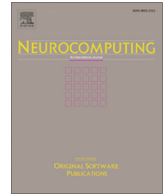




Contents lists available at ScienceDirect

Neurocomputing

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

Deep time series models for scarce data

Qiyao Wang^{*}, Ahmed Farahat, Chetan Gupta, Shuai Zheng

Industrial AI Laboratory, Hitachi America, Ltd. R&D, Santa Clara, CA, USA

ARTICLE INFO

Article history:

Received 16 April 2020

Revised 26 September 2020

Accepted 10 December 2020

Available online xxxx

Keywords:

Time series analysis

Scarce data

Deep learning models

Functional data analysis

ABSTRACT

Time series data have grown at an explosive rate in numerous domains and have stimulated a surge of time series modeling research. A comprehensive comparison of different time series models, for a considered data analytics task, provides useful guidance on model selection for data analytics practitioners. Data scarcity is a universal issue that occurs in a vast range of data analytics problems, due to the high costs associated with collecting, generating, and labeling data as well as some data quality issues such as missing data. In this paper, we focus on the temporal classification/regression problem that attempts to build a mathematical mapping from multivariate time series inputs to a discrete class label or a real-valued response variable. For this specific problem, we identify two types of scarce data: scarce data with small samples and scarce data with sparsely and irregularly observed time series covariates. Observing that all existing works are incapable of utilizing the sparse time series inputs for proper modeling building, we propose a model called sparse functional multilayer perceptron (SFMLP) for handling the sparsity in the time series covariates. The effectiveness of the proposed SFMLP under each of the two types of data scarcity, in comparison with the conventional deep sequential learning models (e.g., Recurrent Neural Network, and Long Short-Term Memory), is investigated through mathematical arguments and numerical experiments.

© 2021 Published by Elsevier B.V.

1. Introduction

Nowadays, time series data that consist of repeated data measurements over a bounded time range have become ubiquitous in numerous domains, such as meteorology, epidemiology, transportation, agriculture, industry, bioinformatics, and the world wide web. Time series data often have intrinsic temporal structures such as auto-correlation, trend, and seasonality. For example, for industrial equipment, the sensor measurements over the lifespan are correlated. Also, the sensor time series often gradually increase/decrease due to performance degradation. When modeling time series data, it is of paramount importance to leverage the internal temporal information to achieve good data analytical performance.

In the literature, three types of time series approaches have been widely explored, including (1) the classical time series models (e.g., the Auto-Regressive Integrated Moving Average and the autoregressive exogenous models) [3], (2) the sequential deep learning models such as the Recurrent Neural Network (RNN), the Long Short-Term Memory (LSTM), and the Gated Recurrent

Unit (GRU) in the machine learning community [4–6], and the (3) emerging functional data analysis in the statistical field [7]. Fundamentally, the classical time series models exploit the autoregressive (AR) and moving average (MA) techniques to encode the dependency of later observations on the prior data and the regression error at previous timestamps. The deep sequential learning models use hidden states to hold the up-to-present memory and recurrently conduct the same transformation on the internal memory and input data along the timestamps to sequentially process the temporal information. Rather than considering time series as a sequence of scalar-valued observations, functional data analysis (FDA) models treat the observed time series data as discrete realizations given rise by a continuous underlying random function of time (i.e., random curve) [7] and directly analyze a sample of such finitely evaluated random functions. This functional setting naturally accounts for temporal information. There is substantial literature on modeling and estimation for functional data, including functional principal component analysis [8,9], regression with functional responses, functional predictors or both [10–13], functional classification and clustering [14,15], and functional quantile analysis [16–18]. Ramsay [7] offers a comprehensive perspective of FDA methods.

The aforementioned time series analytic modeling techniques handle temporal information from different perspectives. Corre-

^{*} Corresponding author.

E-mail addresses: Qiyao.Wang@hal.hitachi.com (Q. Wang), Ahmed.Farahat@hal.hitachi.com (A. Farahat), Chetan.Gupta@hal.hitachi.com (C. Gupta), Shuai.Zheng@hal.hitachi.com (S. Zheng).

spondingly, they impose varying requirements on data and achieve different performances under diverse scenarios. A comprehensive comparison of time series models, for a considered data analytics task, provides useful guidance on model selection for data analytics practitioners. Some papers [19,20] conducted comparative studies of RNN and ARIMA for the time series forecasting task, but the literature still lacks a comprehensive comparison between FDA and these approaches.

In this paper, we focus on the crucial problem of building supervised learning models in time series analysis. In particular, the goal is to build a mathematical mapping from one or several real-valued time series within a bounded period to a discrete class label or a real-valued response variable. Note that we consider the general case where the response variable is different from the temporal covariates. The classical time series models such as ARIMA and the autoregressive exogenous models are infeasible for the considered temporal classification/regression problem, as they are incapable of building mappings for time series to a heterogeneous response. In the other two categories of time series models, several methods are feasible to solve the considered problem, of which the most important two are the sequential learning models [4,21,6], and the Functional Multilayer Perceptron (FMLP), a counterpart of the classical MLP for continuous random functions over a continuum such as time series [22,23,2].

Due to the nature of deep learning models, the sequential and functional deep learning models can both be used to train an end-to-end temporal classification/regression model with reasonable generalizability if we have access to a large number of training samples that cover the diverse variability in data. However, due to the high cost associated with collecting, labeling, storing, processing, and modeling a large amount of training data, building effective deep learning models with a limited amount of samples (i.e., scarce data) is an appealing and meaningful topic in the time series analysis field. In particular to the considered problem, we identify two types of scarce data which can be described as follows. The ideal case is that the number of samples N is sufficiently large and each time series input is densely and regularly observed time series, as shown in Fig. 1. Scarce data occur when any of these requirements are not satisfied: scarce data with a limited number of samples as illustrated in Fig. 1b and scarce data with sparsely and irregularly evaluated time series covariates as illustrated in Fig. 1c.

This paper first presents a review of the existing sequential and functional temporal predictive models. Observing that the existing method's limitations in dealing with the sparse and irregularly observed covariates, we introduce two sparse functional MLP based on the univariate and multivariate sparse functional principal component analysis [24–26].¹ The contributions of this paper are summarized as follows:

1. We discuss the different types of data scarcity for the temporal classification/regression problem.
2. We introduce a new temporal predictive model specially designed to handle scenarios with sparsely and irregularly observed time series inputs.
3. We use mathematical arguments and numerical experiments to investigate each model's feasibility and efficiency in building temporal predictive models under various types of data scarcity.

The rest of the paper is organized as follows. Some preliminaries, including the problem formulation, and review of sequential learning models and the conventional FMLP that only works for

dense and regular time series [22,2] are presented in Section 2. The proposed sparse FMLP for sparse time series inputs is described in Section 3. The performance of the candidate models under the two types of scarce data scenarios is respectively investigated in Section 4 and 5. The paper is concluded in Section 6.

2. Preliminaries

2.1. Problem formulation

The goal of temporal predictive models is to build a mapping from multivariate time series covariates to a scalar response variable with good accuracy and generalizability by leveraging the temporal patterns and dependencies.

Suppose that we have access to N independent training samples. For each subject $i \in \{1, \dots, N\}$, R features are repeatedly measured within a bounded time range $\mathcal{T} \subseteq \mathbb{R}$. In practice, the measuring timestamps can vary across different covariates and different subjects. As a result, the subject and feature indexes need to be included in the mathematical notations. In particular, the observed data for the r -th covariate of subject i is represented by a $M_{i,r}$ dimensional vector $\mathbf{Z}^{(i,r)} = [Z_1^{(i,r)}, \dots, Z_j^{(i,r)}, \dots, Z_{M_{i,r}}^{(i,r)}]^T$, which correspond to observations at timestamps $T_1^{(i,r)}, \dots, T_j^{(i,r)}, \dots, T_{M_{i,r}}^{(i,r)}$, with $T_j^{(i,r)} \in \mathcal{T}$, for $j = 1, \dots, M_{i,r}$. The response variable is Y_i , which is binary for temporal classifications and real-valued for temporal regressions. In summary, the observed data are $\{\mathbf{Z}^{(i,1)}, \dots, \mathbf{Z}^{(i,R)}, Y_i\}_{i=1}^N$, based on which the temporal predictive models aim at constructing the mapping in Eq. (1).

$$Y_i = F(\mathbf{Z}^{(i,1)}, \dots, \mathbf{Z}^{(i,R)}). \quad (1)$$

The sample size N , the number of observations per curve $M_{i,r}$, and the measuring timestamps $T_1^{(i,r)}, \dots, T_j^{(i,r)}, \dots, T_{M_{i,r}}^{(i,r)}$ jointly determine the level of data availability. A large data scenario depicted in Fig. 1 means that not only N and $M_{i,r}$ are sufficiently large but also the measuring timestamps uniformly cover the time domain \mathcal{T} for all subjects and covariates. Scarce data occur when at least one of these requirements is not satisfied. In particular, scarce data with a limited number of data points happen if the sample size N is small. Whereas, scarce data with sparsely evaluated time series features correspond to situations where $M_{i,r}$ is a small number and/or there exist large gaps among the measuring times $T_1^{(i,r)}, \dots, T_j^{(i,r)}, \dots, T_{M_{i,r}}^{(i,r)}$, for at least one subject and one covariate. This paper focuses on examining the advantages and disadvantages of several time series models in solving the problem in Eq. (1) given scarce data. We respectively present the candidate models in the remainder of this section.

2.2. Sequential learning models

Sequential learning models such as the Recurrent Neural Network (RNN), the Long Short-Term Memory (LSTM), and the Gated Recurrent Unit (GRU) are generalizations of the fully connected MLP that have internal hidden states to process the sequences of inputs [21]. The key idea is that they employ a series of MLP-based computational cells with the same architecture and parameters to build a directed neural network structure. Any individual cell in the network takes the actual observations at the current index and the hidden states obtained at the previous step (i.e., memory) to produce the updated hidden states that serve as the input for the next computational cell. When the goal is to predict the scalar response associated with the time series, the achieved hidden states at the last index are fed into a nonlinear function to compute the output state. In RNN, each computational cell has

¹ A preliminary version of this method appears as [1].

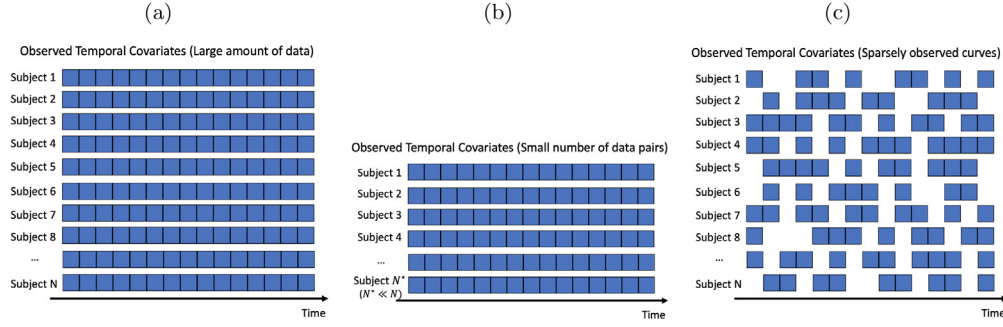


Fig. 1. Left: An example of large samples. Middle: Scarce data with a limited number of data pairs. Right: Scarce data with sparsely evaluated time series covariates.

one MLP, while there are multiple interacting MLPs in each recurrent unit in LSTM [4,21].

The sequential learning models originally designed for text data mining are capable of capturing the order information and the interactions among observations in sequential inputs [4,21]. Recently, these sequential learning models are also frequently adopted to model time series data [27,19,20]. However, when using these models to handle time series data, it is explicitly required that the multivariate inputs are evaluated at an equally-spaced time grid shared by all subjects. This is because the sequential learning models cannot encode the concrete measuring timestamps associated with the individual observations in the time series inputs. Mathematically, in the observed data samples $\{\mathbf{Z}^{(i,1)}, \dots, \mathbf{Z}^{(i,R)}, Y_i\}_{i=1}^N$, the covariate vectors $\mathbf{Z}^{(i,r)}$ are of the same length M across all subjects and all covariates. Also, they are evaluated at equally-spaced time grid within time range \mathcal{T} , which is denoted as $T_1, \dots, T_j, \dots, T_M$. In practice, data pre-processing procedures such as interpolation are often implemented to obtain the required regular time series when the raw data is sparse and irregular. It is noteworthy that the conventional data pre-processing techniques significantly contaminate the training data when the individual time series are highly sparse and irregular.

Let's use RNN as an example to present the mathematics behind the sequential learning models. Suppose that $\mathbf{Z}_j^{(i)} = [Z_j^{(i,1)}, \dots, Z_j^{(i,R)}]^T$ represents the R dimensional vector containing the R features at time T_j . Let the number of hidden units in MLP be L_{RNN} and the L_{RNN} dimensional hidden state at time T_j be $\mathbf{h}_j^{(i)}$. For $j = 0, \dots, M$, the following calculation is recursively conducted

$$\mathbf{h}_j^{(i)} = U_{\text{act1}}(\mathbf{W}_{\text{hh}}\mathbf{h}_{j-1}^{(i)} + \mathbf{W}_{\text{hz}}\mathbf{Z}_j^{(i)}), \quad (2)$$

where \mathbf{W}_{hh} is a L_{RNN} by L_{RNN} dimensional parameter matrix, and \mathbf{W}_{hz} is a L_{RNN} by R dimensional parameter matrix, and $U_{\text{act1}}(\cdot)$ is a non-linear activation function. Let \mathbf{W}_{yh} denote the parameter matrix that associates the last hidden state $\mathbf{h}_M^{(i)}$ with the response variable Y_i . In the output layer, the output is computed by

$$Y_i = U_{\text{act2}}(\mathbf{W}_{\text{yh}}\mathbf{h}_M^{(i)}), \quad (3)$$

where $U_{\text{act2}}(\cdot)$ is a non-linear activation function.

2.3. Functional data analysis and dense functional MLP

Functional data analysis (FDA) is an emerging branch in statistics that specializes in the analysis and theory of data dynamically evolves over a continuum. In general, FDA deals with data subjects that can be viewed as a functional form $X_i(t)$ over a continuous index t . Frequently encountered FDA-type data include time series data, tracing data such as hand-writings, and image data [7].

From the FDA modeling perspective, the $M_{i,r}$ observations of the r -th time series feature of subject i (i.e., $\mathbf{Z}^{(i,r)}$) are treated as discretized realizations from a continuous underlying curve $X^{(i,r)}(t)$ contaminated with zero-mean random errors.

$$Z_j^{(i,r)} = X^{(i,r)}(T_j^{(i,r)}) + \epsilon_{i,r,j}. \quad (4)$$

FDA predictive models directly handle the continuous time series features and solve the problem defined in Eq. (1) by learning

$$Y_i = F(X^{(i,1)}(t), \dots, X^{(i,R)}(t)). \quad (5)$$

Under certain assumptions on the smoothness of the underlying random functions $X^{(i,r)}(t)$ (e.g., continuous second derivatives exist), the conventional functional linear classification or regression models [7,28] assume and learn the unknown real-valued parameters in a linear-form mapping

$$Y_i = b + \sum_{r=1}^R \int_{t \in \mathcal{T}} W_r(\boldsymbol{\beta}_r, t) X^{(i,r)}(t) dt, \quad (6)$$

where $b \in \mathbb{R}$ is the unknown intercept, $\boldsymbol{\beta}_r$ is a finite-dimensional vector that quantifies the parameter function $W_r(\boldsymbol{\beta}_r, t)$, and $\int_{t \in \mathcal{T}} W_r(\boldsymbol{\beta}_r, t) X^{(i,r)}(t) dt$ is a generalization of vector inner product to $L^2(t)$ space and it aggregates the time-varying impact of the time series input on the response. Given Eq. (6), it can be seen that, unlike the sequential learning models, the functional models do not require equally-spaced time series observations and can be effectively trained end-to-end as long as the integral can be consistently approximated based on the actual observations $\mathbf{Z}^{(i,r)}$.

To capture more complex relationships, a functional MLP that embeds the linear calculation in Eq. (6) into the network structure of the conventional MLP [29] is introduced by [30,23] and later explored further by [2]. In particular, functional MLP proposed a new functional neuron that consists of the linear transformation in Eq. (6) and an additional non-linear activation step, as shown in Fig. 2. To build functional MLP, multiple functional neurons that take functional inputs and calculates a numerical output are placed on the first layer. The outputs from the functional layer are supplied into subsequent numerical neuron layers whose inputs and outputs are both scalar values, for further manipulations till the output layer that holds the response variable. An example FMLP with three functional neurons on the first layer and two numerical neurons on the second layer is illustrated in Fig. 2 [2].

For the simplicity of mathematical notations, let's consider the case where K functional neurons in the first layer and one numerical neuron in the second layer. Let $U_k(\cdot)$ be an activation function from \mathbb{R} to \mathbb{R} , $a_k, b_k \in \mathbb{R}$ for $k = 1, \dots, K$. The weight function for the r -th functional feature in the k -th functional neuron is assumed to be quantified by a finite dimensional unknown vector $\boldsymbol{\beta}_{k,r}$ and is denoted as $W_{k,r}(\boldsymbol{\beta}_{k,r}, t)$ for $k = 1, \dots, K$ and $r = 1, \dots, R$. Let

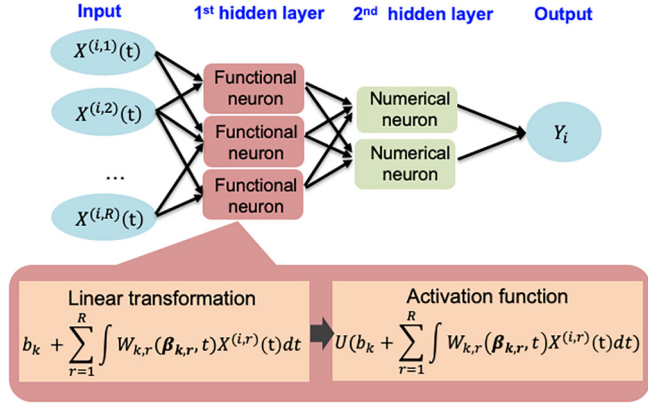


Fig. 2. The architecture of a functional MLP with three functional neurons on the first layer and two numerical neurons on the second layer [2].

$\beta = [\beta_{1,1}, \dots, \beta_{1,R}, \dots, \beta_{K,1}, \dots, \beta_{K,R}]^T$, $\mathbf{X}^{(i)} = [X^{(i,1)}(t), \dots, X^{(i,R)}(t)]^T$. Then the scalar output of the first layer is

$$H(\mathbf{X}^{(i)}, \beta) = \sum_{k=1}^K a_k U_k(b_k + \sum_{r=1}^R \int_{t \in \mathcal{T}} W_{k,r}(\beta_{k,r}, t) X^{(i,r)}(t) dt). \quad (7)$$

For more details about the functional MLP including the specifications of the weight functions $W_{k,r}(\beta_{k,r}, t)$ in the functional neurons, and assumptions and implementations of the gradient descent based training procedure, we list the literature [30,2] as good references. It is noteworthy that this conventional functional MLP requires that the number of observations per curve is large enough and the individual observations are relatively regularly spaced (i.e., Fig. 1b and 1c), as stated in Theorem 1 in [30]. In the next section, we propose an effective way of generalizing functional MLP to sparsely and irregularly observed time series inputs.

3. Proposed sparse functional MLP

When there is a limited amount of irregularly-spaced data available per feature curve, as shown in Fig. 1c, the existing models described in the previous section are no longer feasible solutions. The deep sequential models need dense and regular observations to model the temporal information through the recurrent network structure. A common practice is to conduct data interpolation to obtain the required dense regular data, however, the conventional interpolation techniques such as the cubic B-spline and the Gaussian process regression often produce biased curve estimates. Analogously, the numerical integration calculation in the conventional FMLP is problematic under sparse data scenarios. In this section, we propose a novel algorithm for generalizing Multilayer Perceptron (MLP) to handle sparse functional data, wherein for a given subject there are multiple observations available over time and these observations are sparsely and irregularly distributed within the considered time range.

3.1. Sparse functional MLP based on univariate PACE

To derive and define a functional neuron that is calculable for sparse functional data, we propose to go one step further than the existing functional neuron in Eq. (7) with the help of the functional principal component analysis [31,9,24]. Let the r -th feature be a random process with an unknown mean function $\mu_r(t)$ and an unknown covariance function $G_r(t, t')$, $t, t' \in \mathcal{T}$. Mathematically, the non-increasing eigenvalues $\{\lambda_{r,p}\}_{p=1}^\infty$ and the corresponding eigenfunctions $\{\phi_{r,p}(t)\}_{p=1}^\infty$ are solutions of

$$\lambda \phi(t) = \int_{t' \in \mathcal{T}} G_r(t, t') \phi(t') dt'. \quad (8)$$

The eigenfunctions are orthonormal in the sense that $\int_{t \in \mathcal{T}} \phi_{r,p}(t) \phi_{r,p'}(t) dt = 0$ for $p \neq p'$ and $\int_{t \in \mathcal{T}} \phi_{r,p}^2(t) dt = 1$. Based on the orthonormality of the eigenfunctions, the r -th feature of subject i can be represented as

$$X^{(i,r)}(t) = \sum_{p=1}^\infty \eta_{i,r,p} \phi_{r,p}(t), \quad (9)$$

where $\eta_{i,r,p} = \int_{t \in \mathcal{T}} X^{(i,r)}(t) \phi_{r,p}(t) dt$. A common practice for basis expansion-based methods in FDA [32,15] is to truncate the expansion at the first several directions. This practice is supported by the fact that the core information regarding $X^{(i,r)}(t)$ is mostly captured by the first several basis functions when the curve is smooth in a certain degree. Strict theoretical proofs can be found in [24]. For the r -th feature, let's truncate at the first P_r dimensions and plug the $X^{(i,r)}(t)$'s approximated representation into the FMLP in Eq. (7), we have

$$H(\mathbf{X}^{(i)}, \beta) \approx \sum_{k=1}^K a_k U_k(b_k + \sum_{r=1}^R \int_{t \in \mathcal{T}} W_{k,r}(\beta_{k,r}, t) \times \sum_{p=1}^{P_r} \eta_{i,r,p} \phi_{r,p}(t) dt). \quad (10)$$

Practically, P_r can be selected using the fraction of variance explained approach, AIC or BIC criterion based approach, or the leave-one-curve-out cross-validation method [33].

Given Eq. (10), it can be seen that the model still cannot directly go through as we cannot consistently estimate $\eta_{i,r,p} = \int_{t \in \mathcal{T}} X^{(i,r)}(t) \phi_{r,p}(t) dt$ from the sparse observations $\mathbf{Z}^{(i,r)}$. Borrowing the idea from sparse data Principal Components Analysis through Conditional Expectation (PACE) [24], we propose to estimate $\eta_{i,r,p}$ by its best linear unbiased predictor, $E[\eta_{i,r,p} | \mathbf{Z}^{(i,r)}]$. This is a reasonable choice to estimate $\eta_{i,r,p}$, because if we take the randomness of measuring time t into account, then

$$E_t[E[\eta_{i,r,p} | \mathbf{Z}^{(i,r)}]] = E[\eta_{i,r,p}]. \quad (11)$$

That is to say the random quantities $E[\eta_{i,r,p} | \mathbf{Z}^{(i,r)}]$ and $\eta_{i,r,p}$ share the same expectation. Motivated by a special case where the observations $\mathbf{Z}^{(i,r)}$ and the random errors $\epsilon_{i,r,j}$ are jointly Gaussian distributed, given Eqs. (4) and (9), we can get the explicit formula for $E[\eta_{i,r,p} | \mathbf{Z}^{(i,r)}]$,

$$E[\eta_{i,r,p} | \mathbf{Z}^{(i,r)}] = \gamma_{r,p} + \lambda_{r,p} \phi_{i,r,p}^T [\phi_{i,r} \text{diag}(\lambda_r) \phi_{i,r}^T + \sigma_r^2 I]^{-1} (\mathbf{Z}^{(i,r)} - \boldsymbol{\mu}_{i,r}), \quad (12)$$

where $\gamma_{r,p} = \int \phi_{r,p}(t) \mu_r(t) dt$, $\lambda_r = [\lambda_{r,1}, \dots, \lambda_{r,P_r}]^T$, and $\phi_{i,r,p}$ is the eigenfunction $\phi_{r,p}(t)$ evaluated at the $M_{i,r}$ observing time points, i.e., $\phi_{i,r,p} = [\phi_{r,p}(t_{i,r,1}), \dots, \phi_{r,p}(t_{i,r,M_{i,r}})]^T$. $\phi_{i,r}$ is a $M_{i,r} \times P_r$ matrix, with the p -th column being $\phi_{i,r,p}$. σ_r is the standard deviation of the random noise $\epsilon_{i,r,j}$. By plugging Eq. (12) back into Eq. (10), we achieve the output of the first layer of our *proposed sparse functional MLP* with K neurons,

$$\tilde{H}(\mathbf{X}^{(i)}, \beta) = \sum_{k=1}^K a_k U_k(b_k + \sum_{r=1}^R \int_{t \in \mathcal{T}} W_{k,r}(\beta_{k,r}, t) \sum_{p=1}^{P_r} E[\eta_{i,r,p} | \mathbf{Z}^{(i,r)}] \phi_{r,p}(t) dt). \quad (13)$$

The sparse functional neuron in the above equation plays the same role as the functional neuron in Eq. (7) for the conventional functional MLP [30].

Before training the proposed sparse functional MLP model, we need to first estimate the unknown values in Eq. (10) and Eq. (12), including the eigenfunctions $\phi_{r,p}(t)$, eigenvalues $\lambda_{r,p}$, standard deviation of random error σ_r . They can be estimated using the restricted maximum likelihood estimation in [33], the local linear smoothing based method in [24], or the EM algorithm in [34]. Since This is not the main focus in this paper, we skip the details. The point is that our algorithm needs consistent estimations for the functional components. All the three methods have been proved to produce consistent estimates for these eigen components for sparse functional data and they can be used to perform the estimation step. In our numerical experiments in Section 5, we used the local linear smoothing based method in [24]. Let's denote the estimated values as $\hat{\phi}_{r,p}(t)$, $\hat{\lambda}_{r,p}$ and $\hat{\sigma}_r$. Then we have the empirical counterparts for Eqs. (10) and (12), denoted as $\hat{H}(\mathbf{X}^{(i)}, \beta)$ and $\hat{E}[\eta_{i,r,p}|\mathbf{Z}^{(i,r)}]$.

After the estimation step, similar to the dense functional MLP in [30], we propose to use gradient based algorithms to train our sparse functional MLP model. The forward propagation step can go through as follows. First, in the functional neurons, the integral $\int_{t \in \mathcal{T}} W_{k,r}(\beta_{k,r}, t) \sum_{p=1}^{P_r} \hat{E}[\eta_{i,r,p}|\mathbf{Z}^{(i,r)}] \hat{\phi}_{r,p}(t) dt$ is approximated by the numerical integration techniques. First layer's output $\hat{H}(\mathbf{X}^{(i)}, \beta)$ can then calculated by the formula in Eq. (10). The forward propagation calculation in subsequent numerical layers is straightforward. In the backward propagation step, the partial derivatives from the output layer up to the second hidden layer (i.e., the numerical layer after the functional neuron layer) can be easily calculated as before. Whereas, it is essential to ensure that the partial derivatives of the values at the second layer (i.e., $\hat{H}(\mathbf{X}^{(i)}, \beta)$) with respect to the parameters β exist. This requires that $\partial W_{k,r}(\beta_{k,r}, t) / \partial \beta_{k,r,q}$ exists almost everywhere for $t \in \mathcal{T}$. Under this assumption, $\partial \hat{H}(\mathbf{X}^{(i)}, \beta) / \partial \beta_{k,r,q}$ for any $k = 1, \dots, K; r = 1, \dots, R; q = 1, \dots, Q_r$ can be estimated using numerical approximations of the following quantity

$$\begin{aligned} \frac{\partial \hat{H}(\mathbf{X}^{(i)}, \beta)}{\partial \beta_{k,r,q}} &\approx a_k U'_k \left(\left(b_k + \sum_{r=1}^R \int_{t \in \mathcal{T}} W_{k,r}(\beta_{k,r}, t) \right. \right. \\ &\quad \times \sum_{p=1}^{P_r} \hat{E}[\eta_{i,r,p}|\mathbf{Z}^{(i,r)}] \hat{\phi}_{r,p}(t) dt \Big) \times \int_{t \in \mathcal{T}} \\ &\quad \times \frac{\partial W_{k,r}(\beta_{k,r}, t)}{\partial \beta_{k,r,q}} \sum_{p=1}^{P_r} \hat{E}[\eta_{i,r,p}|\mathbf{Z}^{(i,r)}] \hat{\phi}_{r,p}(t) dt. \end{aligned} \quad (14)$$

To justify the validity of our proposal, we have provided brief arguments regarding the consistency of using estimated values as well as the equivalence between our sparse and the proposed dense MLP under dense regular data scenarios.

3.2. Sparse functional MLP based on multivariate FPCA

The functional neurons in Eq. (10) (i.e., an equivalent of the dense functional neuron in Section 2.3) and Eq. (13) (i.e., the proposed sparse functional neuron) are based on the univariate functional principal component analysis. When separately conducting FPCA, the joint variations among the R variables $\{X^{(i,1)}(t), \dots, X^{(i,R)}(t)\}$ are not captured, which makes the random scores from different variables (i.e., $\eta_{i,r,p}$ and $E[\eta_{i,r,p}|\mathbf{Z}^{(i,r)}]$) being correlated and causes multicollinearity issues during modeling. An example of correlated random scores are illustrated in Fig. 12 in Section 5. To overcome this issue, we propose functional neural networks based on the multivariate FPCA [26,25] that are described as follows.

Let $\mathbf{W}_{k,r}(\beta_k, t) = [W_{k,1}(\beta_{k,r}, t), \dots, W_{k,r}(\beta_{k,R}, t)]^T$, then Eq. (7) can be written as in the following vector format

$$H(\mathbf{X}^{(i)}, \beta) = \sum_{k=1}^K a_k U_k([b_k + \int_{t \in \mathcal{T}} \mathbf{W}_{k,r}(\beta_k, t)^T \mathbf{X}^{(i)}(t)]). \quad (15)$$

Let's denote the $R \times R$ matrix that quantifies the covariance of each variable and the joint variation between variables as $\mathbf{G}(t, t')$, with the (r, r') -th element being $G_{r,r'}(t, t') = \text{Cov}(X^{(i,r)}(t), X^{(i,r')}(t'))$. According to the multivariate FPCA, there exists a set of R dimensional orthonormal eigenfunction vectors $\tilde{\phi}_p(t) = [\tilde{\phi}_{1,p}(t), \dots, \tilde{\phi}_{R,p}(t)]^T$, for $p = 1, \dots, \infty$, such that

$$\int \mathbf{G}(t, t') \tilde{\phi}_p(t') dt' = \tilde{\lambda}_p \tilde{\phi}_p(t), \text{ with } \lim_{p \rightarrow \infty} \tilde{\lambda}_p = 0, \quad (16)$$

where $\tilde{\lambda}_p \in \mathbb{R}$ is the eigenvalue corresponding to the p -th eigenfunction vector $\tilde{\phi}_p(t)$. Accordingly, it has been shown that the R dimensional data $\mathbf{X}^{(i)}(t)$ can be represented by

$$\mathbf{X}^{(i)}(t) = \sum_{p=1}^{\infty} \tilde{\eta}_{i,p} \tilde{\phi}_p(t) \approx \sum_{p=1}^P \tilde{\eta}_{i,p} \tilde{\phi}_p(t), \quad (17)$$

where $\tilde{\eta}_{i,p} = \sum_{r=1}^R \int X^{(i,r)}(t) \tilde{\phi}_{r,p}(t) dt$. Comparing the separate FPCA in Section 3.1 with the multivariate FPCA in Eqs. (16) (17), it can be seen that univariate FPCA is a special case of the multivariate FPCA that assumes zero joint variation between variables. Theoretically, we expect the functional regression models in this section to outperform those in Section 3.1. The magnitude of improvement is affected by the size of joint variation between variables and the complexity of the underlying mapping.

Given Eq. (15)(16) (17), when all the variables $\{X^{(i,r)}(t), i = 1, \dots, N; r = 1, \dots, R\}$ are densely and regularly evaluated, the multivariate functional neuron is defined as follows

$$H_M(\mathbf{X}^{(i)}, \beta) \approx \sum_{k=1}^K a_k U_k([b_k + \int_{t \in \mathcal{T}} \mathbf{W}_{k,r}(\beta_k, t)^T \sum_{p=1}^P \tilde{\eta}_{i,p} \tilde{\phi}_p(t) dt]). \quad (18)$$

For sparsely evaluated data, analogous to Eq. (10), the sparse multivariate function neuron is

$$H_M(\mathbf{X}^{(i)}, \beta) \approx \sum_{k=1}^K a_k U_k([b_k + \int_{t \in \mathcal{T}} \mathbf{W}_{k,r}(\beta_k, t)^T \sum_{p=1}^P E[\tilde{\eta}_{i,p}|\mathbf{Z}^{(i,1)}, \dots, \mathbf{Z}^{(i,R)}] \tilde{\phi}_p(t) dt])$$

$$E[\tilde{\eta}_{i,p}|\mathbf{Z}^{(i,1)}, \dots, \mathbf{Z}^{(i,R)}] = \sum_{r=1}^R E[\tilde{\eta}_{i,p}|\mathbf{Z}^{(i,r)}]. \quad (19)$$

Note that $E[\tilde{\eta}_{i,p}|\mathbf{Z}^{(i,r)}]$ is calculated by replacing $\phi_{r,p}(t)$ and $\lambda_{r,p}$ with $\tilde{\phi}_{r,p}(t)$ and $\tilde{\lambda}_p$ respectively. Based on the multivariate function neurons discussed above, we propose multivariate FMLP by embedding these neurons in the architecture in Fig. 2.

Next, we describe how to numerically implement the multivariate FMLP as follows. The core theoretical result in [25] is that there is an analytical relationship between the univariate FPCA and the multivariate FPCA, which implied that $\tilde{\phi}_{r,p}(t)$ and $\tilde{\lambda}_p$ can be calculated by estimating $\phi_{r,p}(t)$ and $\lambda_{r,p}$ for all $r = 1, \dots, R$. In particular, let the estimated score from univariate FPCA be $\hat{s}_{i,r,p} = \hat{\eta}_{i,r,p}$ or $\hat{s}_{i,r,p} = \hat{E}[\eta_{i,r,p}|\mathbf{Z}^{(i,r)}]$, for $i = 1, \dots, N; r = 1, \dots, R; p = 1, \dots, P_r$. Let $P_+ = \sum_{r=1}^R P_r$ and Ξ is a $P_+ \times P_+$ consisting of blocks $\Xi^{(rr')} \in \mathbb{R}^{P_r \times P_{r'}}$ with the (p, p') -th entry being

$$\begin{aligned} \Xi_{pp'}^{(rr')} &= \text{Cov}(\hat{s}_{i,r,p}, \hat{s}_{i,r',p'}) \\ &= \frac{1}{N-1} \sum_{i=1}^N (\hat{s}_{i,r,p} - \bar{\hat{s}}_{i,r,p})(\hat{s}_{i,r',p'} - \bar{\hat{s}}_{i,r',p'}) \end{aligned} \quad (20)$$

Let's conduct eigen decomposition on matrix Ξ and denote the p -th eigenvector as \mathbf{c}_p . Note that \mathbf{c}_p can be considered as a vector consisting of R blocks, with the r -th block being denoted as $[\mathbf{c}_p]_m^{(r)} \in \mathbb{R}^{P_r}$. According to the proposition in [25], we can estimate $\hat{\phi}_{r,p}(t)$ by

$$\hat{\phi}_{r,p}(t) = \sum_{m=1}^{P_r} [\mathbf{c}_p]_m^{(r)} \hat{\phi}_{r,m}(t), \quad (21)$$

where $\hat{\phi}_{r,m}(t)$ is the achieved eigenfunction from conducting univariate FPCA on the r -th variable. The joint eigenvalue $\hat{\lambda}_p$ is the same as the eigenvalue of matrix Ξ .

In summary, we propose two types of extensions of the FMLP in [22,2]. The model in Section 3.1 equipped with the sparse functional principal component analysis is proposed to handle sparse data cases. The models in Section 3.2 are generalizations of the FMLP to explicitly account for the correlations among variables. The FMLPs using the conventional FPCA are more straightforward to implement, while the multivariate FMLPs are expected to be more accurate, especially when the correlations among the variables are large.

4. Scarce data with limited number of samples

In this section, we present a comparative study of the sequential learning models and the regular FMLP (equivalently, the proposed sparse FMLP) under scenarios where the sample size is small, i.e., Fig. 1b. We first theoretically compare the minimum sample size required by each model by calculating the number of parameters. We also discuss their feasibility and efficiency in dealing with two different types of time series inputs. Finally, we conduct numerical experiments to demonstrate their performance in solving the challenging remaining useful life prediction task in the Predictive Maintenance domain, given the limited amount of training data.

4.1. Theoretical comparison

4.1.1. Comparing the number of parameters

To understand the minimum sample size required to train each candidate model, we first present the mathematical formula for the number of unknown parameters in the considered deep learning models. For the simple RNN described in Section 2.2, supposing that there are L_{RNN} hidden states, the total number of unknown parameters $\text{Count}_{\text{RNN}}$ is given in Eq. (22). The first term $L_{\text{RNN}}(L_{\text{RNN}} + R)$ in the equation corresponds to the sequential processing by memory cells in Eq. (2), while the second term is related to the output state in Eq. (3). The number of unknown parameters for LSTM is four times that of RNN, as there are an input gate, an output gate, and a forget gate in addition to the RNN-like memory cell. Gated recurrent units (GRUs) are a type of hidden activation function in recurrent neural networks. GRU is like a long short-term memory (LSTM) with a forget gate, but has fewer parameters than LSTM, as it lacks an output gate. GRU has been used to model speech, music, and language data by maximizing the conditional probability of a target sequence given a source sequence. For the functional MLP in Section 2.3 and 3, let's denote the length of unknown parameters in the parameter functions $W_{k,r}(\beta_{k,r}, t)$ as $Q_{k,r}$ and the number of hidden functional neurons as L_{FMLP} . Then the number of unknown parameters in FMLP $\text{Count}_{\text{FMLP}}$ can be obtained by the formula in Eq. (22). Note that the first term represents the connections between layers and the second term corresponds to the biases in every layer.

$$\begin{aligned} \text{Count}_{\text{RNN}} &= L_{\text{RNN}}(L_{\text{RNN}} + R) + L_{\text{RNN}} = O(L_{\text{RNN}}^2) \\ \text{Count}_{\text{LSTM}} &= 4(L_{\text{LSTM}}(L_{\text{LSTM}} + R) + L_{\text{LSTM}}) = O(L_{\text{LSTM}}^2) \\ \text{Count}_{\text{GRU}} &= 3(L_{\text{GRU}}(L_{\text{GRU}} + R) + L_{\text{GRU}}) = O(L_{\text{GRU}}^2) \\ \text{Count}_{\text{FMLP}} &= \left(\sum_{k=1}^{L_{\text{FMLP}}} \sum_{r=1}^R Q_{k,r} + L_{\text{FMLP}} \right) + (L_{\text{FMLP}} + 1) = O(L_{\text{FMLP}}) \end{aligned} \quad (22)$$

Based on Eq. (22), it can be seen that the total number of unknown parameters of RNN, LSTM, and GRU are in a quadratic order of the number of hidden units, while the number of parameters of FMLP is in the same order with L_{FMLP} , the number of functional neurons. This indicates that the minimal sample size required by the sequential learning model is theoretically much larger than FMLP when the underlying mapping $F(\cdot)$ in Eq. (1) and (5) is complex and needs a comparable large number of hidden units in the sequential learning models and FMLP to be well approximated. On the other hand, when L_{RNN} and L_{FMLP} are significantly different (i.e., $L_{\text{RNN}} \ll L_{\text{FMLP}}$, or $L_{\text{RNN}} \gg L_{\text{FMLP}}$), the model that can more efficiently capture the temporal information in the covariates requires less number of training samples. As a next step, we study the candidate model's feasibility and efficiency in terms of temporal information capturing under different circumstances. LSTM is the popularly used in many applications, in the rest of the paper, we use LSTM as a representative of the sequential learning models.

4.1.2. Comparison under different scenarios

In real practice, the observed time series data often contains certain zero-mean noises. That is there exist the additive relationship among the actual observation $Z_j^{(i,r)}$, the underlying continuous process $X^{(i,r)}(t)$ that give rises to the time-specific observation, and the random noise $\epsilon_{i,r,j}$, as indicated by Eq. (4). Depending on its mathematical properties, the individual function $X^{(i,r)}(t)$ can be divided into two categories, consisting of smooth functions whose continuous second derivatives exist and non-smooth functions otherwise. Both smooth and non-smooth processes are frequently encountered in real-world applications. Examples of time series data from smooth underlying functions include child growth over time, traffic flow during the day, accumulative number of positive cases over time for a certain pandemic. Continuous processes of the non-smooth nature include the vibration data and acoustic data, of which the rapidly changing dynamics contain the key information to distinguish a given function $X^{(i,r)}(t)$ from other random samples $\{X^{(i,r)}(t)\}_{i \neq j}$.

According to Section 2, the sequential learning models are more appropriate models for building predictive models with non-smooth time series covariates. This is because the sequential learning models conduct computations on the individual observations at each timestamp and offer a good way of capturing the non-linear dependencies within the highly dynamic time series as well as the complex relationship between the time series and the response. Whereas, the functional predictive models attempt to model the infinite-dimensional continuous curve and therefore essentially rely on certain smoothness assumptions on each underlying process $X^{(i,r)}(t)$ to overcome the curse of dimensionality in the formulation in Eq. (6) and (7) [7]. More specifically, the assumption that the parameter function $W_{k,r}(\beta_{k,r}, t)$ can be determined by a finite-dimensional vector $\beta_{k,r}$, i.e., the assumed model can be trained end-to-end accordingly, only if the underlying process is smooth.

When it comes to smooth time series inputs, the FMLP has advantages over the sequential learning models in general thanks to the basis expansion technique in FDA. The key idea of basis expansion is to set the weight function $W_{k,r}(\beta_{k,r}, t)$ as a linear combination of a set of fixed or data-driven basis functions $\{\phi_{k,r,p}(t)\}_{p=1}^{Q_{k,r}}$.

$$W_{k,r}(\beta_{k,r}, t) = \sum_{p=1}^{Q_{k,r}} \beta_{k,r,p} \phi_{k,r,p}(t). \quad (23)$$

Then the core temporal information processing unit $\int_{t \in \mathcal{T}} W_{k,r}(\beta_{k,r}, t) X^{(i,r)}(t) dt$ in FMLP becomes

$$\int_{t \in \mathcal{T}} W_r(\beta_r, t) X^{(i,r)}(t) dt = \sum_{p=1}^{Q_{k,r}} \beta_{k,r,p} \int_{t \in \mathcal{T}} \phi_{k,r,p}(t) X^{(i,r)}(t) dt. \quad (24)$$

This means that the infinite-dimensional variation in $X^{(i,r)}(t)$ is transferred to the $Q_{k,r}$ scalar random variables $\{\int_{t \in \mathcal{T}} \phi_{k,r,p}(t) X^{(i,r)}(t) dt\}_{p=1}^{Q_{k,r}}$, the projection scores of $X^{(i,r)}(t)$ onto the pre-defined basis functions. The basis function-based parameter specification in Eq. (23) is supported by theoretical arguments that the infinite continuous stochastic process $X^{(i,r)}(t)$ can be consistently represented by a finite set of basis functions given a certain degree of smoothness [7,31]. The benefit of the formulation in Eq. (24) is two-folds. First, the temporal information in $X^{(i,r)}(t)$ can be more succinctly captured by the integrals, which take less time to compute than RNN, especially when the time series inputs are long time series. Second, it allows us to embed prior domain knowledge about the characteristic of time series features into the temporal predictive model. For instance, if we know there are sparse jumps, spikes, peaks in the time series covariates, the wavelet basis is a good choice to extract meaningful projection scores. On the other hand, the sequential learning models need more complex architectures to capture this useful knowledge. This means that the number of hidden units and accordingly the minimal sample size required in FMLP is in general much less than the sequential learning models for a target level model performance.

4.2. Comparison of performance on equipment remaining useful life prediction

Remaining Useful Life (RUL) of equipment or one of its components is defined as the time left until the equipment or component reaches its end of useful life. Accurate RUL prediction is exceptionally beneficial to Predictive Maintenance, and Prognostics and Health Management (PHM). Recently, data-driven solutions that utilize historical sensor and operational data to estimate RUL are gaining popularity. In particular, the RUL prediction problem is usually formulated as a temporal regression problem defined in Eqs. (1) and (5). In this section, we compare the performance of functional MLP ('FMLP'), LSTM, and the traditional multivariate regression models that treat the measurements in time series as features ('MLP' and 'SVR') in solving the RUL prediction task with a limited number of training samples in a widely-used benchmark data set called NASA C-MAPSS (Commercial Modular Aero-Propulsion System Simulation) data [35].

4.2.1. Background and data pre-processing

Background: C-MAPSS data set consists of simulated 21 sensor readings, 3 operating condition variables for a group of turbofan engines as they running until some critical failures happen. There are four data subsets in C-MAPSS that correspond to scenarios with different numbers of operating conditions and fault modes [35].

Each subset is divided into the training and testing sets. The training sets contain run-to-failure data where engines are fully observed from an initial healthy state to a failure state. The testing sets consist of prior-to-failure data where engines are observed until a certain time before failure. Table 1 provides a summary for each subset in C-MAPSS, including the number of operating conditions and fault modes, and the number of subjects in the training and testing phase.

Removing the effect of operating conditions: For the second and the fourth sub-datasets, there are six operating conditions reflected by the three operating condition variables. To remove the effects of operating conditions, for each of the 21 sensors, we train an MLP model to learn a mapping from the operating condition variables to the sensor variable. In particular, for a given sensor in 'FD002', we use data from all the 260 engines in the training set to train the MLP model. The sample size is the summation of the number of observations across the 260 engines. The input data is the three operation condition variables and the output is the considered sensor data. Then we normalize the considered sensor variable by deducting the fitted value of the MLP model from the raw sensor readings. The raw sensor trajectory and normalized trajectory of the second sensor for a randomly selected engine in the training set of FD002 are visualized in Fig. 3.

Window-sliding and RUL labeling: In the C-MAPSS data set, the engines in the training and testing sets are observed for a different number of time cycles. Moreover, the full sensor data trajectories in the testing sets are blinded for a variety of periods, therefore the true RUL labels are distributed variously. To handle this phenomenon, we propose to use the window sliding technique used in [36,37]. Let's denote the smallest number of sensor measurements for the individual engines in data subset d as \mathcal{M}_d for $d = 1, \dots, 4$. The values for $\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3, \mathcal{M}_4$ are 31, 21, 38, 19 respectively. The functional inputs and RUL labels are generated as follows. For the d -th subset, trajectories corresponding to each engine in the training and testing data sets are cut into multiple data instances of length \mathcal{M}_d . For instance, the first engine in the training set of FD001 fails at the 144th cycle. A total of 114 training data instances are generated from this engine, with the c -th data instance being the sensor measurements between time cycle c and $c + \mathcal{M}_d - 1$. To specify the RUL labels for the 114 data instances of this engine, we adopt the widely-used piece-wise labeling approach in relevant literature [38,27]. Under the observation that the degradation in the performance is negligible at the beginning period and it starts to degrade linearly at some point T , the RUL label is defined as

$$RUL_{c, \text{piecewise}} = \min\{T, RUL_{c, \text{linear}}\}. \quad (25)$$

Note that in our experiment, we set $T = 130$, following the specifications in the prior art [38,27].

4.2.2. Implementations and results

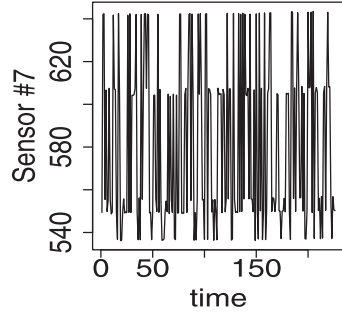
Implementation of FMLP: Following the implementation of LSTM in [27], we use the Min-Max normalization to scale the individual sensor sequences to the range of [0, 1]. The specific mathematical formula can be found in [27,2]. A FMLP with a two-layered architecture is deployed to learn the mapping from 21 sensors to the RUL label. There are four functional neurons (i.e., $K = 4$) on the first layer and two numerical neurons on the second layer in the FMLP. The activation function on both layers are the standard logistic function, i.e., $U_k(u) = \frac{1}{1+e^{-u}}$. To better deal with the complex sensor data, we propose to specify data-driven weight functions by calculating the eigenfunctions from data. Let the estimated eigenfunction from the N samples of the r -th sensor be $\hat{\phi}_{r,p}(t)$. The weight functions are then specified as

Table 1

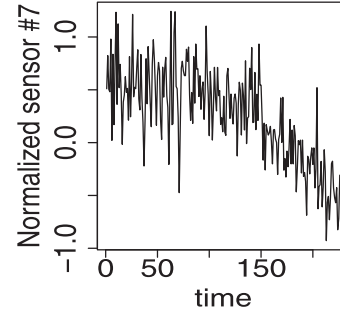
Summary of the subsets in C-MAPSS data set.

	FD001	FD002	FD003	FD004
# of engines in training	100	260	100	249
# of engines in testing	100	259	100	248
# of operating conditions	1	6	1	6
# of fault modes	1	1	2	2

(a) Raw sensor data for one randomly selected engine in the training set of FD002.



(b) Normalized sensor data after removing the effect of operating conditions.

**Fig. 3.** Removing the effect of operating conditions on sensor data.

$$W_{k,r}(\beta_{k,r}, t) = \sum_{p=1}^{P_{k,r}} \beta_{k,r,p} \hat{\phi}_{r,p}(t), \quad (26)$$

where $P_{k,r}$ is selected from the regularly-used fraction of variance explained (FVE) with a 80% cutoff. In practice, there are four commonly-used cutoff values, i.e., 80%, 90%, 95%, and 99%. In this experiment, we choose the smallest rule-of-thumb value. This is because choosing a smaller FVE helps retaining the key smooth patterns in sensor data and removing the random noises shown in Fig. 3.

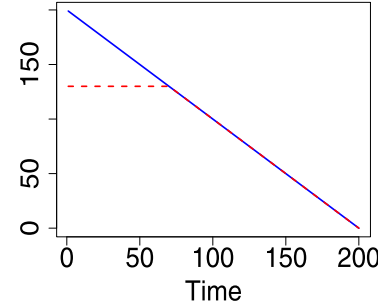
Evaluation metrics: We evaluate the performance of functional MLP with the same evaluation strategy used in [38,27]. Suppose that there are N subjects in the testing set, and the true RUL since the last observation of engine i is $RUL_{i,true}$ and the estimated RUL is $RUL_{i,est}$. The root mean squared error (RMSE) calculated from the N engines is defined as Fig. 4

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (RUL_{i,est} - RUL_{i,true})^2}. \quad (27)$$

Results: The RMSE of FMLP together with the results of LSTM and the multivariate regression models including the Support vector regression ('SVR') and the multilayer perceptron ('MLP') from previous literature are summarized in Tables 2. For all the four subsets, functional MLP significantly outperforms the baseline methods in terms of RMSE. The average improvement over LSTM [27] is 26.89%. For industrial equipment like turbofan engines, the sensor signals over time are often correlated with the smooth degradation process and thus can be assumed to be generated by smooth continuous functions. The experimental results in Tables 2 numerically justify our discussion about the advantage of FMLP over the state-of-art models in handing smooth time series covariates in Section 4.1.

5. Scarce data with sparse time series features

In this section, we consider circumstances where the time series inputs are sparsely and irregularly observed over the time domain. The sparse functional MLP (SFMLP) in Section 3.1 and the SFMLP

**Fig. 4.** Remaining useful life label for a given engine: the red line represents the piece-wise RUL label capped at $T = 130$.

equipped with multivariate FPCA in Section 3.2 are temporal predictive models that are specially designed to handle this type of scenario. The sequential learning models are incapable of directly utilizing the sparsely evaluated temporal information for proper model building. They require to preliminarily fill in the gaps in the raw sparse data based on certain interpolation techniques. The performance of the sequential learning models heavily relies on the accuracy of interpolation. A description of this interpolation plus sequential learning approach is provided in the first part of this Section. Next, we conduct three numerical experiments to demonstrate the superior performance of the proposed sparse FMLPs for sparse data scenarios.

5.1. Sequential learning models under sparse data

As discussed in Section 2.2, the sequential deep learning models essentially require the time series covariates to be densely observed at an equally-spaced time grid. Under sparse data cases, certain data interpolation techniques such as the cubic spline interpolation are often deployed to get data readings of the same interval. The interpolated data are then fed into sequential learning models to build temporal predictive models. The performance of

Table 2

RMSE comparison on C-MAPSS data and improvement ('IMP') of functional MLP over LSTM [27].

Model	FD001	FD002	FD003	FD004
MLP[38]	37.56	80.03	37.39	77.37
SVR[38]	20.96	42.00	21.05	45.35
LSTM[27]	16.14	24.49	16.18	28.17
FMLP	13.36	16.62	12.74	17.76
IMP	17.22%	32.14%	21.26%	36.95%

* IMP w.r.t LSTM is $(\text{RMSE}_{\text{LSTM}} - \text{RMSE}_{\text{FMLP}}) / \text{RMSE}_{\text{LSTM}}$.

the sequential learning models heavily relies on the accuracy of interpolation.

The conventional way of performing interpolation is to separately apply techniques such as cubic B-spline and Gaussian process regression on the sparse observations from each subject. This common practice is illustrated in Fig. 5. However, the recovered curves using data from the individual curves alone are often not consistent estimates for the true underlying curves, due to the limited amount of information available per curve. For instance, as shown by the left figure in Fig. 6, the interpolated curves are significantly dispersed from the actual sin-shaped random functions with two full cycles. As a consequence, the sequential learning models built based upon the biased input data are not reliable.

Observing the benefit of jointly using time series across samples to estimate or model the temporal variations within time series data in functional data analysis (i.e., the right figure in Fig. 6), in this paper, we also consider the functional data approach PACE in Eqs. (9) and (12) as an interpolation method for the sequential learning models. As shown in the numerical experiments in Section 5.2, the performance of LSTM when the recovered data are generated by PACE is significantly better than the other interpolations for the considered problems. In general, FDA type of models is valid when the time series of different subjects can be considered as random samples from an underlying random process, which is a reasonable assumption under many real-world use cases. Therefore, we recommend using FDA-type modelings as interpolation when dealing with sparse time series.

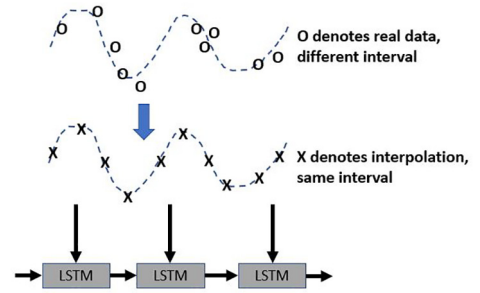
5.2. Performance comparison in numerical experiments

In this section, we conduct three numerical studies, including classification of synthetic curves, prediction of patient's survival beyond a given period, and prediction of engine's remaining time to failure. We compare our sparse functional MLP ('sparse FMLP' or 'SFMLP') with the 'LSTM' as the baseline methods. Our proposed method outperforms LSTM in all three numerical studies, where the input variables are observed multiple times for each subject, and the observing times are irregularly spaced and are not shared across subjects.

5.2.1. Curve classification on simulated data

In this subsection, we conduct a numerical experiment on synthetic data, with the objective of showing the benefit of modeling repeated data over time in the functional fashion as well as the validity of the proposed sparse functional MLP. Specifically, we consider a curve classification problem. For each subject, we have a variable of interest measured at multiple random times within time range \mathcal{T} . Each subject has an associated group label. The problem is to build a model to predict the group label using the repeatedly observed feature within time window \mathcal{T} .

The synthetic curves and labels are generated as follows. There are two distinct groups, i.e., $g = 1, 2$. The j -th observation of the i -th subject in group g is denoted as $Z_{g,i,j}$, which is generated using

**Fig. 5.** Data pre-processing in sequential learning models.

$Z_j^{(g,i)} = X^{(g,i)}(T_j^{(g,i)}) + \epsilon_{g,i,j}$ Eq. (4) and the Karhunen-Loève expansion $X^{(g,i)}(T_j^{(g,i)}) = \mu_g(T_j^{(g,i)}) + \sum_{p=1}^{\infty} \xi_{g,i,p} \phi_p(T_j^{(g,i)})$ [7], with $\xi_{g,i,p} \sim N(0, \lambda_p)$, for $i = 1, \dots, N_g, j = 1, \dots, M_{g,i}$. Without loss of generality, we set time window $\mathcal{T} = [0, 1]$. The number of subjects in both groups are $N_1 = N_2 = 300$. The number of observation on each curve is 10, i.e., $M_{g,i} = M = 10$. Given $M_{g,i}$, the observing times are i.i.d samples from the Uniform distribution within $[0, 1]$. The p -th eigenfunction $\phi_p(t)$ equals $\sqrt{2} \sin(p\pi t)$ for $p = 1, \dots, \infty$. The first four eigenvalues are $(0.1, 0.045, 0.01, 0.001)$ and $\lambda_p = 0$ for $p > 4$. The mean functions are $\mu_1(t) = \sin(4\pi t)$ and $\mu_2(t) = -\sin(4\pi t)$. The standard deviation of the i.i.d random errors is 0.3. Two randomly selected subjects from each of the two groups are visualized in Fig. 7a and b. In each plot, the true curve $X_{g,i}(t)$ is the brown dashed line and the observations to be modeled are the black dots. We can see that the data on each curve is sparse, and the big gaps between observations prohibit the pre-smoothing step in LSTM and the dense functional MLP to successfully recover the two-cycle sine functions using the limited amount of data available. Our proposed sparse functional MLP is specifically designed to handle this kind of scenarios.

Cubic spline, the Gaussian process regression, and functional data analysis based PACE interpolations are used in this experiment to get data readings of the same interval. We then feed these interpolation data readings into an LSTM network. This LSTM uses two layers LSTM (32, 64) and one layer of neural network (8). Note that grid search is used to tune the hyperparameters in LSTM. Fig. 9 shows the achieved interpolation at 100 points ($M = 100$) within the period for one randomly selected subject in group 1. It can be seen that cubic splines and the Gaussian process regression produce biased estimates of the true curve, while the result of PACE is consistent. The same observation is shown by the RMSE for all the 300 curves in group 1 in the first table in Table 3. Given the simulation result, in real practice, we recommend trying PACE as an alternative curve fitting method when the measurements in each time series are sparse and irregular. The leave-one-out cross validation results are given by the first three rows of Table 3.

To implement our proposed sparse FMLP, we first estimate all the required components in the proposed sparse FMLP. The first three dimensions of the achieved eigen projection scores, i.e.,

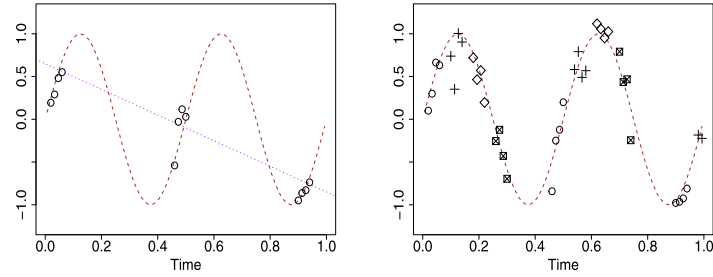


Fig. 6. Left: An example when doing interpolation based on a single subject is problematic. Brown line is the underlying true curve, black dots are the observations to be used, and purple lines is the biased pre-smoothing results. Right: Visualization of the benefit of performing interpolate using combined the observations from different subjects.

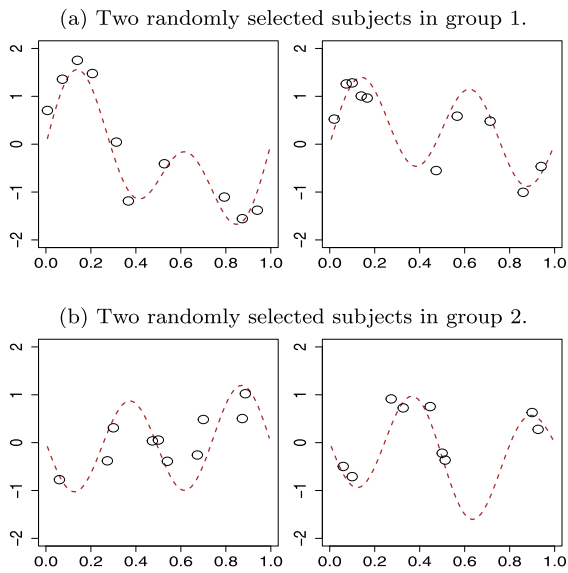


Fig. 7. Visualizations of simulated data in Section 5.2.1.

$E[\eta_{g,i,1}|\mathbf{Z}_{g,i}]$, $E[\eta_{g,i,2}|\mathbf{Z}_{g,i}]$, and $E[\eta_{g,i,3}|\mathbf{Z}_{g,i}]$ for $g = 1, 2$; $i = 1, \dots, N_g$, are visualized in Fig. 8. In the plots, two different types of dots correspond to the two different groups. As shown by the plots, the extracted eigen projection scores better distinguishes the two evolution curves $\sin(4\pi t)$ and $-\sin(4\pi t)$. We specify the weight function of the k -th functional neuron as $W_k(\beta_k, t) = \sum_{p=1}^P \beta_{k,p} \hat{\phi}_p(t)$, with $\hat{\phi}_p(t)$ being the p -th estimated eigenfunction. The architecture of the sparse functional MLP is that there are 4 functional neurons in the first layer followed by a layer with two numerical neurons. The activation functions in both layers are logistic function. The leave-one-out cross-validation results are given by the last two rows of Table 3.

Based on the two tables in Table 3, we have the following observations. First, the functional data based interpolation that jointly

consider the data from all samples significantly outperforms the alternative techniques. Second, for this example, the performance of our sparse functional MLP under $P = 2$ and $P = 3$ is comparable, with the result under $P = 3$ being slightly worse. This makes sense because the first two and three components contain a comparable amount of information in the curves, according to our setting (i.e., eigenvalues are (0.1, 0.045, 0.01, 0.001)). Also, the reason why $P = 3$ slightly worse is that the third component with a very small eigenvalue and can be considered as additional noises introduced to the first two components. Third, the performances of LSTM with different interpolations are similar, although the fitting performance of PACE is significantly better than the other approaches. We think this is because the interpolation within each group might be biased in a similar way for all subjects, such that the two groups of curves are still distinguishable. Last but not least, the performance of LSTM with PACE and the proposed sparse FMLP is comparable. We think this is because the distinction between the two groups of data is large and different models tend to have similar performances.

5.2.2. Prediction of PBC patient's long-term survival

In this section, we consider the problem of predicting the long-term survival of patients with primary biliary cirrhosis (PBC) using their serum bilirubin measurements (in mg/dl) at the beginning period of the study. This enables the doctors to get early warnings and to take corresponding actions to increase patient's survival possibility. The PBC data set we used are results of a Mayo Clinic trial from 1974 to 1984. This data set is publicly available in an R package called 'survival' and has been investigated by numerous researchers including [15].

We consider the following problem setting. We use the patient's bilirubin measurements, which is known to be an important indicator of the presence of chronic liver cirrhosis, within the first 910 days of the study to predict whether the patients survive beyond 10 years after entering the study. There are 260 patients included in the analysis, with 84 died between 910 and 3650 days and 176 being alive after 10 years. Following [15], the bilirubin measurements are log-transformed. The number of bilirubin measurements per patient within the 910 days ranges from 1 to 5, with

Table 3

The interpolation accuracy of different methods for time series in group 1 and leave-one-out cross-validation results for synthetic curve classification task in Section 5.2.1.

Interpolation method				RMSE
Cubic spline				0.562
Gaussian process regression				0.440
PACE				0.169
Model	Spec	# of samples	Architecture	Accuracy
LSTM _{spline}	$M = 100$	600	L(32,64) N(8)	99.4%
LSTM _{GP}	$M = 100$	600	L(32,64) N(8)	99.5%
LSTM _{PACE}	$M = 100$	600	L(32,64) N(8)	99.3%
Sparse FMLP	$P = 2$	600	F(4) N(2)	99.5%
Sparse FMLP	$P = 3$	600	F(4) N(2)	99.3%

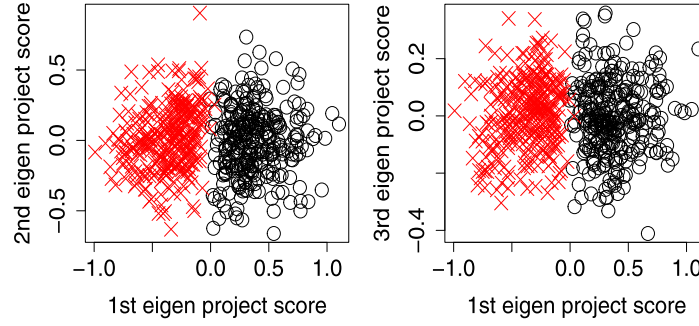


Fig. 8. Extracted features from sparse FMLP($P = 3$). Black dots represent group 1 and red dots represent group 2.

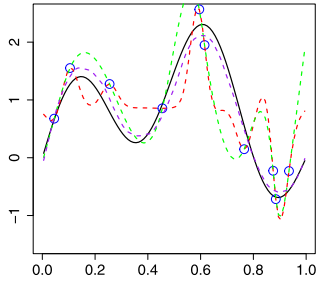
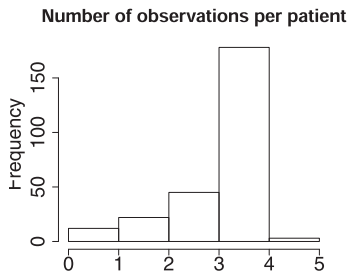


Fig. 9. Comparison of interpolation results for a randomly selected sample. The blue dots are the sparse observations available in the data set. The black line is the ground truth. The red, green, and purple lines respectively corresponding to the Gaussian process regression, cubic B-spline and PACE.

histogram given in Fig. 10a. The sparse observations of two randomly selected patients are plotted as the black dots in Fig. 11. Given these two plots, it can be seen that the functional data is sparse, and the number of observations and the observing time are not the same across patients.

Our proposed sparse functional MLP is implemented in the same fashion as described in the previous numerical experiment. The number of projection is chosen through the leave-one-curve-out cross-validation approach described in [24]. The selected $\hat{P} = 2$. The estimated eigen-functions using the restricted maximum likelihood estimate method in [33], denoted as $\hat{\phi}_p(t)$ for $p = 1, 2$, are given in Fig. 10b. The best predicted curves using $\hat{X}_i(t) = \sum_{p=1}^{\hat{P}} \hat{E}[\eta_{i,p} | \mathbf{Z}_i] \hat{\phi}_p(t)$ for two randomly selected patients are given in Fig. 11. In each of the two plots in Fig. 11, the observed dots are closely distributed around the predicted log(bilirubin) curve $\hat{X}_i(t)$ with some random errors, which visually justified the validity of our curve approximator in Eqs. (13) and (14). The archi-

(a) Histogram of number of observation per patient within the first 910 days in the PBC study.



(b) Estimated eigenfunctions with first two highest eigenvalues from sparse FMLP.

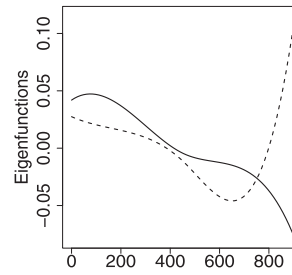


Fig. 10. Some results for the PBC long-term survival prediction study.

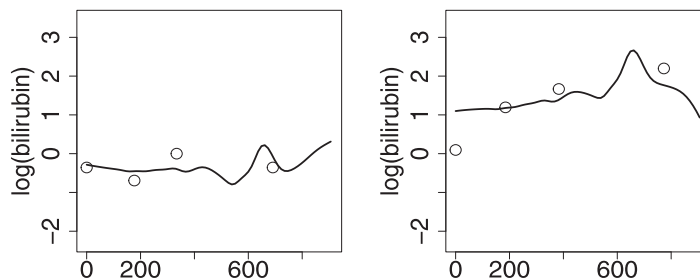


Fig. 11. The raw data and the recovered log(bilirubin) curve using sparse FMLP for two randomly selected patients.

Table 4

Leave-one-out cross-validation results for PBC long-term survival prediction using sparse FMLP.

Classified	True	
	Survived	Died
Survived	151	25
Died	45	39

structure of the sparse functional MLP we used is that there are four functional neurons on the first layer, i.e., $K = 4$ in Eqs. (13) and (14), followed by another layer of two numerical neurons. The activation functions in both layers are logistic function. The weight function $V_k(\beta_k, t)$ for the k -th functional neuron in the first layer

$$W_k(\beta_k, t) = \sum_{p=1}^{\hat{p}} \beta_{k,p} \hat{\phi}_p(t), \quad (28)$$

with $\hat{\phi}_p(t)$ being the p -th estimated eigenfunction. The leave-one-out cross-validation results using the specifications mentioned above are given in Table 4. The overall accuracy is 73.08%. We also implemented the cubic B-spline interpolation plus LSTM strategy described in the previous subsection. The accuracy of LSTM with $M = 910$ is 67%, which is lower than our proposed method.

5.2.3. Predicting remaining useful life for aircraft engines

In this subsection, we re-consider the RUL prediction problem in Section 4. The C-MAPSS data set is a simulated data set without any irregularity and missing values. To mimic real scenarios where there are usually a certain level of irregularity in time series trajectories, we sparsify the data set by randomly keep a certain percentage of the raw data (30%, 50% or 100%) for each of the engines in the training data. Note that the sampled timestamps are different for different variables and subject. The last observation of each engine is always kept to indicate the failure time.

For each of the 21 sensor variables, the correlations among the projection scores from separately conducting FPCA on different sensors are visualized in Fig. 12. It can be seen that there are strong correlations among some of the sensors. Therefore, there will be some benefits of considering the multivariate FPCA in Section 3.1. The architecture of the sparse functional MLP ('Sparse FMLP') and the multivariate sparse functional MLP ('Sparse MFMLP') we used is that there are four functional neurons on the first layer, i.e., $K = 4$ in (13) and (14), followed by another layer of two numerical neurons. The activation functions in both layers are logistic function. And the weight function is the same as Eq. (28). The basis functions in Eq. (28) are the eigenfunction from separate FPCA for 'Sparse MFMLP' and the eigenfunctions from multivariate FPCA for 'Sparse MFMLP'. The baselines respectively utilize the cubic spline, Gaussian process regression, PACE and multivariate PACE to prepare data for LSTM ('LSTM_{spline}', 'LSTM_{GP}', 'LSTM_{PACE}', 'LSTM_{MPACE}'). The RMSE are summarized in Table 5. The plots of RMSE over different level of data sparsity for the considered approaches are provided in Fig. 13. The plot is produced with the result for data set 'FD001'.

Here a summary of the major observations. First, the performance of LSTM varies across different interpolation techniques. What's more, the LSTM with FDA type of interpolation approaches significantly outperforms the conventional data interpolation techniques. This is because the fitting error of the FDA type of interpolations is smaller. Second, the proposed models in the paper outperform all the LSTM based approaches. Specifically, the proposed models is more accurate the 'LSTM_{PACE}' and 'LSTM_{MPACE}'. This indicates that the functional way of modeling is better than the sequential calculation in LSTM for this specific problem. Third, the multivariate FMLP performs slightly better than the FMLP based on separate PACE. This is consistent with our intuition, as there are some correlations among the projection scores and 'Sparse MFMLP' is proposed to handle such correlations. Note that the magnitude of improvement is not large. We think it is reason-

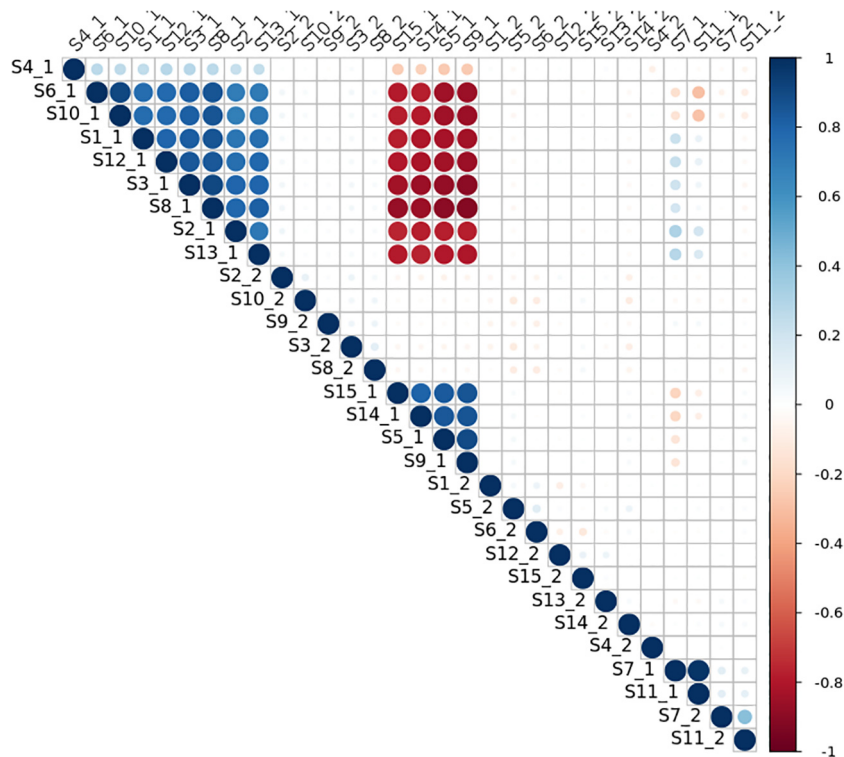


Fig. 12. Correlation matrix for the projection scores of sensors calculated from the univariate PACE.

Table 5
RMSE comparison on C-MAPSS data.

Train	Test	Model	FD001	FD002	FD003	FD004
10%	100%	LSTM _{Spline}	40.54	43.76	39.81	49.02
		LSTM _{GP}	34.22	36.78	35.89	34.33
		LSTM _{PACE}	22.03	23.69	21.89	23.75
		LSTM _{MPACE}	21.85	23.10	21.10	23.54
		Sparse FMLP	20.81	19.53	19.83	19.39
		Sparse MFMLP	20.27	18.67	19.63	19.03
30%	100%	LSTM _{Spline}	40.54	43.76	39.81	49.02
		LSTM _{GP}	29.98	30.28	30.42	28.91
		LSTM _{PACE}	17.79	18.94	20.26	21.64
		LSTM _{MPACE}	17.48	18.36	20.09	21.56
		Sparse FMLP	17.12	17.45	16.41	18.34
		Sparse MFMLP	16.85	17.08	16.06	18.18
50%	100%	LSTM _{Spline}	40.57	49.56	39.82	44.35
		LSTM _{GP}	26.45	24.76	26.43	25.26
		LSTM _{PACE}	18.85	18.84	20.44	20.77
		LSTM _{MPACE}	17.13	18.52	19.49	20.14
		Sparse FMLP	15.47	17.24	14.63	17.44
		Sparse MFMLP	15.02	16.98	14.41	17.19
100%	100%	LSTM [27]	16.14	24.49	16.18	28.17
		FMLP	13.36	16.62	12.74	17.76
		Sparse FMLP	13.73	17.04	12.75	16.92
		Sparse MFMLP	13.11	16.03	11.97	16.33

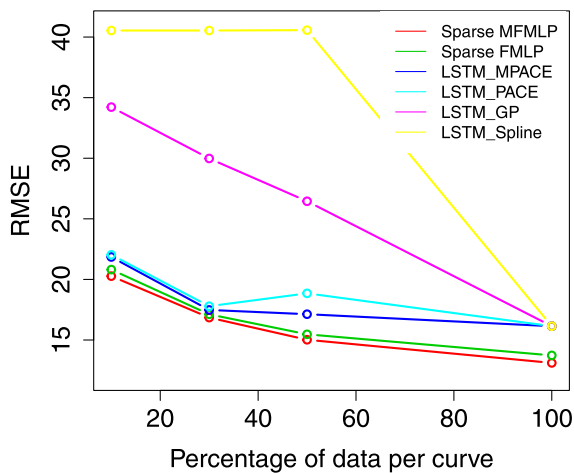


Fig. 13. RMSE under different percentage of data per subject.

able as the correlation exists among small clusters of projections. As another point, under the dense data scenario, i.e., 100% for both training and testing, our proposed sparse FMLP and the dense FMLP [30] has comparable performance. This experimentally justifies their equivalence under dense data circumstance as discussed in Section 3.1. Last, as shown in Fig. 13, the performance of models utilize functional thinking is less affected when the number of data points per curve decreases, compared to models such as 'LSTM_{Spline}', 'LSTM_{GP}'.

6. Conclusion and discussion

In this paper, we focused on the temporal classification/regression problem, the purpose of which is to learn a mathematical mapping from the time series inputs to scalar response, leveraging the temporal dependencies and patterns. In real-world applications, we noticed that two types of data scarcity are frequently encountered: scarcity in terms of small sample sizes and scarcity introduced by sparsely and irregularly observed time series covari-

ates. Noticing the lack of feasible temporal predictive models for sparse time series data in the literature, we proposed two sparse functional MLP ('Sparse FMLP' and 'Sparse MFMLP') to specifically handle this problem. The proposed SFMLP is an extension of the conventional FMLP for densely observed time series data, employing the univariate and multivariate sparse functional principal component analysis. We used mathematically arguments and numerical experiments to evaluate the performance of each candidate model under different types of data scarcity and achieved the following conclusions:

- When the sample size is large and the underlying function that gives rise to the time series observations is non-smooth over time, the sequential learning models are more appropriate models to handle the supervised learning task.
- When the sample size is small and the underlying function is smooth, we expect FMLP or equivalently the proposed SFMLP to outperform the sequential learning in general, since
 1. FMLP requires fewer training samples due to its feed-forward network structure, succinct temporal pattern capturing technique, and its capability of encoding domain expert knowledge.
 2. FMLP is easier to train as the time-varying impact of the covariates on the response is aggregated through integrals of the continuous processes, rather than recursive computations on the individual observations.
 3. FMLP is less restrictive on the data format. In particular, the time series covariates can be regular or irregular. Also, the number of observations as well as the measuring time-stamps can be different across features and subjects.
- The proposed sparse FMLPs are feasible solutions when the individual time series features are sparsely and irregularly evaluated.

CRediT authorship contribution statement

Qiyao Wang: Methodology, Writing - original draft, Software. **Ahmed Farahat:** Supervision, Writing - review & editing, Methodology. **Chetan Gupta:** Conceptualization, Supervision, Project administration, Funding acquisition. **Shuai Zheng:** Writing - review & editing, Software.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

- [1] Q. Wang, S. Zheng, A. Farahat, S. Serita, T. Saeki, C. Gupta, Multilayer perceptron for sparse functional data, in: 2019 International Joint Conference on Neural Networks (IJCNN), IEEE, 2019, pp. 1–10.
- [2] Q. Wang, S. Zheng, A. Farahat, S. Serita, C. Gupta, Remaining useful life estimation using functional data analysis, in: 2019 IEEE International Conference on Prognostics and Health Management (ICPHM), IEEE, 2019, pp. 1–8.
- [3] J.D. Hamilton, *Time Series Analysis*, vol. 2, Princeton New Jersey, 1994..
- [4] T. Mikolov, M. Karafiát, L. Burget, J. Černocký, S. Khudanpur, Recurrent neural network based language model, in: Eleventh Annual Conference of the International Speech Communication Association, 2010.
- [5] J.T. Connor, R.D. Martin, L.E. Atlas, Recurrent neural networks and robust time series prediction, *IEEE Transactions on Neural Networks* 5 (2) (1994) 240–254.
- [6] J. Chung, C. Gulcehre, K. Cho, Y. Bengio, Empirical evaluation of gated recurrent neural networks on sequence modeling, *arXiv preprint arXiv:1412.3555* (2014)..
- [7] J.O. Ramsay, *Functional Data Analysis*, Wiley Online Library, 2006..
- [8] P.E. Castro, W.H. Lawton, E.A. Sylvestre, Principal modes of variation for processes with continuous sample curves 28 (1986) 329–337..
- [9] B.W. Silverman et al., Smoothed functional principal components analysis by choice of norm, *The Annals of Statistics* 24 (1) (1996) 1–24.
- [10] H. Cardot, F. Ferraty, P. Sarda, Spline estimators for the functional linear model 13 (3) (2003) 571–591..
- [11] T. Cai, P. Hall, Prediction in functional linear regression 34 (2006) 2159–2179..
- [12] H.-G. Müller, Y. Zhang, Time-varying functional regression for predicting remaining lifetime distributions from longitudinal trajectories 61 (4) (2005) 1064–1075..
- [13] F. Yao, H.-G. Müller, J.-L. Wang, Functional linear regression analysis for longitudinal data 33 (6) (2005) 2873–2903..
- [14] G.M. James, C.A. Sugar, Clustering for sparsely sampled functional data, *Journal of the American Statistical Association* 98 (462) (2003) 397–408.
- [15] H.-G. Müller, Functional modelling and classification of longitudinal data, *Scandinavian Journal of Statistics* 32 (2) (2005) 223–240.
- [16] H. Cardot, C. Crambes, P. Sarda, Quantile regression when the covariates are functions, *Nonparametric Statistics* 17 (7) (2005) 841–856.
- [17] F. Ferraty, A. Rabhi, P. Vieu, Conditional quantiles for dependent functional data with application to the climatic el ni no phenomenon, *Sankhyā: The Indian Journal of Statistics* (2005) 378–398..
- [18] K. Chen, H.-G. Müller, Conditional quantile analysis when covariates are functions, with application to growth data, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 74 (1) (2012) 67–89.
- [19] A.A. Adebiyi, A.O. Adewumi, C.K. Ayo, Comparison of arima and artificial neural networks models for stock price prediction, *Journal of Applied Mathematics* 2014 (2014).
- [20] S.-L. Ho, M. Xie, T.N. Goh, A comparative study of neural network and box-jenkins arima modeling in time series prediction, *Computers & Industrial Engineering* 42 (2–4) (2002) 371–375.
- [21] S. Hochreiter, J. Schmidhuber, Long short-term memory, *Neural Computation* 9 (8) (1997) 1735–1780.
- [22] F. Rossi, B. Conan-Guez, Functional multi-layer perceptron: a non-linear tool for functional data analysis, *Neural Networks* 18 (1) (2005) 45–60.
- [23] F. Rossi, N. Delannay, B. Conan-Guez, M. Verleysen, Representation of functional data in neural networks, *Neurocomputing* 64 (2005) 183–210.
- [24] F. Yao, H.-G. Müller, J.-L. Wang, Functional data analysis for sparse longitudinal data, *Journal of the American Statistical Association* 100 (470) (2005) 577–590.
- [25] C. Happ, S. Greven, Multivariate functional principal component analysis for data observed on different (dimensional) domains, *Journal of the American Statistical Association* 113 (522) (2018) 649–659.
- [26] J.-M. Chiou, Y.-T. Chen, Y.-F. Yang, Multivariate functional principal component analysis: A normalization approach, *Statistica Sinica* (2014) 1571–1596.
- [27] S. Zheng, K. Ristovski, A. Farahat, C. Gupta, Long short-term memory network for remaining useful life estimation, in: 2017 IEEE International Conference on Prognostics and Health Management (ICPHM), IEEE, 2017, pp. 88–95.
- [28] F. Yao, H.-G. Müller, Functional quadratic regression, *Biometrika* 97 (1) (2010) 49–64.
- [29] S.K. Pal, S. Mitra, Multilayer perceptron, fuzzy sets, classification, 1992..
- [30] F. Rossi, B. Conan-Guez, F. Fleuret, Functional data analysis with multi layer perceptrons, in: *Proceedings of IJCNN*, Citeseer, 2002, pp. 2843–2848..
- [31] P. Hall, H.-G. Müller, J.-L. Wang, Properties of principal component methods for functional and longitudinal data analysis, *The Annals of Statistics* (2006) 1493–1517.
- [32] G.-M. Pomann, A.-M. Staicu, S. Ghosh, A two sample distribution-free test for functional data with application to a diffusion tensor imaging study of multiple sclerosis, *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 65 (3) (2016) 395–414.
- [33] J. Peng, D. Paul, A geometric approach to maximum likelihood estimation of the functional principal components from sparse longitudinal data, *Journal of Computational and Graphical Statistics* 18 (4) (2009) 995–1015.
- [34] G.M. James, T.J. Hastie, C.A. Sugar, Principal component models for sparse functional data, *Biometrika* (2000) 587–602.
- [35] A. Saxena, K. Goebel, Phm08 challenge data set, NASA Ames Prognostics Data Repository (<http://ti.arc.nasa.gov/project/prognostic-data-repository>), NASA Ames Research Center, Moffett Field, CA (consulted 2014-02-15), 2008..
- [36] S.-J. Wu, N. Gebräel, M.A. Lawley, Y. Yih, A neural network integrated decision support system for condition-based optimal predictive maintenance policy, *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* 37 (2) (2007) 226–236.
- [37] Z. Tian, An artificial neural network method for remaining useful life prediction of equipment subject to condition monitoring, *Journal of Intelligent Manufacturing* 23 (2) (2012) 227–237.
- [38] G.S. Babu, P. Zhao, X.-L. Li, Deep convolutional neural network based regression approach for estimation of remaining useful life, in: *International Conference on Database Systems for Advanced Applications*, Springer, 2016, pp. 214–228.



Qiyao Wang received her Ph.D degree in Statistics from the University of Pittsburgh. Dr. Wang is currently a senior research scientist at the Industrial AI Laboratory at Hitachi America, Ltd. Her research interests are time series data analysis, and deep learning techniques.



Ahmed K. Farahat holds a Ph.D. degree from the University of Waterloo in Canada and M.Sc. and B.Sc. degrees from Cairo University in Egypt, all in Computer Engineering. Dr. Farahat is currently a principal research scientist at the Industrial AI Laboratory at Hitachi America, Ltd. Previously, Dr. Farahat was a research scientist at the Big Data Laboratory at Hitachi America, Ltd. (2014–2017). He also worked as a postdoctoral fellow at the University of Waterloo (2013–2014), a research assistant at the University of Waterloo (2007–2012) and a research engineer at IBM Egypt (2005–2007). His research interests lie in the areas of machine learning and data mining, and their applications to industrial data analytics. Dr. Farahat is the recipient of the Best Paper Award Runner-Up at the 2013 IEEE International Conference on Data Mining (ICDM), and two best paper awards in specialized workshops at ICDM09 and NIPS013.



Chetan Gupta is the Vice President, and manages the Industrial AI Lab at Hitachi America R&D. He has more than 15 years of experience in analytics, AI, big data, and related domains. Over his career he has worked both as a machine learning data scientist as well as in designing systems and architectures for big data applications. At Hitachi, he manages a large team of data scientists, architects and developers that is engaged in developing cutting edge solutions and opening new frontiers in the area of industrial analytics. His team builds fundamental horizontal technologies that are then used to build solutions for industry specific verticals. He has led efforts to build horizontal solutions in predictive maintenance, quality, operations monitoring and control, and for verticals such as mobility, mining, building energy management systems, etc. Over the years Chetan has led multiple research and development teams, and mentored young researchers. He has close to 50 patents either granted or under review and more than 40 publications in the area of data mining/machine learning, data stream systems, complex event processing, workload management, etc. Chetan has a Ph.D. in Mathematics and M.S. in Math-

ematical Computer Science and Chemical Engineering from University of Illinois, Chicago.



Shuai Zheng received the PhD degree in Computer Science from the University of Texas at Arlington in 2017. His main research interests are machine learning, data mining, and cloud computing.