\ell,k} z_{t,\ell} + \eta_{t,k}, \quad k = 1, \ldots, s_n \\
b_{t,k} &= \frac{1}{\sqrt{2}} \sum_{j=1}^{p} (u_{j,k} b_{t-j,k} + v_{j,k} a_{t-j,k}) + \sum_{\ell=1}^{d} f_{\ell,k} z_{t,\ell} + \omega_{t,k}, \quad k = 1, \ldots, s_n
\end{aligned}
\tag{3}
$$

By projecting the functional terms into $\{\Theta_{s_n}\}$, the estimation of $\beta_j(\tau)$ and $\gamma_l(\tau)$ can be solved over the finite sieve space.

Let $\boldsymbol{y}_t = (a_{t,0},\ b_{t,1},\ a_{t,1}, \ldots, b_{t,s_n},\ a_{t,s_n})' \in \mathbb{R}^S$ be the Fourier coefficients of $Y_{t,s_n}(\tau)$, and $\boldsymbol{z}_t = (z_{t,\ell}, \ldots, z_{t,d})' \in \mathbb{R}^d$ be the covariates vector, and let $\boldsymbol{e}_t = (\eta_{t,0},\ \omega_{t,1},\ \eta_{t,1}, \ldots, \eta_{t,s_n})' \in \mathbb{R}^S$ be the Fourier coefficients of $\varepsilon_{t,s_n}(\tau)$ at time t. The pFAR model under $\{\Theta_{s_n}\}$ is represented as:

$$
\boldsymbol{y}_t = \sum_{j=1}^{p} \boldsymbol{y}_{t-j} \boldsymbol{\beta}_j + \boldsymbol{z}_{t-1} \boldsymbol{\gamma} + \boldsymbol{e}_t,
$$

where $\boldsymbol{\beta}_j$ is a matrix in $\mathbb{R}^{S \times S}$ and $\boldsymbol{\gamma}$ is a matrix in $\mathbb{R}^{d \times S}$. Here, $\boldsymbol{\beta}_j$ represents the un-

known Fourier coefficients of the AR operator $\beta_{j,s_n}(\tau)$, and $\boldsymbol{\gamma}$ corresponds to the unknown Fourier coefficients of $\gamma_{\ell,s_n}(\tau)$ for $\ell = 1, \ldots, d$. For notational simplicity, we use $\boldsymbol{x}_t = (\boldsymbol{y}'_{t-1}, \cdots, \boldsymbol{y}'_{t-p}, \boldsymbol{z}'_{t-1})$ to denote all the explanatory variables, and $\mathbf{B} = [\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p, \ \boldsymbol{\gamma}]' \in \mathbb{R}^{P \times S}$ to denote the unknown Fourier coefficient matrices of the regression parameter functions. The pFAR model in matrix form will be further simplified as $\mathbf{Y} = \mathbf{XB} + \mathbf{E}$ where $\mathbf{Y} = (\boldsymbol{y}_{p+1}, \ldots, \boldsymbol{y}_n)' \in \mathbb{R}^{(n-p)\times S}$, $\mathbf{X} = (\boldsymbol{x}_{p+1}, \ldots, \boldsymbol{x}_n)' \in \mathbb{R}^{(n-p)\times S}$ and $\mathbf{E} = (\boldsymbol{e}_{p+1}, \ldots, \boldsymbol{e}_n)' \in \mathbb{R}^{(n-p)\times S}$.

We assume that the Fourier coefficients of the innovation function, denoted as $\eta_{t,0}$, $\eta_{t,k}$, $\omega_{t,k}$, are IID with zero mean, that is $\eta_{t,0} \sim \text{IID}(0, \sigma_0^2)$, $\eta_{t,k} \sim \text{IID}(0, \sigma_{\eta k}^2)$, and $\omega_{t,k} \sim \text{IID}(0, \sigma_{\omega k}^2)$ for $k = 1, \ldots, s_n$. To estimate the Fourier coefficients $\mathbf{B}$, we consider the least-square (LS) loss under sieve Θ_{s_n}:

$$\begin{aligned} L(\boldsymbol{\beta}, \boldsymbol{\gamma} | \mathcal{D}_n) &= \sum_{t=p+1}^{n} \|\mathbf{y}_t - \sum_{j=1}^{p} \boldsymbol{y}_{t-j}\boldsymbol{\beta}_j - \boldsymbol{z}_{t-1}\boldsymbol{\gamma}\|_2^2 \\ &= \sum_{t=p+1}^{n} \sum_{i=1}^{S} \{\mathbf{y}_{ti} - \mathbf{x}_t \mathbf{B}_{\cdot i}\}^2 \end{aligned}$$

where $\mathcal{D}_n = \{(Y_t(\tau), \mathbf{z}_t) : t = 1, \ldots, n\}$ are observations, and $\mathbf{B}_{\cdot i}$ denotes the i-th column of $\mathbf{B}$. Although an LS estimator can be obtained with fixed s_n, direct estimation when d and s_n are large is inefficient and lacks interpretability. It is therefore necessary to regularize the LS estimation by selecting and estimating the nonzero coefficients in $\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_p$ and $\boldsymbol{\gamma}$. Next, we represent how to perform regularization and sparse estimation under two-layer sparsity applying to both the groups and the elements within an active group.

2.3 Estimation via sparse regularization

We need to estimate the regression parameter functions and detect the essential dependence in the model. To enable estimation in the large-dimensional case, we assume sparsity in

the Fourier coefficients matrix $\mathbf{B}$. In other words, not every covariate is associated with the response. Moreover, it is useful to impose the group structures on the lagged functional and large-dimensional scalar covariates to understand the impact of a certain category. For example, for the gas data to be analyzed in Section 5, the covariates are naturally grouped to several categories by domain knowledge, i.e., prices, renewable energies and temperature. Similarly, we consider each lagged functional covariate as a group. As such, there are three kinds of sparsity:

- Group sparsity: all elements in the group are zeros;
- Element sparsity: some elements in the group are zeros;
- Null sparsity: none of the elements is zero.

Specifically, each lagged functional covariate $Y_{t-j}(\tau)$ is considered as one group for $j = 1, \ldots, p$. Thus $\boldsymbol{\beta}_j$ is considered as a group. The d scalar variables $\boldsymbol{z}_t = (z_{t,1}, \ldots, z_{t,d})$ are separated to m predetermined groups, representing, e.g., price, renewable energy and temperature. As such, the total number of groups is $m+p$. Among them, only some groups are active, while the rest are sparse with all elements being zero. Within the active group, there are some active elements, while the others are zero, leading to the element sparsity. If all the elements are nonzero, there is no sparsity, indicating that all the covariates are significant.

Let $G = \{1, 2, \ldots, m+p\}$ denote the group index set. Let $\mathcal{G}$ denote the group structure set of the functional and scalar covariates on $\mathbf{B}$, that is $\{g : \mathbf{B}_g \in \mathcal{G}\}$ for $\{g \in G\}$. We use the sparse group lasso penalty (Friedman et al. 2010; Li et al. 2015) to achieve the group and individual variable selection. The regularized pFAR model is estimated by optimizing the penalized least squares function under the two-layer sparsity:

$$\min_{B} \left\{ \frac{1}{2N} \|\mathbf{Y} - \sum_{g \in G} \mathbf{X}^{(g)} \mathbf{B}_g\|_F^2 + \lambda \sum_{g \in G} \eta_g \|\mathbf{B}_g\|_2 + \alpha \sum_{i,j} |\mathbf{B}_{i,j}| \right) \right\} \tag{4}$$

where $\|\cdot\|_F$ is the Frobenius norm, $\mathbf{B}$ is the $P \times S$ parameter matrix, $N = n - p$ and $P = pS + d$. $\mathbf{X}^{(g)}$ refers to the submatrix of $\mathbf{X}$ with the columns corresponding to the covariates in group g. $\mathbf{B}_g$ denotes the coefficient submatrix of group g, and $\mathbf{B}_{i,j}$ denotes the (i,j)-th element of $\mathbf{B}$. Specifically, the L_2 penalty term aims to shrink the coefficients of inactive groups to zero and the L_1 penalty term forces inactive entries within an active group to zero. η_g is positive weight for group g to balance size difference among groups. A default choice is usually the square root of the group size, see Huang et al. (2012) and Simon et al. (2013). The tuning parameter $\lambda \geq 0$ is for groups. When $\lambda = 0$, the penalty reduces to lasso; When λ increases, the degree of group sparsity increases and the group sparsity becomes more important. There is another tuning parameter $\alpha \geq 0$ for the individual variables. When $\alpha = 0$, the penalty becomes group lasso; when α increases, the element sparsity involves larger weight and becomes more important.

Let $\hat{\mathbf{B}}$ be the local minimizer of (4), we derive the estimation formula of the Fourier coefficients for $\beta_{g,s_n}(\tau)$ in $\mathbf{B}$, i.e. $\hat{\boldsymbol{\beta}}_g$ when other elements of $\mathbf{B}$ are fixed. Assume $\boldsymbol{\beta}_g$ belongs to group $g \in G$ for $g = 1, \ldots, p$. Denote the tuning parameter $\lambda_g := \lambda \eta_g$. Then all Fourier coefficients in group g will be zero, i.e. $\hat{\boldsymbol{\beta}}_g = \mathbf{0}$ if

$$(\langle \frac{U_{g0}}{N} \rangle_\alpha)^2 + \sum_{gk:k=1,\cdots,s_n} \{(\langle \frac{U_{gk}}{\sqrt{2}N} \rangle_{\sqrt{2}\alpha})^2 + (\langle \frac{V_{gk}}{\sqrt{2}N} \rangle_{\sqrt{2}\alpha})^2\} \leq \lambda_g^2.$$

Otherwise, $\hat{u}_{j0}$ satisfies:

$$\hat{u}_{g0} = \frac{\langle U_{g0} \rangle_{N\alpha}}{\left\|\mathbf{x}_{1+(g-1)S}\right\|_2^2 + N\lambda_g / \left\|\hat{\mathbf{B}}_g\right\|_2},$$

and for $k > 0$, $\hat{u}_{jk}$ and $\hat{v}_{jk}$ respectively satisfy:

$$\hat{u}_{gk} = \frac{\langle U_{gk} \rangle_{2N\alpha}}{\left\|\mathbf{x}_{2k+(g-1)S}\right\|_2^2 + \left\|\mathbf{x}_{2k+1+(g-1)S}\right\|_2^2 + 2N\lambda_g / \left\|\hat{\mathbf{B}}_g\right\|_2}$$

$$\hat{v}_{gk} = \frac{\langle V_{gk} \rangle_{2N\lambda}}{-\left\|\mathbf{x}_{2k+(g-1)S}\right\|_2^2 + \left\|\mathbf{x}_{2k+1+(g-1)S}\right\|_2^2 + 2N\lambda_g / \left\|\hat{\mathbf{B}}_g\right\|_2}$$

where $U_{g0} = \mathbf{x}_{1+(g-1)S}^T(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}_{(-g0)})_{.1}$, and $\hat{\mathbf{B}}_{(-g0)}$ denotes the estimated matrix $\hat{\mathbf{B}}$ with the element $\hat{u}_{g0}$ being replaced by zero. Similarly, $U_{gk} = \mathbf{x}_{2k+(g-1)S}^T(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}_{(-gk)}^{(1)})_{.2k} + \mathbf{x}_{2k+1+(g-1)S}^T(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}_{(-gk)}^{(1)})_{.2k+1}$ where $\hat{\mathbf{B}}_{(-gk)}^{(1)}$ denote the estimated matrix $\hat{\mathbf{B}}$ with the elements $\hat{u}_{gk}$ being replaced by zeros. The subscript $.2k$ refers to the $2k$-th column of the matrix. $V_{gk} = \mathbf{x}_{2k+1+(g-1)S}^T(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}_{(-gk)}^{(2)})_{.2k} - \mathbf{x}_{2k+(g-1)S}^T(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}_{(-g)}^{(2)})_{.2k+1}$ where $\hat{\mathbf{B}}_{(-gk)}^{(2)}$ denotes the estimated matrix $\hat{\mathbf{B}}$ with the element $\hat{v}_{gk}$ being replaced by zeros. And $\langle \cdot \rangle$ denotes the soft thresholding operator defined by $\langle z \rangle_\lambda = \text{sign}(z)\max\{0, |z| - \lambda\}$ for real values z and λ.

Similarly, let $b_{\ell q}$ be the (ℓ, q)-th element of $\mathbf{B}$ with $\ell = pS + 1, \ldots, P$ and $q = 1, \ldots, S$, i.e. the coefficients of $\gamma_{\ell, s_n}(\tau)$. We show the estimation formula of $b_{\ell q}$ when all the other elements in $\mathbf{B}$ are fixed. Assume that $b_{\ell q}$ belongs to a group $g \in G$, then all elements in $\mathbf{B}_g$ will be zero if:

$$\sum_{\{\ell q:\ b_{\ell q} \in \mathbf{B}_g\}} (|S_{\ell q}|/N - \alpha)^2 \leq \lambda_g^2.$$

Otherwise, $\hat{b}_{\ell q}$ satisfies:

$$\hat{b}_{\ell q} = \frac{\langle S_{\ell q} \rangle_{N\alpha}}{\|\mathbf{x}_\ell\|_2^2 + N\lambda_g / \left\|\hat{\mathbf{B}}_g\right\|_2},$$

where $S_{\ell q} = \mathbf{x}_l^T (\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}_{(-l)})_{.q}$ and $\hat{\mathbf{B}}_{(-l)}$ denotes the estimated matrix $\hat{\mathbf{B}}$ with the ℓ-th row being replaced by zeros.

The tuning parameters α and λ control the sparsity of the functional and scalar covariates, as well as the trade-off between forecast loss and penalty. Various criteria have been proposed to choose the penalty parameters. For example, Rice and Silverman (1991) and Yao et al. (2005) considered using cross-validation, Wang, Li and Tsai (2007) showed that BIC generates a consistent selection compared with generalized cross-validation, where the latter may lead to overfitting. Kong, Xue, Yao and Zhang (2016) combined BIC and AIC to select the penalty and truncation parameters.

In our study, we propose a forward-looking criterion to select the tuning parameters with the optimal out-of-sample forecast accuracy. We divide the whole sample into training, validation and testing sets and denote the intervals as T_1, T_2 and T_3, respectively. For practical reason, we focus on the observations at discrete points $\tau_s \in [0, 1]$ where $s = 1, \cdots, L$ of the entire trajectories of curve $Y_t(\tau)$, e.g. the 24 hourly gas flows of the daily curve. Let $\{Y_t(\tau_s), s = 1, \cdots, L\}$ denote the discrete points observed in the curves at time t. Given hyperparameter s_n, we choose the tuning parameters set (α^*, λ^*) by minimizing the root mean squared error over the validation period T_2:

$$(\alpha^*, \lambda^*) = \arg\min_{(\alpha,\lambda)} \left\{ \frac{1}{|T_2|L} \sum_{t \in T_2} \sum_{s=1}^{L} \sqrt{\{Y_t(\tau_s) - (\hat{Y}_{t,s_n}(\tau_s)|\alpha, \lambda)\}^2} \right\} \tag{5}$$

where $|T_2|$ is the length of interval T_2. $(\hat{Y}_{t,s_n}(\tau_s)|\alpha, \lambda)$ is the prediction of $Y_t(\tau_s)$ using the past observations up to time $t-1$ with a given pair (α, λ) under Θ_{s_n}. The fitted dynamics

with the optimal tuning parameters will be fixed and further utilized for the out-of-sample prediction in period T_3.

3 Asymptotic Properties

In this section, we derive the asymptotic properties of the pFAR estimator.

Let $\mathbf{B}_0^{\infty}$ be the true Fourier coefficients matrix of infinite expansion of the parameter functions $\beta(\tau)$ and $\gamma(\tau)$. Denote $\mathbf{B}_{0,s_n}$ as the projection of $\mathbf{B}_0^{\infty}$ on the subspace Θ_{s_n} and $\hat{\mathbf{B}}_{s_n} \in \mathbb{R}^{P\times S}$ be its sieve estimator, we will show the asymptotic properties in the $(P \times S)$-dimensional space followed by the properties in $\mathcal{H}$. For notational simplicity, we replace $\mathbf{B}_{0,s_n}$ and $\hat{\mathbf{B}}_{s_n}$ with $\mathbf{B}_0$ and $\hat{\mathbf{B}}$ respectively, and let $b_{\ell q}$ be the ℓq-th element of $\mathbf{B}$ for $\ell = pS+1, \cdots, P$ and $q = 1, \cdots, S$. That is, $b_{\ell q}$ denotes the Fourier coefficients for scalar covariates. For Fourier coefficients of $\beta_{j,s_n}(\tau)$, we keep the notation of u_{j0}, u_{jk}, and v_{jk} with $j = 1, \ldots, p$, and $k = 1, \ldots, s_n$.

Bickel et al. (2009) proved that both the prediction and the estimation error of the LASSO estimator are bounded if the true model is known for the multiple linear regression model, which is known as the "oracle property". Li et al. (2015) extended the oracle inequality properties to a multivariate linear regression framework with sparse group LASSO penalty. We follow this idea and show that the oracle inequality properties also exist for the Fourier coefficients $\hat{\mathbf{B}}$. The properties are derived with two mild conditions and one assumption holding. The conditions require that the sieve dimensionality s_n goes to infinity with a certain rate with $n \longrightarrow \infty$, and also require that the Fourier coefficients of the response curves and the scalar covariates are subGaussian processes. The assumption imposes certain restriction on the residual matrix $\boldsymbol{\delta} = \hat{\mathbf{B}} - \mathbf{B}_0$ to suffice for the main arguments of the pFAR estimator below. See details in the supplementary materials.

In particular, we investigate the properties of the sieve estimator $\hat{\mathbf{B}}$ under Θ_{s_n}. The

properties of the pFAR estimator $\hat{\beta}(\tau)$ and $\hat{\gamma}(\tau)$ in $\mathcal{H}$ will be derived afterwards. We prove that the difference of the prediction loss of Fourier coefficients of the response using the sieve estimator and theoretical estimator is bounded. We also derive the asymptotic consistency of the sieve estimator and investigate the theoretical bound of the order of sparsity.

Theorem 3.1. *Let $\boldsymbol{B}_0$ be the true unknown Fourier coefficients matrix, consider the estimator $\hat{\boldsymbol{B}}$ in (4), assume $\alpha = 2\sigma r\{log(PS)U/N\}^{1/2}$ for some constant $r > \sqrt{2}$, denote $\varsigma_g = \lambda\eta_g/\alpha$ for $g \in G$. And each column of error matrix $\boldsymbol{E} \in \mathbb{R}^{N\times S}$, i.e. $\boldsymbol{e}_k$ has $\boldsymbol{e}_k \sim N(0, \sigma_k^2\boldsymbol{I}_{N\times N})$, $\sigma = max\{\sigma_0, \sigma_1, \ldots, \sigma_S\}$. Let U be the maximum diagonal element of matrix $\psi = \frac{1}{N}\boldsymbol{X}^T\boldsymbol{X}$, and ϕ_{max} be the largest eigenvalue of ψ. Assume the conditions C1, C2, and Assumption 1 hold with $\kappa_1 = \chi_1(\pi_1, \pi_2, \varsigma)$ and $\kappa_2 = \chi_2(\pi_1, \pi_2, \varsigma)$, then with probability at least $1 - (PS)^{1-r^2/2}$, we have the following oracle bounds:*

$$\frac{1}{N}\|\boldsymbol{X}(\hat{\boldsymbol{B}} - \boldsymbol{B}_0)\|_2^2 \leq 64\frac{\sigma^2 r^2 U log(PS)}{N}\Big(\frac{\pi_1^{1/2}}{\kappa_1} + \frac{(\sum_{g\in J_2(\boldsymbol{B}_0)}\varsigma_g^2)^{1/2}}{\kappa_2}\Big)^2$$

$$|\hat{\boldsymbol{B}} - \boldsymbol{B}_0|_1 \leq 32\sigma r\sqrt{\tfrac{U log(PS)}{N}}\Big(\frac{\pi_1^{1/2}}{\kappa_1} + \frac{(\sum_{g\in J_2(\boldsymbol{B}_0)}\varsigma_g^2)^{1/2}}{\kappa_2}\Big)^2$$

And the order of sparsity, i.e. the number of selected elements is bound as

$$\Pi_1(\hat{\boldsymbol{B}}) \leq 64\phi_{max}\Big(\frac{\pi_1^{1/2}}{\kappa_1} + \frac{(\sum_{g\in J_2(\boldsymbol{B}_0)}\varsigma_g^2)^{1/2}}{\kappa_2}\Big)^2$$

Theorem 3.1 builds the asymptotic properties for the Fourier coefficients under sieve if the true model is given. It shows that the mean squared prediction error of the Fourier coefficients for the response curves $Y_t(\tau)$ is bounded by a factor of order $N^{\varrho-1}$ for some $\varrho \in (0, 1)$ which controls the growing rate of s_n. The ℓ_1 norm of the estimation error

of Fourier coefficients for the regression parameter functions $\beta_j(\tau)$ and $\gamma_l(\tau)$ is bounded by a factor of order $N^{(\varrho-1)/2}$. The order of sparsity is bounded by a constant that is related to Assumption 1. Theorem 3.1 shows the oracle inequality properties of sieve estimator $\hat{\mathbf{B}}$, while the properties are expressed in terms of Fourier coefficients. In fact, with $N, s_n \longrightarrow \infty$, we have $\|\mathbf{B}_0^{\infty} - \mathbf{B}_0\|_{L_1} \longrightarrow 0$ because $\mathbf{B}_0$ is the Fourier truncation of the true $\mathbf{B}_0^{\infty}$ on Θ_{s_n}.

Denote $\hat{\beta}_j(\tau) = \hat{u}_{j0} + \sum_{k=1}^{s_n}[\hat{v}_{jk}\Phi_{2k-1}(\tau) + \hat{u}_{jk}\Phi_{2k}(\tau)]$ and $\hat{\gamma}_l(\tau) = \hat{h}_{l0} + \sum_{k=1}^{s_n}[\hat{f}_{lk}\Phi_{2k-1}(\tau) + \hat{h}_{lk}\Phi_{2k}(\tau)]$ as the parameter estimates, where the estimated coefficients $\{\hat{u}_{j0}, \hat{u}_{jk}, \hat{v}_{jk}\}$ and $\{\hat{h}_{j0}, \hat{h}_{jk}, \hat{f}_{jk}\}$ are from $\hat{\mathbf{B}}$. Denote the forecast as $\hat{Y}_t(\tau) = \hat{a}_{t,0} + \sum_{k=1}^{s_n}[\hat{b}_{t,k}\Phi_{2k-1}(\tau) + \hat{a}_{t,k}\Phi_{2k}(\tau)]$, where the coefficients $\{\hat{a}_{j0}, \hat{a}_{jk}, \hat{b}_{jk}\}$ are computed based on (3). We then build the following asymptotic properties of pFAR estimator on $\mathcal{H}$.

Theorem 3.2. *Let $\beta_j(\tau)$ and $\gamma_\ell(\tau)$ be the true parameter functions. Under the conditions of Theorem 3.1, we have*

$$\begin{aligned}
&\tfrac{1}{N}\textstyle\sum_t \|\hat{Y}_t(\tau) - Y_t(\tau)\|_2^2 \xrightarrow{p} 0, \\
&\|\hat{\beta}_j(\tau) - \beta_j(\tau)\|_1 \xrightarrow{p} 0, \qquad j = 1, \ldots, p, \\
&\|\hat{\gamma}_\ell(\tau) - \gamma_\ell(\tau)\|_1 \xrightarrow{p} 0 \qquad \ell = 1, \ldots, d.
\end{aligned}$$

Theorem 3.2 establishes the oracle inequality properties under $\mathcal{H}$. It shows that the ℓ_1 norm of the estimation error for the regression parameter functions $\beta_j(\tau)$ and $\gamma_l(\tau)$ convergences to zero in probability. The mean square prediction loss of the pFAR estimator also convergences to zero in probability as if the true model is given.

4 Simulation Studies

We assess the finite-sample performance of pFAR in series of simulation studies. Our interests are to investigate the accuracy of the regularized least squares estimator and to check the ability of detecting the active functional and scalar covariates.

Our simulation study considers two scenarios: a moderate case with four lagged functional covariates and 20 exogenous scalar covariates; and a high-dimensional case with four lagged functional covariates and 200 scalar covariates. In both scenarios, there are group sparsity and element sparsity. The forecast accuracy is measured for both in-sample and out-of-sample. The detection of covariates is evaluated using false zero and false nonzero for sparsity in element and group.

4.1 Set up

The data are generated from the partially functional autoregressive model

$$Y_t(\tau) = \sum_{j=1}^{p} \int_0^1 \beta_j(\tau - s) Y_{t-j}(\tau) ds + \sum_{\ell=1}^{d} \gamma_\ell(\tau) z_{t-1,\ell} + \varepsilon_t(\tau),$$

where $p = 4$ functional data and $d = 20$ or 200 scalar covariates. We impose group sparsity on the functional covariates with $j = 1, 2$, meaning only the first two lags are active and $\beta_3(\tau) = \beta_4(\tau) = 0$. The scalar covariates are categorized into four groups. We let $\ell = 1, \cdots, 8$ be active, while the remaining $\gamma_l(\tau) = 0$ for $\ell = 9, \cdots, d$. In the moderate case with $d = 20$, we split the scalar covariates with equal group size of $[5, 5, 5, 5]$ respectively, indicating that the first two groups are active and there is also element sparsity in the second group with $\gamma_9 = \gamma_{10} = 0$. In the high-dimensional case with $d = 200$, the group size of scalar varies as $[20, 20, 80, 80]$ respectively, which means only one group is

active and there is also element sparsity in the group.

The scalar covariates are generated from a multivariate normal distribution $N(0, \Sigma)$ with $\Sigma_{(j,k)} = 0.5^{|j-k|}$ for any (j, k)-pair at each group. For the serial dependence, the functional parameter β_j is obtained from the Fourier expansion with basis Φ_0, $\Phi_{2k}(\tau)$ and $\Phi_{2k-1}(\tau)$ with $k \leq 2$ and $\tau \in [0, 1]$. In other words, we set $s_n = 2$. Moreover, the Fourier coefficients for β_j are randomly generated from uniform distribution with $u_{j0} \sim U[0, 1]$, $u_{jk} \sim U[-\sqrt{2}, \sqrt{2}]$ and $v_{jk} \sim U[-0.1, 0.1]$. For the functional coefficients γ_l, the Fourier coefficients h_{l0}, f_{lk} and h_{lk} are generated from uniform distribution $U[-\sqrt{2}, \sqrt{2}]$. For the innovation, we have $\eta_{t,0} \sim N(0, \sigma_0^2)$, $\eta_{t,k} \sim N(0, \sigma_{\eta k}^2)$, and $\omega_{t,k} \sim N(0, \sigma_{\omega k}^2)$ with σ_0^2, σ_k and $\sigma_{\omega k}^2$ being a random uniform $[0, \sqrt{3}/3]$ to control the signal to noise ratio.

For each scenario, we simulate 100 series of $n = 800$ functional observations. We set $T_1 = [1, 600]$ as training, $T_2 = [601, 700]$ as validation and $T_3 = [701, 800]$ as testing to perform in-sample estimation, tuning parameter selection and out-of-sample prediction respectively.

4.2 Evaluation

We calculate the mean squared error (MSE) for the estimates of $\beta_j(\tau)$ and $\gamma_\ell(\tau)$ over T_1 as:

$$\text{MSE}_\beta = \sum_{j=1}^{p} E(\| \hat{\beta}_j(\tau) - \beta_j(\tau) \|_2^2), \text{ and } \text{MSE}_\gamma = \sum_{l=1}^{d} E(\| \hat{\gamma}_\ell(\tau) - \gamma_l(\tau) \|_2^2),$$

where $\hat{\beta}_j(\tau)$ and $\hat{\gamma}_\ell(\tau)$ are constructed through the Fourier basis and the estimated Fourier coefficients. $\| \cdot \|_2^2$ is the ℓ_2 norm in $\mathcal{H}$. The accuracy of course relies on the choice of s_n. To evaluate the impact of the sieve hyperparameter, we consider a range of $s_n \in \{0, 1, 2, 3, 4, 5, 10\}$, which covers the true value $s_n = 2$.

For a practical comparison, we use the discrete observations extracted from the continuous data, denoted as $Y_t(\tau_i)$ to be taken at $L = 50$ equally spaced time sequence with $\{\tau_i \in [0,1] : i = 1, \cdots, L\}$. We compute the prediction accuracy of the estimated/forecast functional response: the in-sample MSE (inMSE) over T_1 and the one-step-ahead out-of-sample Mean Squared Forecast Error (outMSFE) over T_3 by

$$\text{inMSE} = \frac{1}{|T_1|L}\sum_{t\in T_1}\sum_{i=1}^{L}\{Y_t(\tau_i) - \hat{Y}_{t,s_n}(\tau_i)\}^2, \quad \text{outMSFE} = \frac{1}{|T_3|L}\sum_{t\in T_3}\sum_{i=1}^{L}\{Y_t(\tau_i) - \hat{Y}_{t,s_n}(\tau_i)\}^2,$$

where the prediction $\hat{Y}_{t,s_n}$ depends on α^* and λ^*, the tuning parameters selected as described in Section 2.3. For notational simplification, the hyperparameters are omitted in the notation.

To assess sparsity detection, we compute the rate of false zero (FZ) and false nonzero (FN). The false zero means a key covariate has failed to be detected in the analysis. In other words, an active variable is wrongly estimated to be zero. The false nonzero, on the other hand, means to impose spurious importance on the irrelevant covariate, wrongly identifying it as nonzero active. We discuss sparsity detection in the functional covariates, denoted as FZ_f and FN_f; scalar covariates as FZ_s and FN_s; and groups as gFZ and gFNs:

$$\begin{aligned}
\text{FZ}_f &= \frac{\sum_{j=1}^{p} I(\hat{\beta}_j(\tau) = 0 | \beta_j(\tau) \neq 0)}{\sum_{j=1}^{p} I(\beta_j(\tau) \neq 0)}, & \text{FN}_f &= \frac{\sum_{j=1}^{p} I(\hat{\beta}_j(\tau) \neq 0 | \beta_j(\tau) = 0)}{\sum_{j=1}^{p} I(\beta_j(\tau) = 0)}, \\
\text{FZ}_s &= \frac{\sum_{\ell=1}^{d} I(\hat{\gamma}_\ell(\tau) = 0 | \gamma_\ell(\tau) \neq 0)}{\sum_{\ell=1}^{d} I(\gamma_\ell(\tau) \neq 0)}, & \text{FN}_s &= \frac{\sum_{\ell=1}^{d} I(\hat{\gamma}_\ell(\tau) \neq 0 | \gamma_\ell(\tau) = 0)}{\sum_{\ell=1}^{d} I(\gamma_\ell(\tau) = 0)}, \\
\text{gFZ} &= \frac{\sum_{g=1}^{8} I(\hat{\text{B}}_g = 0 | \text{B}_g \neq 0)}{\sum_{g=1}^{8} I(\text{B}_g \neq 0)}, & \text{gFN} &= \frac{\sum_{g=1}^{8} I(\hat{\text{B}}_g \neq 0 | \text{B}_g = 0)}{\sum_{g=1}^{8} I(\text{B}_g = 0)},
\end{aligned}$$

where B_g denotes the Fourier coefficients in group g, and $\hat{\text{B}}_g$ is the estimate. $I(\cdot)$ is the indicator function. All the metrics take values between 0 and 1. The lower the value of the

metric, the better detection of the active elements/groups and sparsity structure.

4.3 Results

Table 1 summarizes the element sparsity, the group sparsity and the estimation accuracy in the two scenarios: d20_g4_m and d200_g4_h. The tuning parameters (α^*, λ^*) are selected via optimizing the forecast accuracy over the validation set T_2 via Eq. (5). Moreover, we report the performance with respect to the hyperparameter s_n ranging from 0 to 10.

The pFAR model produces a powerful and stable detection of the active elements and groups, with FZ being always zero except one case in d200_g4_h with $s_n = 0$. It indicates a successful identification of all the key features, and independent from the choice of s_n. Meanwhile, FN_s is, in general, small and varies little for the choice of s_n. Even for the high-dimensional case d200_g4_h, which consists of a large amount of inactive scalar covariates, it is around $1 \sim 2\%$, indicating an accurate detection of element sparsity. It does, however, tend to overselect group sparsity and individual functional sparsity. FN_f fluctuates around 50%. A further inspection shows that the over-selection is caused by choosing the lagged functional covariate $Y_{t-3}(\tau)$ though with small coefficients, which is possibly due to the propagation of the serial dependence. This further affects the group selection accuracy, with the functional covariate considered as group, leading to an error rate of around $22\% \sim 31\%$. Although the choice of s_n displays little influence on sparsity detection, it leads to different estimation and prediction accuracy. There is a sharp decrease in the MSE and MSFE when reaching to the true value $s_n = 2$, and keeps promising and stable afterwards. This suggests that the forward-looking criterion is appropriate for the hyperparameter choosing and a larger value of s_n has no harm.

In conclusion, the pFAR model is able to accurately select the active covariates and groups with both moderate and a large dimensional scalar covariates. The sparsity detection is stable with varying s_n. It produces accurate estimation and prediction performance

when the hyperparameter selection is larger or equal to the true value of s_n.

		Element sparsity				Group sparsity		Coefficients		Prediction	
(α^*, λ^*)	s_n	FZ_f	FNZ_f	FZ_s	FN_s	groupFZ	groupFN	MSE_β	MSE_γ	inMSE	outMSFE
d20_g4_m											
(0.025, 0.004)	0	0	0.17	0	0.02	0	0.09	4.25	18.80	8.74 (0.35)	8.73 (0.78)
(0.025, 0.004)	1	0	0.51	0	0.02	0	0.26	2.03	9.43	5.70 (0.28)	5.71 (0.66)
(0.025, 0.004)	2	0	0.51	0	0.03	0	0.26	0.00	0.13	1.14 (0.01)	1.12 (0.01)
(0.021, 0.004)	3	0	0.47	0	0.05	0	0.24	0.00	0.11	1.14 (0.01)	1.12 (0.01)
(0.021, 0.003)	4	0	0.46	0	0.05	0	0.23	0.00	0.12	1.14 (0.01)	1.12 (0.01)
(0.020, 0.003)	5	0	0.46	0	0.05	0	0.23	0.00	0.10	1.14 (0.01)	1.12 (0.01)
(0.020, 0.003)	10	0	0.44	0	0.06	0	0.22	0.00	0.12	1.14 (0.01)	1.13 (0.01)
d200_g4_h											
(0.028, 0.007)	0	0	0.19	0.01	0.02	0	0.25	2.72	15.49	11.71 (0.86)	11.85 (2.05)
(0.025, 0.004)	1	0	0.56	0	0.01	0	0.32	0.94	8.54	9.30 (1.00)	9.32 (2.32)
(0.025, 0.004)	2	0	0.66	0	0.01	0	0.35	0.00	0.11	1.24 (0.01)	1.22 (0.02)
(0.025, 0.004)	3	0	0.61	0	0.01	0	0.31	0.00	0.12	1.24 (0.01)	1.22 (0.02)
(0.025, 0.004)	4	0	0.56	0	0.01	0	0.29	0.00	0.12	1.24 (0.01)	1.22 (0.02)
(0.025, 0.004)	5	0	0.53	0	0.01	0	0.22	0.00	0.12	1.24 (0.01)	1.22 (0.02)
(0.025, 0.004)	10	0	0.45	0	0.01	0	0.23	0.00	0.15	1.25 (0.01)	1.22 (0.02)

Table 1: Performance comparison for two scenarios with $p = 4$ functional covariates, $d = 20$ (moderate d20_g4_m) and $d = 200$ (high-dimensional d200_g4_h) scalar covariates. The average value of the sparsity detection rate, estimation accuracy and the performance accuracy are reported for different choice of s_n. The standard errors of the in-sample and out-of-sample accuracy are reported in parentheses. The tuning parameters (α^*, λ^*) are selected via optimizing the forecast accuracy over the validation set T_2.

5 Real Data Analysis

We consider the natural gas flows of two years denoted as 1 October Y1 to 30 September Y3 in the German high-pressure natural gas transmission network. Details are omitted due to a non-disclosure agreement. The gas flows record both demand and supply with hourly time resolution for 24 hours, seven days a week. There are three types of nodes (locations), which exhibit different characteristics. Municipal energy suppliers (labeled M) serve residential and small commercial constituents. Power plants/industry (P) represent electricity generation and factory production nodes. Border Points (B) are large nodes with natural gas imported and exported via Germany. The question is who are the essential

driving factors for the daily gas flow curves and whether we can accurately predict the day-ahead gas flows of different types.

As an illustration, we take two nodes of each type and perform the dynamic analysis of the different functional series data. On each day, the high-resolution data are converted to daily flow curves by smoothing over the 24 hourly gas flows over each day. Figure 1 displays the flow curves of the six nodes. To provide an interpretable comparison among different types, the gas flows are normalized for each node to have zero mean and unit variance. There is stronger weekly seasonality in nodes M driven by the working routines of households. Spikes are observed in nodes P due to emergent demand on electricity generation using gas. Nodes B are generally stable and less sensitive to seasonal shift. Though differing in trends and seasonality, there is always significant serial dependence across all the nodes. The sample autocorrelations and cross-correlations, showing significant values, are included in the supplementary material.

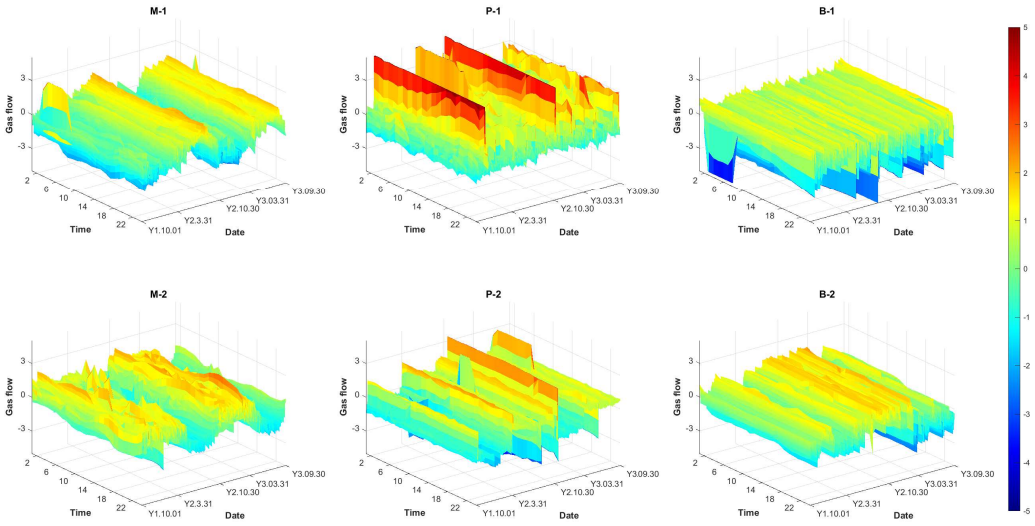

Figure 1: Daily gas flow curves at six representative nodes from 1 October Y1 to 30 September Y3.

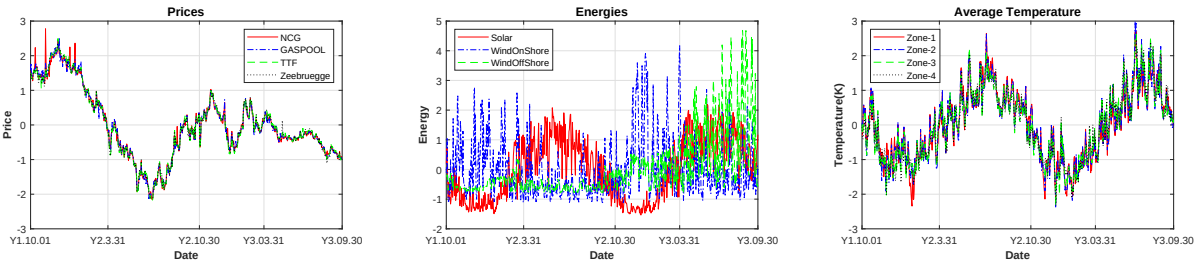

Figure 2: Exogenous scalar variables: four prices, i.e., NCG, GASPOOL, TTF and Zeebruegge; three renewable energies, i.e., solar, wind-onshore and wind-offshore and average temperature of four zones from 1 October Y1 to 30 September Y3.

We are also provided a large number of exogenous variables, namely 85 environmental and economic variables, including four market price values, i.e., NCG, GASPOOL, TTF and Zeebruegge, in different networks, three renewable energy variables, i.e., solar, wind-onshore and wind-offshore, and daily temperatures in 78 locations. The 78 locations are further split into four geographic zones and each zone is considered as a group. Since the variables vary in scale, the maximum price value of TTF, for example, is 2.90 and the maximum value of wind-on-shore is $30,158$, the exogenous scalars are also normalized. Figure 2 presents the normalized observations of the 85 scalar covariates, which shows the similar dynamic features within a group.

We divide the sample into three phases with 365 days from 1 October Y1 to 30 September Y2 as the training (labeled T_1), 183 days from 1 October Y2 to 31 March Y3 as validation (T_2), and 182 days dated on 1 April Y3 to 30 September Y3 as forecasting period (T_3). We find $s_n = 1$ and larger values deliver similar results. Thus, we fix $s_n = 1$ motivated by our limited experience in the simulation study. The tuning parameters (α^*, λ^*) are selected and fixed as before over T_2. At each forecast origin, forecasting is conducted by the updated model with available data using selected covariates. Specifically, we obtain

the h-day-ahead forecasts for each hour τ_s using iterative approach:

$$\hat{Y}_{t+h}(\tau_s) = \sum_{j\in J}\int_0^1 \hat{\beta}_j(\tau_s - u)Y_{t-j+1}(u)du + \sum_{\ell\in D}\hat{\gamma}_\ell(\tau_s)z_{t,\ell}, \quad s = 1,\ldots,24,$$

where $J \subset \{1,\cdots,7\}$ and $D \subset \{1,\cdots,85\}$ represent the sets of selected functional and scalar covariates respectively. We consider h=1, 7 and 14, which covers 1-day, 1-week and 2-weeks-ahead forecasts.

5.1 Essential factors

In general, the fitted model shows that gas flows of municipal energy suppliers are more influenced by weekly seasonality, price and temperature, while flows in power plants/industry and border nodes are driven by serial dependence. Table 2 reports the estimated Fourier coefficients for the lagged curves (autocorrelation) and those for prices and renewable energy. Here only the first Fourier coefficients are provided, as the rest, in most cases, are small. The complete results can be found in the supplementary material. We classify the dependence into three categories: "weak" when the absolute coefficient falls in interval (0,0.05), "semi-strong" if it is in [0.05, 0.1], and "strong" otherwise. It shows that the serial dependence of gas flows is dominated by the lag-1 gas flow curves, and this is stable for various type of nodes. As expected, there is a stronger weekly dependence for nodes M, with lag 7 in general larger than for nodes of type P and B. Surprisingly, the price effects are weak. Renewable energy has more effect on nodes M than P and B. For instance, solar has a semi-strong and negative effect at M-1, while it either disappears or becomes weak for P and B.

Figure 3 visualizes the effects of temperature associated with geographical zones. In the heatmap, columns represent gas nodes and rows are locations of temperature. The estimated first Fourier coefficients are colored in blue (red) if negative (positive). In general,

nodes M are more sensitive to temperature than P and B. Specifically, M nodes are almost independent to temperature in zone-2 and related to other zones, while the effects are generally weak and mixed, although M nodes are closer to zone-3. It indicates that, given the other functional and exogenous variables impact, the temperature effects are mixed among locations, and the closer positions are not necessary to have stronger effects for M nodes. Nodes P and B are almost independent from the variation in temperature, which is understandable given their functions.

covariates	M-1	M-2	P-1	P-2	B-1	B-2
lag 1	0.78	0.52	0.74	0.80	0.43	0.76
lag 2	-0.07	-0.07	-	-	-	-
lag 3	-	0.03	0.03	-	-	-
lag 4	-	0.06	-	-	-	-
lag 5	0.05	0.05	-	0.03	0.07	0.02
lag 6	-	0.06	-	-	-	0.10
lag 7	0.07	0.02	0.03	0.03	-	0.02
NCG	0.03	0.02	-	0.01	-	-
GASPOOL	0.02	0.04	-	-	-	-
TTF	0.01	0.01	-	-	-0.03	-
Zeebruegge	0.02	0.03	-	0.01	-0.03	0.01
Solar	-0.05	-0.02	-	-0.03	-	-
Wind-onshore	0.01	0.09	-	-	-	-0.01
Wind-offshore	0.01	0.04	-	-	-	-0.01

Table 2: The first estimated Fourier coefficients of parameter functions corresponding to the lagged gas flows, price variables and energy variables for three types of gas nodes, where the symbol "-" represents 0.

5.2 Forecast

5.2.1 Evaluation

We measure forecast accuracy in terms of level and direction as well as the out-of-sample goodness-of-fit. For the feasibility of comparison, we convert back to the original gas flow data when computing the forecast accuracy. Specifically, we compute the daily mean

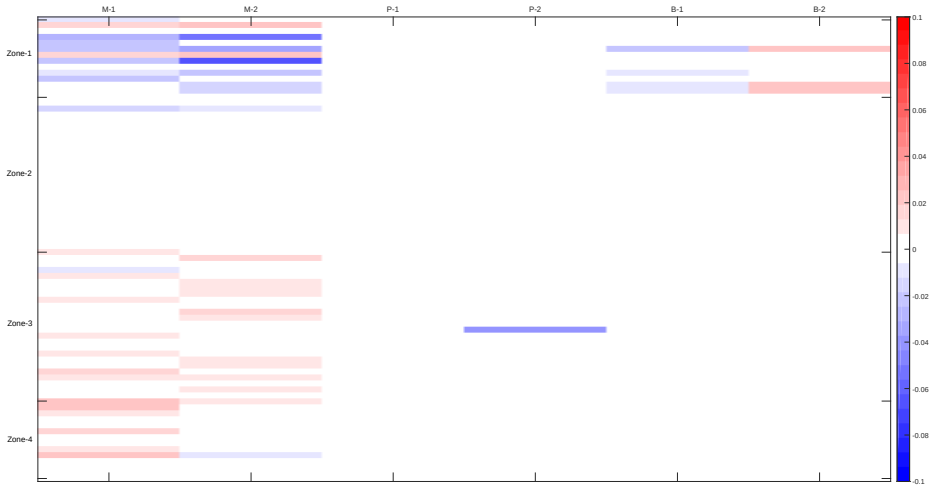

Figure 3: Heatmap of estimated effects (the first estimated Fourier coefficients) of temperature on gas where the rows are gas nodes and columns are zones with blue (red) indicating negative (positive) effects.

absolute percentage error (MAPE), the mean sign correctness proportion (MSCP) and the out-of-sample R^2 (OoR^2) for h-step-ahead forecast as:

$$\text{MAPE}_h = \frac{1}{24 \cdot |T_3|} \sum_{s=1}^{24} \sum_{t \in T_3} \left| \frac{Y_{t+h}(\tau_s) - \hat{Y}_{t+h}(\tau_s)}{Y_{t+h}(\tau_s)} \right|,$$

$$\text{MSCP}_h = \frac{1}{24 \cdot |T_3|} \sum_{s=1}^{24} \sum_{t \in T_3} I[(Y_{t+h}(\tau_s) - Y_{t+h}(\tau_{s-1}))(\hat{Y}_{t+h}(\tau_s) - \hat{Y}_{t+h}(\tau_{s-1})) > 0],$$

$$\text{Oo}R^2_h = 1 - \frac{\sum_{s=1}^{24} \sum_{t \in T_3} (Y_{t+h}(\tau_s) - \hat{Y}_{t+h}(\tau_s))^2}{\sum_{s=1}^{24} \sum_{t \in T_3} (Y_{t+h}(\tau_s) - \bar{Y}_{t+h}(\tau_s))^2},$$

where $|T_3| = 182$ is the forecast period length. $\bar{Y}_{t+h}(\tau_s)$ is the historical average of the hourly flows up to time t, and we set $Y_t(\tau_0) = Y_{t-1}(\tau_{24})$.

We also consider several alternative models as comparison. The AR type models include AR(1), AR(7) and ARX, which estimate and forecast the gas flows at each hour independently. The ARX model incorporates all the exogenous variables and LASSO is implemented for variable selection. The FAR type models, including FAR(1) and FAR(7),

consider the serial- and cross-dependence of the high-resolution gas flow data, but without taking exogenous covariates into account. See supplementary materials for the models' specific formula.

For comparison, we use the pFAR model as benchmark, and compute the relative MAPE of an alternative model i by:

$$\text{rMAPE}_h = \frac{\text{MAPE}_h(\text{model } i)}{\text{MAPE}_h(\text{pFAR})} - 1,$$

where a positive (negative) rMAPE_h represents better (worse) performance of pFAR than the alternative model i, and the amount produces the percentage change against the benchmark. The significance of accuracy difference between pFAR model and the alternative is measured by the Diebold and Mariano (DM) test (Diebold and Mariano 2002). Let e^h_{1t} and e^h_{2t} denote the h-day-ahead forecast error of the two models respectively, y_t is the observed response at time t, and $d^h_t = |e^h_{1t}/y_t| - |e^h_{2t}/y_t|$ denotes the absolute percentage loss-differential. The DM test statistic is defined:

$$\text{DM}_h = \frac{\frac{1}{n}\sum_{t=1}^{n} d^h_t}{\sqrt{(\hat{\sigma}_0 + 2\sum_{k=1}^{h-1}\hat{\sigma}_k)/n}},$$

where $\hat{\sigma}_0$ is the sample standard deviation and $\hat{\sigma}_k$ is the autocovariance at lag k of the loss differential series d^h_t with $k \geq 1$. The null hypothesis is $\text{H}_0 : \text{DM}_h = 0$ which means there is no significant difference between two h-step-ahead forecasts.

5.2.2 Result

Figure 4 displays the 1-day-ahead forecasts of the hourly gas flows over T_3 from 1 April Y3 to 30 September Y3. Although the dynamics of gas flows differ for node type, the pFAR model is able to successfully capture the evolution of gas flows with different features, and

produce a powerful and stable performance. Table 3 details MSCP, OoR^2 and MAPE of the pFAR model, as well as rMAPE of the alternative models for 1−, 7− and 14−day-ahead forecasting horizon. The best forecast model is marked in bold. Asterisk highlights statistically significant outperformance of pFAR at 10% level based on the DM test.

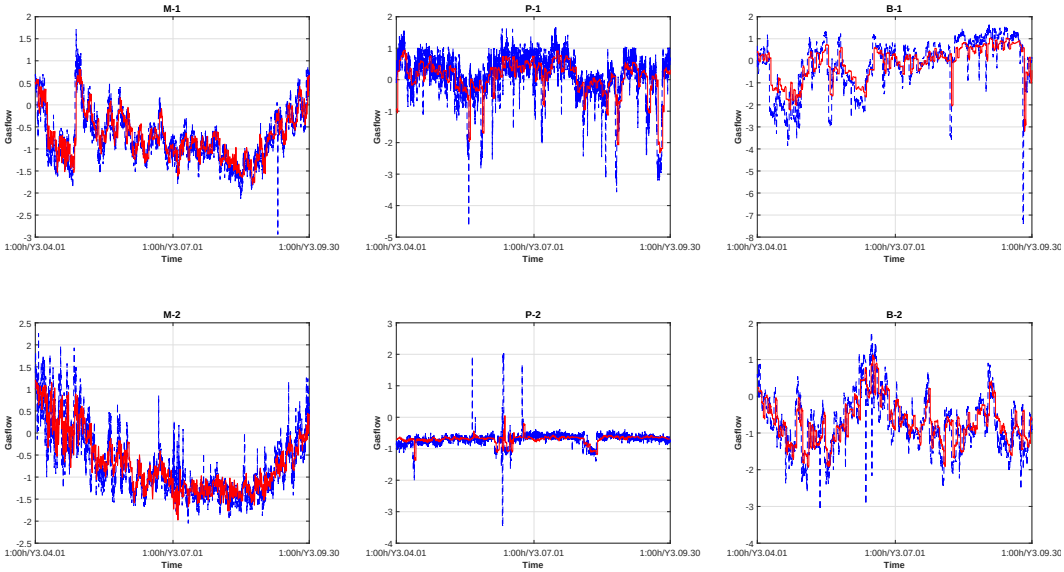

Figure 4: The hourly gas flow observations (dashed lines in blue) and 1-step-ahead forecast (solid lines in red) at six nodes from 1:00 hour at 1 April Y3 to 24:00 hour at 30 September Y3. The data are normalized to display.

pFAR produces good forecast accuracy with MAPE in a range of 0.043 (P-2) and 0.165 (M-2) for 1-day-ahead forecast. The accuracy naturally decreases with the forecasting horizon h. The out-of-sample R^2 is promising for nodes M, with above 88.7% and 91.50% for one-day-ahead forecast, and 70.75% and 80.47% for 2-weeks ahead. The direction prediction is reasonable and stable terms of in MSCP. For example, the pFAR predicts the moving direction of gas flows well with MSCP being between 71.18% and 74.06% for 1- to 14- day-ahead forecast at nodes M, and larger than 50% for nodes P and B.

Moreover, pFAR significantly outperforms all the alternative models and horizons in terms of rMAPE for nodes M, and the relative performance further improves when h

increases. According to rMAPE, pFAR improves 11.2% and 18.4% against ARX, 15.5% and 34.5% against AR(1) for the two M nodes at one-day-ahead forecast, the improvement rates become 13.8% and 21.3% against ARX and 26.0% and 35.3% against AR(1) when h=14. This is due to the poor accuracy generated by the alternative. We observe similar outperformance for nodes of type B, but with less improvement. The results are however mixed for nodes of type P. For instance, at $h = 1$ and 7, pFAR shows better accuracy than the AR type models with a significant improvement between 0.9% $\sim$ 43.7%, and outperforms FAR models with improvement up to 7.2%, while there is a reduction of 1.6% $\sim$ 3.8% in P-1 for 2-week-ahead forecast compared with all the alternative models. Moreover, it attracts attention of the poor performance in P-1, where OoR^2 is relatively low, varying from 32.28% to 0.21% among three forecast horizons and MAPE performance is also poor for h=14. A further inspection shows that the spikes in P-1 cause overfitting. Specifically, it turns out that all the prices, renewable energies and some temperatures are selected as active covariates for h=14, which is in contrast to the case for $h = 1$. Although over-selection makes for a nice in-sample estimation, it usually leads to poor out-of-sample accuracy.

In summary, pFAR in general delivers good and stable out-of-sample forecast accuracy across different types of nodes with various features and forecast horizons. The general forecasting excellence benefits from the consideration of cross-dependence of multiple stochastic processes, and the detection and incorporation of essential functional and scalar covariates together.

6 Conclusion

We propose a partially functional autoregressive model (pFAR) to investigate the dynamics of the functional response, depending on the lagged functional covariates and the causal

	Model	M-1	M-2	P-1	P-2	B-1	B-2
h=1							
	AR(1)	+15.5%*	+34.5%*	+10.6%*	+31.3%*	+5.4%*	+8.8%*
	AR(7)	+7.8%*	+5.4%*	+9.7%*	+21.2%*	+0.6%*	+4.0%*
rMAPE	ARX	+11.2%*	+18.4%*	+10.3%*	+33.4%*	+7.0%*	+8.8%*
	FAR(1)	+6.1%*	+12.7%*	0	0	+2.4%*	+1.5%*
	FAR(7)	+1.4%	+1.6%*	+0.4%*	+2.4%*	**-1.1%***	**-1.5%***
MAPE		**0.088**	**0.165**	**0.103**	**0.043**	0.093	0.127
OoR^2	pFAR	91.50%	88.71%	32.28%	90.03%	49.70%	75.78%
MSCP		74.06%	72.08%	51.40%	53.41%	52.85%	53.57%
h=7							
	AR(1)	+13.3%*	+33.6%*	+0.9%*	+43.7%*	+7.6%*	+16.7%*
	AR(7)	+9.2%*	+9.6%*	+1.6%*	+17.1%*	+4.1%*	+11.6%*
rMAPE	ARX	+7.3%*	+22.0%*	+1.2%*	+38.4%*	+6.0%*	+9.7%*
	FAR(1)	+8.0%*	+13.3%*	0	+20.4%*	+7.2%*	+12.3%*
	FAR(7)	+8.4%*	+7.6%*	+0.6%*	+7.2%*	+3.7%	+9.0%*
MAPE		**0.152**	**0.228**	**0.128**	**0.064**	**0.138**	**0.172**
OoR^2	pFAR	73.61%	80.59%	5.98%	86.87%	16.29%	61.05%
MSCP		73.43%	71.83%	52.19%	51.98%	50.40%	53.57%
h=14							
	AR(1)	+26.0%*	+35.3%*	-3.4%	+29.3%*	+6.7%*	+16.5%*
	AR(7)	+20.5%*	+16.8%*	-2.4%	-1.6%	+7.0%*	+15.0%*
rMAPE	ARX	+13.8%*	+21.3%*	−3.0%*	+21.4%*	+6.6%*	+6.6%*
	FAR(1)	+23.8%*	+21.7%*	**-3.8%***	+14.2%*	+6.5%*	+11.4%*
	FAR(7)	+21.0%*	+16.7%*	-3.4%	**-4.6%**	+6.9%*	+11.2%*
MAPE		**0.169**	**0.247**	0.140	0.081	**0.136**	**0.206**
OoR^2	pFAR	70.75%	80.47%	0.21%	83.03%	12.86%	43.48%
MSCP		72.92%	71.18%	52.26%	52.98%	51.91%	53.05%

Table 3: Forecast accuracy: Out-of-sample R^2, MSCP, MAPE of the pFAR model and the relative MAPE of the alternative models for 1−, 7− and 14−day-ahead forecasts of gas flows. The best performing model is marked in bold. The asterisk represents significant forecast accuracy difference compared to the pFAR model according to the DW test. The positive (negative) value of rMAPE represents better (worse) performance of pFAR, and the amount produces the percentage change of alternative against pFAR.

relation with the high-dimensional exogenous scalar covariates. Under a two-layer sparsity imposed to group and individual covariates, a regularized least square estimation is derived to detect the essential lagged functional and scalar variables, and achieves variable selection at both the group and elementary levels. A forward-looking criterion is proposed to tune the penalty parameters. Furthermore, we develop the asymptotic oracle inequality properties for the pFAR estimator and investigate its finite sample performance under both moderate and high-dimensional scenarios. The pFAR model has found to be able to select the

active covariates and groups accurately, and to provide good estimation and prediction performance.

We applied the pFAR model to analyze the natural gas flows in the high-pressure gas pipeline network in Germany. We found the lag-1 daily gas flow curve dominates the serial dependence. The municipal energy suppliers nodes are more sensitive to the weekly seasonality, price, renewable energy and temperature, while the power plants/industry nodes and the border nodes are mainly driven by the serial dependence. Moreover, the pFAR delivers good and stable out-of-sample forecast accuracy, with superior relative performance than alternative models across different samples and forecast horizons, by taking the cross-dependence in the functional time series and the causal dependence driven by mixed type exogenous covariates into consideration.

SUPPLEMENTARY MATERIAL

The supplementary materials to this paper contain a set of supplemental figures and tables, the derivation procedures of parameter estimation and mathematical proofs of theorems.

References

Besse, P. C., Cardot, H., and Stephenson, D. B. (2000), "Autoregressive Forecasting of Some Functional Climatic Variations," *Scandinavian Journal of Statistics*, 27, 673–687.

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009), "Simultaneous Analysis of LASSO and Dantzig Selector," *The Annals of Statistics*, 37, 1705–1732.

Bosq, D. (1991), "Modelization, Nonparametric Estimation and Prediction for Continuous

Time Processes," in *Nonparametric functional estimation and related topics*, Vol. 335 Dordrecht: Springer, pp. 509–529.

Bosq, D. (2000), *Linear Processes in Function Spaces: Theory and Applications*, Vol. 149 New York: Springer.

Brockhaus, S., Melcher, M., Leisch, F., and Greven, S. (2017), "Boosting Flexible Functional Regression Models with a High Number of Functional Historical Effects," *Statistics and Computing*, 27, 913–926.

Chen, X. (2007), "Large Sample Sieve Estimation of Semi-Nonparametric Models," *Handbook of Econometrics*, 6, 5549–5632.

Chen, Y., Chua, W. S., and Koch, T. (2018), "Forecasting Day-Ahead High-Resolution Natural-Gas Demand and Supply in Germany," *Applied Energy*, 228, 1091–1110.

Chen, Y., and Li, B. (2017), "An Adaptive Functional Autoregressive Forecast Model to Predict Electricity Price Curves," *Journal of Business & Economic Statistics*, 35, 371–388.

Crainiceanu, C. M., Staicu, A.-M., and Di, C.-Z. (2009), "Generalized Multilevel Functional Regression," *Journal of the American Statistical Association*, 104, 1550–1561.

Damon, J., and Guillas, S. (2002), "The Inclusion of Exogenous Variables in Functional Autoregressive Ozone Forecasting," *Environmetrics*, 13, 759–774.

Diebold, F. X., and Mariano, R. S. (2002), "Comparing Predictive Accuracy," *Journal of Business & Economic Statistics*, 20, 134–144.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), "Least Angle Regression," *The Annals of Statistics*, 32, 407–499.

Fan, J., and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle properties," *Journal of the American Statistical Association*, 96, 1348–1360.

Fan, Y., James, G. M., and Radchenko, P. (2015), "Functional Additive Regression," *The Annals of Statistics*, 43, 2296–2325.

Friedman, J., Hastie, T., and Tibshirani, R. (2010), "A Note on the Group LASSO and a Sparse Group LASSO," *arXiv:1001.0736*, .

Gertheiss, J., Maity, A., and Staicu, A.-M. (2013), "Variable Selection in Generalized Functional Linear Models," *Statistics*, 2, 86–101.

Goldsmith, J., Bobb, J., Crainiceanu, C. M., Caffo, B., and Reich, D. (2011), "Penalized Functional Regression," *Journal of Computational and Graphical Statistics*, 20, 830–851.

Grenander, U. (1981), *Abstract Inference* New York: Wiley.

Harezlak, J., Coull, B. A., Laird, N. M., Magari, S. R., and Christiani, D. C. (2007), "Penalized Solutions to Functional Regression Problems," *Computational Statistics & Data Analysis*, 51, 4911–4925.

Huang, J., Breheny, P., and Ma, S. (2012), "A Selective Review of Group Selection in High-dimensional Models," *Statistical Science*, 27, 481–499.

Kong, D., Staicu, A.-M., and Maity, A. (2016), "Classical Testing in Functional Linear Models," *Journal of Nonparametric Statistics*, 28, 813–838.

Kong, D., Xue, K., Yao, F., and Zhang, H. H. (2016), "Partially Functional Linear Regression in High Dimensions," *Biometrika*, 103, 147–159.

Li, Y., Nan, B., and Zhu, J. (2015), "Multivariate Sparse Group Lasso for the Multivariate Multiple Linear Regression with an Arbitrary Group Structure," *Biometrics*, 71, 354–363.

Liu, X., Xiao, H., and Chen, R. (2016), "Convolutional Autoregressive Models for Functional Time Series," *Journal of Econometrics*, 194, 263–282.

Lu, Y., Du, J., and Sun, Z. (2014), "Functional Partially Linear Quantile Regression Model," *Metrika*, 77, 317–332.

Malfait, N., and Ramsay, J. O. (2003), "The Historical Functional Linear Model," *Canadian Journal of Statistics*, 31, 115–128.

Matsui, H., and Konishi, S. (2011), "Variable Selection for Functional Regression Models via the L1 Regularization," *Computational Statistics & Data Analysis*, 55, 3304–3310.

Müller, H.-G., and Stadtmüller, U. (2005), "Generalized Functional Linear Models," *The Annals of Statistics*, 33, 774–805.

Müller, H.-G., Wu, Y., and Yao, F. (2013), "Continuously Additive Models for Nonlinear Functional Regression," *Biometrika*, 100, 607–622.

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

Ramosay, J., and Dalzell, C. (1991), "Some Tools for Functional Data Analysis (with discussion)," *Journal of the Royal Statistical Society, Series B*, 53, 539–572.

Rice, J. A., and Silverman, B. W. (1991), "Estimating the Mean and Covariance Structure Nonparametrically When the Data are Curves," *Journal of the Royal Statistical Society, Series B*, 53, 233–243.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013), "A Sparse-Group LASSO," *Journal of Computational and Graphical Statistics*, 22, 231–245.

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the LASSO," *Journal of the Royal Statistical Society, Series B*, 58, 267–288.

Wang, H., Li, R., and Tsai, C.-L. (2007), "Tuning Parameter Selectors for the Smoothly Clipped Absolute Deviation Method," *Biometrika*, 94, 553–568.

Wang, L. f., Chen, G., and Li, H. (2007), "Group SCAD Regression Analysis for Microarray Time Course Gene Expression Data," *Bioinformatics*, 23, 1486–1494.

Xu, M., Li, J., and Chen, Y. (2017), "Varying Coefficient Functional Autoregressive Model with Application to the US Treasuries," *Journal of Multivariate Analysis*, 159, 168–183.

Yao, F., Müller, H.-G., and Wang, J.-L. (2005), "Functional Linear Regression Analysis for Longitudinal Data," *The Annals of Statistics*, 33, 2873–2903.

Yuan, M., and Lin, Y. (2006), "Model Selection and Estimation in Regression with Grouped Variables," *Journal of the Royal Statistical Society, Series B*, 68, 49–67.

Zou, H. (2006), "The Adaptive Lasso and Its Oracle Properties," *Journal of the American Statistical Association*, 101, 1418–1429.