IDENTIFICATION AND ESTIMATION OF GENERAL NONLINEAR STRUCTURED LATENT FACTOR MODEL FOR FUNCTIONAL DATA

Xiaorui Wang[1,2], Yimang Zhang[2] and Jian Qing Shi[2]

[1] *Nanjing University of Information Science and Technology*

[2] *Southern University of Science and Technology*

Abstract: Nonlinear structured latent factor model captures the relationship between observed variables and latent variables in a nonlinear way, offering greater flexibility compared to a linear factor model. Functional data characterizes features of data that vary continuously over time or space and is widely applied across various fields. This paper proposes a nonlinear structured latent factor model for functional data. We consider correlations for the latent factor to account for the dependence in functional data at different time points. The structured identifiability of latent factors is studied to ensure uniqueness, thereby allowing these factors to have a physical interpretation. A Gaussian process (GP) prior is utilized to estimate the unknown nonlinear link functions. To improve computational efficiency, an efficient algorithm is developed by using the nearest neighbor Gaussian process (NNGP). The consistency of the latent factors and the unknown parameters, as well as the posterior consistency of the unknown link functions, was established. Simulation studies were conducted to demonstrate the finite-sample performance and the flexibility of the proposed model, and the significant computational time savings achieved by NNGP compared to GP. The method was applied to analyse the gait data collected in our laboratory for early detection of neurodegenerative diseases.

Xiaorui Wang and Yimang Zhang contributed equally.
Corresponding author, email: shijq@sustech.edu.cn
Jian Qing Shi 's ORCID ID is 0000-0002-2924-1137

Key words and phrases: Nonlinear structured latent factor model; Identifiability; Gaussian process prior; Nearest neighbor Gaussian process;

1. Introduction

Latent factor model (Spearman, 1904) is a commonly used statistical tool for multivariate data analysis, which describes the common dependence among multiple observed variables using only a few latent variables. Compared to another commonly used tool for handling multivariate data, principal components analysis (PCA, McCabe, 1984), it provides better physical interpretability. For instance, in the gait dataset collected in our laboratory, numerous gait features may be summarized into four latent factors: pace, rhythm, symmetry, and variability (SD). The pace factor directly correlates with walking speed, rhythm reflects the regularity of gait, symmetry measures bilateral symmetry between two feet, and SD indicates gait stability and consistency. Latent factor analysis greatly improves the interpretability of the data, enabling researchers to gain deeper insights into complex gait patterns. The details of the dataset will be discussed in Section 5. In practice, factor analysis has applications across diverse fields, including malfunction detection, diagnostics, psychometrics, and medical research.

Identifiability is a key property in latent factor models, and is essential for substantiating the physical interpretation of latent factors. An identifiable model allows for the unique extraction of specific latent factors from observed data, thus enriching the comprehension of the interactions between observed variables and latent factors. In this paper, we consider the identifiability of the structured latent factors. The term "structured" signifies that the model incorporates specific patterns or constraints on associations between latent factors and observed variables. Traditionally, this structure is often enforced by setting certain factor loadings to zero, implying that each observed variable is influenced by specific latent factors only. In a more recent study, Chen et al. (2020) explored the identifiability and estimation of structured factor models with predefined generalized link functions, including both linear and logistic models as special cases. Zhang et al. (2025) studied the identifiability and estimation of nonlinear structured latent factor models with an unknown link function from a Bayesian perspective.

Existing latent factor models assume that observed variables are related to a set of linear combinations of latent factors through either a linear function, a known nonlinear link function, or an unknown link function similar to a multiple index model (Chen et al., 2020; Zhang et al., 2025). However, real-world data often exhibit unknown and complex nonlinear relationships. In this paper, we propose a novel general nonlinear latent factor model, similar to a multivariate nonparametreic model. Additionally, we examine the methods within the context of functional data analysis, where data collected at different time points are dependent. This scenario often occurs in various practical applications, such as stock price fluctuations over time, human motion trajectories, and variations in patients' brainwaves. Gait data also falls under this category. Several studies have been conducted on latent factor models for functional data. Tu et al. (2014) and Gu et al. (2024) adopted the dynamic mode decomposition method, which is a popular approach of linearizing the one-step-ahead transition operator of nonlinear dynamical systems and reconstructing the dynamics by eigenpairs of the linear mapping matrix. Another type of method is to model each latent factor using a Gaussian process in coregionalization models (Gu and Shen, 2020; Lin et al., 2025). However, none of them considered the identifiability of latent factors, resulting in a lack of physical interpretability. This paper addresses this issue and considers a more general model under unknown link functions.

Gaussian processes (GPs) and related methods have been widely applied to build nonlinear model, see e.g. Shi and Choi (2011). Lawrence (2005) proposed the Gaussian process latent variable model (GPLVM), which utilizes the concept of Gaussian process regression (GPR) models for nonparametric and nonlinear dimensionality reduction. Their approach removed the irrelevant relationships between the observed variables and latent factors, offering a more flexible framework compared to traditional methods such as functional PCA. Wang et al. (2005) studied the Gaussian process dynamical models (GPDM), assuming that samples are indexed by t (time series) but remain independent and identically distributed. In our analysis of dependent gait data, we aim to reduce the dimensionality of the observed data by identifying low-dimensional interpretable factors. This necessitates exploring nonlinear structured latent factor models for dependent data, a topic that has not received sufficient attention in the literature.

One of primary challenges in implementing GPR, as well as in GPLVM and GPDM, is the significant computational cost associated with the calculation of the inverse of covariance matrices, which is $O(N^3)$ and N is the number of time points. Damianou et al. (2011) tackled this issue by using variational approximations for GPDM; however, this method fails to ensure consistent parameter estimates as the sample size increases. Liu et al. (2018b) introduced scalable Gaussian processes designed to enhance the scalability of full GP models while maintaining good prediction quality for large datasets. Another strategy for reducing computation time is local approximation, which only needs a local subset of size m_0 (where $m_0 \ll N$) data points (Gramacy and Lee, 2008; Gramacy, 2016). Among these approaches, the nearest neighbor Gaussian process (NNGP, Datta et al., 2016) is particularly promising due to its ability to balance computational efficiency with model accuracy. The NNGP method effectively combines the principles of Gaussian processes and nearest neighbor techniques, concentrating on local data subsets and utilizing nearest neighbor approximations to significantly lower the computational complexity of full GP models, while still delivering strong predictive performance. Consequently, NNGP is a valuable tool for researchers and practitioners, especially in scenarios involving large datasets and complex, nonlinear relationships. More discussion can be found in Liu et al. (2018a); Wu et al. (2022); Coube-Sisqueille and Liquet (2022); Saha et al. (2022); Villarraga and Daziano (2025) among others. A variational version of this method will be developed and used in this paper.

In this paper, we delve into the identifiability and estimation of the nonlinear structured latent factor model for functional data. The efficiency of our proposed methodology is rigorously substantiated through both asymptotic properties and simulation studies. In contrast to the extant literature, our work delineates several notable contributions. Firstly, we introduce a nonlinear structured latent factor model with an unknown link function for functional data. This transition from linear to nonlinear factor analysis methodologies signifies a pivotal evolution in the field, enabling the capture of more complex and nuanced relationships within functional data. Secondly, compared to conventional latent factor models, we account for the dependency among observations, allowing us to explore the underlying latent structure of functional data. Thirdly, we apply nonparametric Bayesian methods to latent factor models, with a specific focus on employing the GPR approach for estimating unknown functions. Unlike the GPLVM, which fails to offer a lucid substantive interpretation of latent factors, our approach not only establishes the identifiability of these factors but also provides them with a clear and meaningful interpretation. In addition, to address computational challenges, we incorporate the concept of the NNGP.

The rest of the paper is organized as follows. In Section 2, we introduce the nonlinear structured latent factor model for functional data, and then discuss the structured identifiability of the latent factors as well as the estimation of unknown parameters, latent factors, and nonlinear function. In Section 3, We establish the consistency of the proposed latent factor estimates and unknown parameters, along with the posterior consistency of the estimator of the unknown link function. Simulation studies are conducted to investigate the finite-sample performance in Section 4. In Section 5, the proposed method is applied to a gait dataset collected in our laboratory. Finally, concluding remarks and directions for further studies are provided in Section 6. All technical proofs are given in the online Supplementary Materials.

2. Model and Methodology

2.1 General Nonlinear Structured Latent Factor Model for Functional Data

Suppose that $\boldsymbol{Y}(t_i) = (Y_1(t_i), \ldots, Y_J(t_i))^\top \in \mathcal{R}^J$ is a J-dimensional functional manifest (observed) variable for $i = 1, \ldots, N$. The component $Y_j(t_i)$ denotes the value of the j-th manifest variable at the i-th time point. For each time point t_i, we assume the manifest variable is associated with a K-dimensional latent vector, denoted as $\boldsymbol{x}(t_i) = (x_1(t_i), \ldots, x_K(t_i))^\top$. In the gait dataset of Section 5, $\boldsymbol{Y}(t_i)$ represents gait speed, step length, time for one stride, one step or one stance, and so on, while $\boldsymbol{x}(t_i)$ represents pace, rhythm, asymmetry, variability (SD).

Each component $Y_j(t_i)$ usually depends on one or more latent variables. To model such a structure, we defined the following general nonlinear structured latent factor model (GNSLFM):

$$Y_j(t_i) = f_j(\boldsymbol{x}^\top(t_i)\boldsymbol{R}_j) + \varepsilon_j(t_i),\ \varepsilon_j(t_i) \sim \mathcal{N}(0, \sigma^2), j = 1, \ldots, J, i = 1, \ldots, N, \qquad (2.1)$$

where f_j's are unknown multivariate link function, $\varepsilon_j(t_i)$'s are i.i.d. random errors, $\boldsymbol{R}_j = \mathrm{diag}\{r_{j1}, \ldots, r_{jK}\}$ is a $K \times K$ diagonal matrix, and $r_{jk} = 1$ means the j-th manifest variable is associated with the latent variable $x_k(t_i)$ and $r_{jk} = 0$ otherwise.

Remark 1. Model (2.1) uses a multivariate nonparametric function to describe the relationship between manifest variables and latent variables. Chen et al. (2020) and Zhang et al. (2025) established the relationship between $Y_j(t_i)$ and a set of linear combinations of $\boldsymbol{x}(t_i)$ using a known or unknown nonlinear link function similar to a multi-index model. Instead, Model (2.1) establishes a very general factor model. In addition, we assume normal random errors to facilitate the subsequent use of Gaussian process regression (GPR) for estimating the unknown link functions f_j, which is a common assumption in the GPR literature. When the data contain outliers or exhibit heavy-tailed distributions, alternative assumptions such as the t-distribution can be used (Wang and Shi, 2014; Wang et al., 2017, 2021); however, this makes the posterior distribution analytically intractable and necessitates the use of numerical methods for inference. In this paper, we focus on developing a tractable modeling and estimation framework, and leave the investigation of such robust extensions to future work.

Remark 2. The role of $\boldsymbol{R}_j$ is to select the set of latent factors associated with the manifest variable $Y_j(t_i)$. In this paper, we assume that $\boldsymbol{R}_j$ is pre-specified according to the identifiability conditions of the latent factors, which will be discussed later. Regarding the automatic selection of $\mathbf{R}_j$, we have investigated this problem in Zhang (2025) and provided practical guidance for settings with a small number of latent factors ($K = 2, 3, 4$), which holds true in most applications. While these approaches work well in such low-dimensional cases, a general automatic procedure remains challenging due to the structural identifiability constraints of our framework. We therefore leave the development of a fully general method for selecting $\mathbf{R}_j$ for future research.

Let $\boldsymbol{x}^{(j)}(t_i)$ be the subset of latent factors associated with $Y_j(t_i)$, i.e., selection of the elements with $r_{jk} = 1$ for $k = 1, \ldots, K$, then the model can be rewritten as

$$Y_j(t_i) = f_j(\boldsymbol{x}^{(j)}(t_i)) + \varepsilon_j(t_i). \tag{2.2}$$

The structure of the model in (2.1) or (2.2) is illustrated by an example in Figure 1. Here we consider two latent factors ($K = 2$) and $J = 6$, the structure of the model is defined by $\boldsymbol{R}_j$, where $\boldsymbol{R}_j = \mathrm{diag}\{1, 0\}$ for $j = 1, 2, 3$; $\boldsymbol{R}_j = \mathrm{diag}\{1, 1\}$ for $j = 4$ and $\boldsymbol{R}_j = \mathrm{diag}\{0, 1\}$ for $j = 5, 6$. Thus, $\boldsymbol{x}^{(1)}(t_i) = \boldsymbol{x}^{(2)}(t_i) = \boldsymbol{x}^{(3)}(t_i) = (x_1(t_i), 0)$, $\boldsymbol{x}^{(4)}(t_i) = (x_1(t_i), x_2(t_i))$, $\boldsymbol{x}^{(5)}(t_i) = \boldsymbol{x}^{(6)}(t_i) = (0, x_2(t_i))$. Based on the conditions discussed in Theorem 1, this structure can guarantee the identifiability of $x_1(t)$ and $x_2(t)$ for each t.

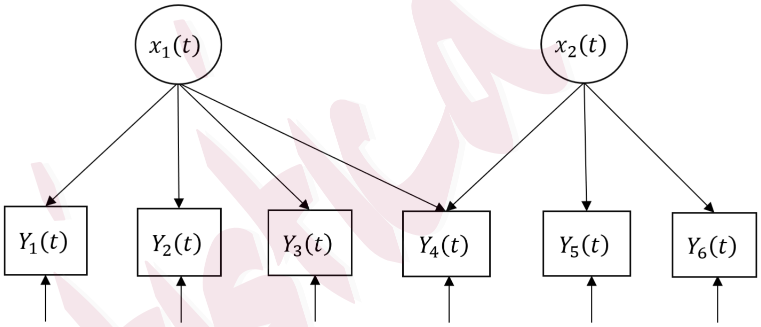

Figure 1: Model structure— Latent factor model

For latent factor models, existing literature typically assumes that the latent factors are independent across different time points t_i, $i = 1, \ldots, N$. However, in practice, this assumption is often too strong. In this paper, we consider $\boldsymbol{Y}(t_i)$ as functional data, i.e., $\boldsymbol{Y}(t_i)'s$ are dependent at different data time $t_i's$. Consequently, $\boldsymbol{x}(t_i)'s$ are also dependent, which is commonly encountered in practical applications. To describe the internal dependence within each factor across different values of t, we may assume each curve $x_k(\cdot)$ in the latent space follows a GP process:

$$x_k(\cdot) \sim \mathcal{GP}_k\left(0, \boldsymbol{\Sigma}_k(\cdot, \cdot; \boldsymbol{\Theta}_{xk})\right) \text{ for } k = 1, \ldots, K, \tag{2.3}$$

with squared exponential covariance function, i.e., the (i,l)-th element of $\boldsymbol{\Sigma}_k$ is given by

$$\mathrm{Cov}(x_k(t_i), x_k(t_l)) = v_{xk} \exp\left\{-\frac{1}{2} w_{xk}(t_i - t_l)^2\right\}, \tag{2.4}$$

where the kernel-parameters $\boldsymbol{\Theta}_{xk} = (v_{xk}, w_{xk})^\top$. Time series models, such as $AR(1)$ model, can also be used to describe the dependence between different time points. The $AR(1)$ model is a parametric model that assumes a linear dependency structure, limited to the immediate preceding time point. In contrast, the GP is a nonparametric model that allows for the flexible modeling of arbitrarily complex dependency structures, not restricted to linearity.

2.2 Structured Identifiability

Identifiability is crucial in structured latent factor models to provide meaningful interpretation of the latent factors. Under Model (2.2), it is well known that the latent factor is not identifiable if no constraints is applied to $\boldsymbol{R}_j$. For positive integers K, N and J, let $\boldsymbol{r}_j = (r_{j1}, \ldots, r_{jK})^\top, j = 1, \ldots, J$ be vectors in $\{0,1\}^K$, and define the structured index matrix as $\boldsymbol{Q}^\top = (\boldsymbol{r}_1, \ldots, \boldsymbol{r}_J) \in \mathbb{R}^{K \times J}$. To ensure the identifiability of the latent factors, we will impose zero constraints on some elements of $\boldsymbol{Q}$.

Here we consider the cases with $N \to \infty$, and the identifiability of the k-th latent factor $\boldsymbol{X}_{[k]} = (x_k(t_1), x_k(t_2), \ldots)^\top \in \mathbb{R}^{\mathbb{Z}_+}$ is equivalent to the identification of the direction of an infinite dimensional vector. Define the following as

$$\sin_+ \angle(\boldsymbol{u}, \boldsymbol{v}) = \limsup_{N\to\infty} \sin \angle \left(\boldsymbol{u}_{[1:N]}, \boldsymbol{v}_{[1:N]}\right),$$

to quantify the angle between two vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ where $\boldsymbol{u} = (u_1, u_2, \ldots)^\top, \boldsymbol{v} = (v_1, v_2, \ldots)^\top \in \mathbb{R}^{\mathbb{Z}_+}$ are two vectors with countably infinite components. When $\sin_+ \angle(\boldsymbol{u}, \boldsymbol{v})$ is 0, we say the angle between $\boldsymbol{u}$ and $\boldsymbol{v}$ is 0. Before presenting the structured identifiability of the latent factor in Model (2.2), we will first introduce the definition of structured identifiability.

Definition 1 (Structured identifiability of a latent factor). Consider the k-th latent factor, where $k \in \{1, \ldots, K\}$, and a nonempty parameter space $\mathcal{S} \subset \mathbb{R}^{\mathbb{Z}_+ \times \{1,\ldots,K\}}$ for $\boldsymbol{X}$. We say the k-th latent factor is structurally identifiable in the parameter space $\mathcal{S}$ if for any $\boldsymbol{X}, \boldsymbol{X}' \in \mathcal{S}, P_{\boldsymbol{X}} = P_{\boldsymbol{X}'}$ implies $\sin_+ \angle\left(\boldsymbol{X}_{[k]}, \boldsymbol{X}'_{[k]}\right) = 0$, where $P_{\boldsymbol{X}}$ is the probability distribution of $\{Y_j(t_i), i \in \mathbb{Z}_+, j \in \{1, \ldots, J\}\}$, given factor scores $\boldsymbol{X}$, structured index matrix $\boldsymbol{Q}$ and link function f_j.

For $\Delta \subset \{1, \ldots, K\}$, denote $R_{\boldsymbol{Q}}(\Delta) = \{j : r_{jk} = 1, \text{ if } k \in \Delta \text{ and } r_{jk} = 0, \text{ if } k \notin \Delta, 1 \leq j \leq J\}$. Denote $\overline{K} = \{1, \ldots, K\}$. Defined the parameter space for $\boldsymbol{X}$ as $\mathcal{S}$, where

$$\mathcal{S} = \left\{\boldsymbol{X} \in \mathbb{R}^{\mathbb{Z}_+ \times \overline{K}} : \|\boldsymbol{x}(t_i)\| \leq C, \text{the columns of } \boldsymbol{X} \text{ are linearly independent}\right\}$$

with C being a positive constant. Then under the above parameter space, the following Theorem 1 provides a necessary and sufficient condition on the structured index matrix $\boldsymbol{Q}$ for the structured identifiability of the k-th latent factor.

Theorem 1. *We assume the columns of $\boldsymbol{Q}$ are linearly independent. Under Defini-tion 1, given the link function f_j , the k-th latent factor is structurally identifiable in the parameter space $\mathcal{S}$ if and only if*

$$\{k\} = \bigcap_{k\in\Delta,\ |R_{\boldsymbol{Q}}(\Delta)|=\mathcal{O}(J)} \Delta,$$

where we define $\bigcap_{k\in\Delta,\ |R_{\boldsymbol{Q}}(\Delta)|=\mathcal{O}(J)} \Delta = \emptyset$ if $R_{\boldsymbol{Q}}(\Delta) = \emptyset$ for all Δ that contains k, and where $|\cdot|$ denotes the cardinality of a set.

Theorem 1 is not only valid under the double-asymptotic ($N \to \infty$ and $J \to \infty$) setting but also when N and J are sufficiently large finite values provided that $|R_{\boldsymbol{Q}}(\Delta)| = \mathcal{O}(J)$ is replaced by $R_{\boldsymbol{Q}}(\Delta) \neq \emptyset$. Compared to Chen et al. (2020) and Leeb (2021) in which they discussed the conditions for generalized linear factor models, we discuss the identifiability of latent factors for a general nonlinear mod-els and the dependency of the k-th latent factor $\boldsymbol{X}_{[k]}$ across different values of t. The proof is given in the online Supplementary Materials. Theorem 1 is proved by contradiction. Because the latent factors are dependent across different observation points $i = 1, \ldots, N$, the construction of $\widetilde{\boldsymbol{X}}$ and $\boldsymbol{X}'$ requires accounting for correla-tions across rows. This differs from Chen et al. (2020), where the the latent factors are assumed to be independent across observation points $i = 1, \ldots, N$.

Remark 3. We discuss what happens when a global factor is present and thank one anonymous referee for raising this issue. Under the identifiability conditions of Theorem 1, the presence of a global factor renders some latent factors unidentifiable.

Nevertheless, the proposed general nonlinear structural latent factor model remains applicable. In such cases, although the consistency and interpretability of the unidentifiable factors become meaningless, the identifiable factors remain well defined and substantively interpretable. Scenario 2 in part (i) of the Simulation section provides an illustration. For further discussion of models incorporating a global factor, one may instead consider alternative specifications such as the bifactor model (Fang et al. (2021)).

2.3 Estimation and Algorithm

In this section, we provide the estimation procedure for the unknown link function, latent factors, and unknown parameters under the identifiability c onditions. First we consider using a Gaussian process (GP) prior to estimate the unknown link function $f_j(\cdot)$, and then use maximum a posteriori (MAP) estimation to obtain the estimators of the latent factors and unknown model parameters. The computation involves solving the inverse of the covariance matrix, which incurs high computational costs. We address this issue by using the nearest neighbor Gaussian process (NNGP) method.

2.3.1 Estimation of the nonlinear link function

Frequentist methods such as kernel estimation, local linear approaches, and B-splines are commonly employed to fit u nknown n onlinear f unctions. H owever, t hese techniques often face challenges like the "curse of dimensionality". Consequently, Gaussian process regression has become increasingly popular for estimating nonlinear functions. Specifically, w e a ssume a G P p rior f or t he u nknown f unction $f_j(\cdot)$ to quantify the smoothness of the nonlinear function between the observed variables and the latent factors, i.e.,

$$f_j(\cdot) \overset{prior}{\sim} \mathcal{GP}\left(0, \boldsymbol{C}_j(\cdot,\cdot;\boldsymbol{\Theta}_{fj})\right),$$

where $\boldsymbol{C}_j(\cdot,\cdot;\boldsymbol{\Theta}_{fj})$ is a covariance kernel function and $\boldsymbol{\Theta}_{fj}$ is referred to as the hyper-parameters. For any $t_i, t_l \in \mathcal{R}$, we use squared exponential kernel function, that is,

$$\begin{aligned} \boldsymbol{C}_j(i,l;\boldsymbol{\Theta}_{fj}) &= \operatorname{Cov}\left(f_j\left(\boldsymbol{x}^{(j)}(t_i)\right), f_j\left(\boldsymbol{x}^{(j)}(t_l)\right)\right) \\ &= v_{fj}\exp\left[-\frac{1}{2}\sum_{k=1}^{K} w_{fjk} r_{jk}\{x_k(t_i)-x_k(t_l)\}^2\right] \end{aligned}$$

where $\boldsymbol{\Theta}_{fj} = (v_{fj}, w_{fj1}, \ldots, w_{fjK})^\top$. Other types of covariance kernels can be found in Rasmussen and Williams (2006) and Shi and Choi (2011).

Denote $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_K) \in \mathbb{R}^{N\times K}$ with $\boldsymbol{x}_k = (x_k(t_1), \ldots, x_k(t_N))^\top$, and $\boldsymbol{Y} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_J) \in \mathbb{R}^{N\times J}$ with $\boldsymbol{Y}_j = (Y_j(t_1), \ldots, Y_j(t_N))^\top$. Let $\boldsymbol{f}_j = \left(f_j\left(\boldsymbol{x}^{(j)}(t_1)\right), \ldots, f_j\left(\boldsymbol{x}^{(j)}(t_N)\right)\right)^\top$ and $\boldsymbol{f} = (\boldsymbol{f}_1, \ldots, \boldsymbol{f}_J) \in \mathbb{R}^{N\times J}$. Followed by Shi and Choi (2011), we estimate the unknown nonlinear link functions $f_j(\cdot)$ using the posterior mean of Gaussian process regression. Specifically, for each $f_j\left(\boldsymbol{x}^{(j)}(t_i)\right)$, its estimator is obtained through the conditional expectation given all observations except the i-th time point, as described in Shi and Choi (2011) (p.19, formula (2.7)):

$$\mathrm{E}\left(f_j\left(\boldsymbol{x}^{(j)}(t_i)\right) \mid \mathcal{D}_{\setminus i}\right) = \boldsymbol{\psi}_j^\top(\boldsymbol{x}^{(j)}(t_i))_{\setminus i}\left(\boldsymbol{C}_{j\setminus i} + \sigma^2 \boldsymbol{I}_{N-1}\right)^{-1}\boldsymbol{Y}_{j\setminus i}, \tag{2.5}$$

where $\boldsymbol{\psi}_j\left(\boldsymbol{x}^{(j)}(t_i)\right)$ denotes the covariance vector between $f_j\left(\boldsymbol{x}^{(j)}(t_i)\right)$ and all other latent function values. This posterior mean serves as a plug-in estimator for $f_j(\cdot)$

in subsequent parameter estimation. The notation $\backslash i$ indicates the removal of the ith sample, meaning estimating $f_j\left(\boldsymbol{x}^{(j)}(t_i)\right)$ using all samples except the ith sample. That is, $\boldsymbol{\psi}_j^\top(\boldsymbol{x}^{(j)}(t_i))_{\backslash i} = (\psi_{j,1}, \ldots, \psi_{j,i-1}, \psi_{j,i+1}, \ldots, \psi_{j,N})$ and $\boldsymbol{Y}_{j\backslash i} = (Y_j(t_1), \ldots, Y_j(t_{i-1}), Y_j(t_{i+1}), \ldots, Y_j(t_N))^\top$. Here, we use the posterior mean of $f_j\left(\boldsymbol{x}^{(j)}(t_i)\right)$ in Expression (2.5) as the estimate of $f_j\left(\boldsymbol{x}^{(j)}(t_i)\right)$.

2.3.2 Estimation of latent factors and unknown parameters using nearest neighbor Gaussian process

Recall from the previous section that we assumed the latent factors follow a Gaussian process or a time series to model the dependence between different time points t. For the former, direct estimation of the latent factors using MAP method requires computing the inverse of the latent factor covariance matrix in the iterative algorithm, leading to a computational cost of $\mathcal{O}(N^3)$. In this paper, we define a NNGP (Datta et al., 2016) to address this issue. That is:

$$x_k(\cdot) \sim NNGP_k\left(0, \widetilde{\boldsymbol{\Sigma}}_k\left(\cdot, \cdot; \boldsymbol{\Theta}_{xk}\right)\right)$$

where $\widetilde{\boldsymbol{\Sigma}}_k\left(\cdot, \cdot; \boldsymbol{\Theta}_{xk}\right)$ is derived from the parent $\mathcal{GP}_k\left(0, \boldsymbol{\Sigma}_k\left(\cdot, \cdot; \boldsymbol{\Theta}_{xk}\right)\right)$ for $k = 1, \ldots, K$.

NNGP is a non-degenerate stochastic process that preserves the structural properties of conventional Gaussian processes while introducing sparsity through innovative neighborhood conditioning. Derived from a parent Gaussian process, NNGP operates as a distinct stochastic process alongside standard GP frameworks, retaining all essential theoretical properties of Gaussian processes and achieving computational acceleration through sparse covariance matrix construction. Its positive definite covariance structure ensures full-rank matrices, avoiding pathological distributions and maintaining non-degeneracy. The sparsity-aware methodology constrains the conditional dependence structure of spatial or temporal random effects to their m nearest neighbors, and Datta et al. (2016) rigorously demonstrated that this localized conditioning preserves multivariate Gaussian characteristics without compromising theoretical validity. The computational superiority of NNGP is highlighted through efficient algorithmic optimizations, with Finley et al. (2019) reducing the covariance matrix inversion complexity to $\mathcal{O}(Nm^3)$ floating-point operations. Next, the implementation of NNGP will be discussed in detail, and in Section 4, its computational efficiency will be compared to the standard Gaussian process framework through simulation studies.

Under the Bayesian analysis of the model, there are three types of parameters: $\{\boldsymbol{x}(t_i),\ i=1,\ldots,N\}$ for factor scores, σ^2 and $\{\boldsymbol{\Theta}_{xk},\ k=1,\ldots,K\}$ for the model parameters and $\{\boldsymbol{\Theta}_{fj},\ j=1,\ldots,J\}$ for the hyperparameters involved in the covariance function. Hyperparameters $\{\boldsymbol{\Theta}_{fj},\ j=1,\ldots,J\}$ are conventionally presumed to be known, and our goal is to estimate the factor scores $\boldsymbol{x}(t_i)$ and the model parameters $\boldsymbol{\Theta}_{xk}$. We will provide a method for estimating these hyperparameters later.

Using NNGP, we estimate the latent factors $\boldsymbol{X}$ and model parameters $\boldsymbol{\Theta}_{xk}$ through MAP estimation. Under the assumption that $f_j(\cdot)$ is given and $\boldsymbol{X}$ is independent of $\boldsymbol{f}$, the joint posterior distribution reduces to

$$p(\boldsymbol{X},\boldsymbol{\Theta}_{xk}|\boldsymbol{Y},\boldsymbol{f}) \propto p\left(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{f}\right)p(\boldsymbol{X}|\boldsymbol{\Theta}_{xk})p(\boldsymbol{\Theta}_{xk}). \tag{2.6}$$

We assume that a non-informative prior distribution is designated for $\boldsymbol{\Theta}_{xk}$. Then from (2.6) we need to maximize

$$\log p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{f})p(\boldsymbol{X}|\boldsymbol{\Theta}_{xk}) \propto \log p(\boldsymbol{X}|\boldsymbol{\Theta}_{xk}) + \log p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{f}), \tag{2.7}$$

where the latent factor likelihood term decomposes as

$$\log p(\boldsymbol{X}|\boldsymbol{\Theta}_{xk}) = \sum_{k=1}^{K} \log p(\boldsymbol{x}_k) = \sum_{k=1}^{K} \Big[-\frac{N}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}_k| - \frac{1}{2}\mathrm{tr}\big(\boldsymbol{\Sigma}_k^{-1}\boldsymbol{x}_k\boldsymbol{x}_k{}^{\top}\big)\Big]. \tag{2.8}$$

The above equation is derived from Equation (2.3). Assuming $\boldsymbol{X}$ is given, we use this Formula (2.8) to estimate $\boldsymbol{\Theta}_{xk}$. To address computational challenges with large N, we implement NNGP approximation by replacing the full covariance matrix $\boldsymbol{\Sigma}_k$ with its sparse counterpart $\widetilde{\boldsymbol{\Sigma}}_k$. This yields the modified log-likelihood

$$\ell(\boldsymbol{\Theta}_{xk}|\boldsymbol{X}) = \sum_{k=1}^{K} \left[-\frac{N}{2}\log(2\pi) - \frac{1}{2}\log\left|\widetilde{\boldsymbol{\Sigma}}_k\right| - \frac{1}{2}\mathrm{tr}\left(\widetilde{\boldsymbol{\Sigma}}_k^{-1}\boldsymbol{x}_k\boldsymbol{x}_k{}^{\top}\right)\right], \tag{2.9}$$

where the specific form of $\widetilde{\boldsymbol{\Sigma}}_k$ will be provided later.

To estimate $\boldsymbol{X}$, the maximum likelihood estimation in (2.7) for the model can be divided into a sum of two parts $\log p(\boldsymbol{X}|\boldsymbol{\Theta}_{xk})$ and $\log p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{f})$. The part of $\log p(\boldsymbol{X}|\boldsymbol{\Theta}_{xk})$ can be decomposed into the sum of K terms, where the density $p(\boldsymbol{x}_k)$ can be expressed as the product of conditional densities

$$p(\boldsymbol{x}_k) = p(x_k(t_1))p(x_k(t_2)|x_k(t_1))\dots p(x_k(t_N)|x_k(t_1),\dots x_k(t_{N-1})). \tag{2.10}$$

Based on Equation (2.3), we choose m nearest neighbors for each $x_k(t_i)$, $i = 1,\dots,N$, and replace the large conditioning sets on the right-hand side of (2.10) with m

previous nearest neighbors

$$\widetilde{p}(\boldsymbol{x}_k) = p(x_k(t_1)) \prod_{i=2}^{N} p\left(x_k(t_i) \mid \boldsymbol{x}_k\left(N\left(t_i\right)\right)\right), \tag{2.11}$$

where $\boldsymbol{x}_k\left(N\left(t_i\right)\right)$ is m previous nearest neighbors of $x_k(t_i)$. The selection of m can be refered to Datta et al. (2016) and Guinness (2018). It can be shown that $\widetilde{p}(\boldsymbol{x}_k(t))$ in (2.11) is a multivariate Gaussian density with covariance matrix $\widetilde{\boldsymbol{\Sigma}}_k$. The term $p\left(x_k(t_i) \mid \boldsymbol{x}_k\left(N\left(t_i\right)\right)\right)$ is the conditional density $N\left(x_k(t_i) \mid \boldsymbol{a}_{i,k}^\top \boldsymbol{x}_k\left(N\left(t_i\right)\right), d_{i,k}\right)$, where

$$\boldsymbol{a}_{i,k} = \boldsymbol{\Sigma}_k\left(N\left(t_i\right), N\left(t_i\right)\right)^{-1} \boldsymbol{\Sigma}_k\left(N\left(t_i\right), t_i\right), \tag{2.12}$$

$$d_{i,k} = \boldsymbol{\Sigma}_k\left(t_i, t_i\right) - \boldsymbol{\Sigma}_k\left(t_i, N\left(t_i\right)\right) \boldsymbol{a}_{i,k}. \tag{2.13}$$

Finley et al. (2019) provides efficient algorithms to calculate sparse $\widetilde{\boldsymbol{\Sigma}}_k = (\boldsymbol{I} - \boldsymbol{A}_k)^{-1}\boldsymbol{D}_k(\boldsymbol{I} - \boldsymbol{A}_k)^{-\top}$, where $\boldsymbol{A}_k = (\boldsymbol{a}_{1,k}, \ldots, \boldsymbol{a}_{N,k})$ and $\boldsymbol{D}_k = \operatorname{diag}\{d_{1,k}, \ldots, d_{N,k}\}$, the expression of $\boldsymbol{a}_{i,k}$ and $d_{i,k}$ are given in (2.12) and (2.13). Thus we can obtain the expression for $\left|\widetilde{\boldsymbol{\Sigma}}_k\right|$ and $\widetilde{\boldsymbol{\Sigma}}_k^{-1}$ in Equation (2.9), that is

$$\widetilde{\boldsymbol{\Sigma}}_k^{-1} = (\boldsymbol{I} - \boldsymbol{A}_k)^\top \boldsymbol{D}_k^{-1}(\boldsymbol{I} - \boldsymbol{A}_k) \tag{2.14}$$

and $\widetilde{\boldsymbol{\Sigma}}_k^{-1}$ admits a Cholesky decomposition $\widetilde{\boldsymbol{\Sigma}}_k^{-1} = \boldsymbol{L}_k^\top \boldsymbol{L}_k$ with the lower-triangular Cholesky factor $\boldsymbol{L}_k = \boldsymbol{D}_k^{-1/2}(\boldsymbol{I} - \boldsymbol{A}_k)$. Then

$$\left|\widetilde{\boldsymbol{\Sigma}}_k\right| = \prod_{i}^{N} l_{kii}^{-2}, \tag{2.15}$$

where l_{kii} is the diagonal element of matrix $\boldsymbol{L}_k$.

The sequential NNGP algorithm proposed in Datta et al. (2016) updates the components of $x_k(t_i)$ individually for $i = 1, \ldots, N$. That is, updates $x_k(t_i)$ from $N(0, \widetilde{d}_{i,k})$, where

$$\frac{1}{\widetilde{d}_{i,k}} = \frac{1}{d_{i,k}} + \sum_{l=i+1}^{i+m} \frac{\boldsymbol{a}_{l,k}^2[i]}{d_{l,k}}, \tag{2.16}$$

where $\boldsymbol{a}_{l,k}[i]$ is the ith element of vector $\boldsymbol{a}_{l,k}$. Thus, for $\log p(\boldsymbol{X}|\boldsymbol{\Theta}_{xk})$, we can update $x_k(t_i)$ from $\sum_{k=1}^K \log \widetilde{p}\left(x_k(t_i) \mid N(0, \widetilde{d}_{i,k})\right)$.

Equation (2.7) involves the part of $\log p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{f})$ as well. From Model (2.1), $Y_j(t_i)|f_j(\cdot), \boldsymbol{X} \sim N\left(f_j(\boldsymbol{x}^{(j)}(t_i)), \sigma^2\right)$ independently for $i = 1, \ldots, N$, that is

$$\log p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{f}) \propto \sum_{i=1}^N \sum_{j=1}^J \frac{1}{\sigma^2}\left[Y_j(t_i) f_j\left(\boldsymbol{x}^{(j)}(t_i)\right) - \frac{1}{2}(f_j\left(\boldsymbol{x}^{(j)}(t_i)\right))^2\right] - NJ\log\sigma.$$

Combining the two parts above, when the estimation of $f_j(\cdot)$ and σ^2 are given, we can estimate $\boldsymbol{x}(t_i)$ sequentially, i.e.

$$\widehat{\boldsymbol{x}}(t_i) = \underset{\boldsymbol{x}(t_i)}{\arg\max}\ \ell(\boldsymbol{x}(t_i)),$$

where the objective function combines the log-likelihood of observations and NNGP distribution:

$$\begin{aligned}\ell(\boldsymbol{x}(t_i)) \quad &\propto \quad \sum_{j=1}^J \frac{1}{\sigma^2}\left[Y_j(t_i) f_j\left(\boldsymbol{x}^{(j)}(t_i)\right) - \frac{1}{2}(f_j\left(\boldsymbol{x}^{(j)}(t_i)\right))^2\right] - J\log\sigma \\ &\quad + \sum_{k=1}^K \log\widetilde{p}\left(x_k(t_i) \mid N(0, \widetilde{d}_{ik})\right), \end{aligned} \tag{2.17}$$

where $\boldsymbol{x}^{(j)}(t_i)$ is given in (2.2), and $\widetilde{d}_{ik}$ is defined in (2.16).

2.3.3 Iterative algorithm

In the following, we introduce an iterative algorithm to estimate the unknown link function, latent factors, and model parameters. The iterative approach starts with giving initial values of $\boldsymbol{X}$, σ^2 and $\boldsymbol{\Theta}_{xk}$. The initial values of $\boldsymbol{X}$ are computed using a linear latent factor model, with the same constraints imposed on the loading matrix. The initial values of $\boldsymbol{\Theta}_{xk}$ and σ^2 are generated from a random uniform distribution. We first assume the hyperparameters $\boldsymbol{\Theta}_{fj}$ are given and aim to estimate the factor scores $\boldsymbol{X}$ and the model parameters $(\sigma^2, \boldsymbol{\Theta}_{xk})$. In the subsequent remark, we will discuss how to estimate the hyperparameters using the empirical Bayesian method. Algorithm 1 provides the detailed steps of the iterative algorithm, which is implemented by alternately updating $\boldsymbol{\Theta}_{xk}$ and $\boldsymbol{X}$.

This iterative procedure ensures consistent estimation of both latent factors and nonlinear link functions while maintaining computational tractability for functional data.

Remark 4. (i) The NNGP approximation in Step 1 replaces the full GP covariance $\boldsymbol{\Sigma}_k$ with a sparse version $\widetilde{\boldsymbol{\Sigma}}_k$, reducing computational complexity from $\mathcal{O}(N^3)$ to $\mathcal{O}(Nm^3)$. (ii) The optimization in Step 2 leverages the block-diagonal structure induced by NNGP, allowing parallel computation across time points. (iii) Hyperparameters are typically considered to be known. If they are unknown, we apply the empirical Bayesian method to estimate their values using the same approach as in Step 1 expression (2.18) of Algorithm 1, which can be further refined by in-

Algorithm 1 Iterative Estimation via NNGP

```
Initialize: Set initial values for $\boldsymbol{X}^{(0)}$, $\sigma^{2(0)}$ and $\boldsymbol{\Theta}_{xk}^{(0)}$
while not converged do
    Step 1: Update parameters $\boldsymbol{\Theta}_{xk}^{(iter)}$ and $\sigma^{2(iter)}$ given $\boldsymbol{X}^{(iter-1)}$
      Estimate $\sigma^{2(iter)}$ by maximizing:
        $\ell(\sigma^2|\boldsymbol{X}^{(iter-1)}) = \sum_{j=1}^{J} \Big[ -\frac{1}{2} \log |\sigma^2 \boldsymbol{I}_N + \boldsymbol{C}_N| - \frac{1}{2} \mathrm{tr}\Big( (\sigma^2 \boldsymbol{I}_N + \boldsymbol{C}_N)^{-1} \mathbf{Y}_j \mathbf{Y}_j{}^{\top} \Big) \Big]. \quad (2.18)$
      Then maximize the log-likelihood of the latent factors using expression (2.9), where the
    expressions of $\widetilde{\boldsymbol{\Sigma}}_k^{-1}$ and $\left|\widetilde{\boldsymbol{\Sigma}}_k\right|$ are given in equations (2.14) and (2.15).
    Step 2: Update latent factors $\boldsymbol{X}^{(iter)}$ given $\boldsymbol{\Theta}_{xk}^{(iter-1)}$ and $\sigma^{2(iter-1)}$
      For each time point $t_i$, solve:
        $\boldsymbol{x}^{(iter)}(t_i) = \arg\max_{\boldsymbol{x}(t_i)} \ell(\boldsymbol{x}(t_i)|\boldsymbol{\Theta}_{xk}^{(iter-1)}, \sigma^{2(iter-1)}),$
    where the objective function is defined in (2.17), and $f_j\left(\boldsymbol{x}^{(j)}(t_i)\right)$ is replaced by its estimator
    $\mathrm{E}\left(f_j\left(\boldsymbol{x}^{(j)}(t_i)\right)\right)$, which is given in (2.5).
    Check convergence: Repeat Steps 1-2 until $\|\boldsymbol{X}^{(iter)} - \boldsymbol{X}^{(iter-1)}\|_2 < \epsilon$ where $\epsilon$ is a
    small convergence threshold.
end while
Estimate unknown link function: Using the final estimate $\widehat{\boldsymbol{X}}$, compute for each $j$ and $t_i$:
    $\widehat{f}_j\left(\boldsymbol{x}^{(j)}(t_i)\right) = \mathrm{E}\left[f_j\left(\boldsymbol{x}^{(j)}(t_i)\right)|\widehat{\boldsymbol{X}}\right]$
  where the expectation is evaluated via (2.5).
Output: Estimated latent factors $\widehat{\boldsymbol{X}}$, model parameters $\widehat{\boldsymbol{\Theta}}_{xk}$, $\widehat{\sigma}^2$, and link functions $\widehat{f}_j(\cdot)$
```

corporating the estimation within Step 1, enabling concurrent estimation of both hyper parameters and model parameters. (iv) Although the true model is nonlinear, when the nonlinearity is not severe, we use a linear factor model to initialize $\boldsymbol{X}$, which performs well in practice. More caution is required in more general cases. In addition, since the conditional objective functions (2.9,2.17,2.18) admit multiple maximizers, we initialize both latent factors and model parameters with multiple starting values and perform a global search, selecting the solution that maximizes the joint likelihood.

3. Asymptotic Properties

In this section, we present the asymptotic properties for the estimators of latent factors, unknown parameters and unknown link function. Theorem 2 shows that the latent factors and unknown parameters are consistently estimated given the true link function f_j^*, if the condition in Theorem 1 is satisfied. Theorem 3 provides posterior consistency of (f_j, σ^2) given the latent factor estimators.

Theorem 2. *Let* $(\boldsymbol{X}^*, \boldsymbol{Q}^*) \in \mathcal{S}$, *defined in Section 2.2, be the true factor scores and informative matrix,* $\boldsymbol{x}^*(t_i)$ *is the* i*-th row of* $\boldsymbol{X}^*$ *and* $\boldsymbol{r}_j^*$ *is the* j*-th row of* $\boldsymbol{Q}^*$*. Let* $(\boldsymbol{\Theta}_{xk}^*, \sigma_*^2)$ *be the true values of model parameters. Let* $\left(\widehat{\boldsymbol{x}}(t_i), \widehat{\boldsymbol{\Theta}}_{xk}, \widehat{\sigma}^2\right)$ *denote the estimators of* $(\boldsymbol{x}^*(t_i), \boldsymbol{\Theta}_{xk}^*, \sigma_*^2)$*. If* $\boldsymbol{Q}$ *satisfies the condition in Theorem 1, given* $(\boldsymbol{x}^*(t_i), \boldsymbol{\Theta}_{xk}^*)$*, then* $\widehat{\sigma}^2 \xrightarrow{p} \sigma_*^2$ *as* $N \to \infty$ *and* $J \to \infty$*. Furthermore, when* $\widehat{\sigma}^2$ *is a consistent estimator of* σ_*^2*, the latent factors* $\widehat{\boldsymbol{x}}(t_i)$ *are consistent, that is,* $\widehat{\boldsymbol{x}}(t_i) \xrightarrow{p} \boldsymbol{x}^*(t_i)$ *as* $J \to \infty$*. When* $\widehat{\boldsymbol{x}}(t_i)$ *is a consistent estimator of* $\boldsymbol{x}^*(t_i)$*, we can also prove that* $\widehat{\boldsymbol{\Theta}}_{xk} \xrightarrow{p} \boldsymbol{\Theta}_{xk}^*$ *as* $N \to \infty$*.*

Theorem 3. *Denote* P_0 *be the joint conditional distribution of* $\{Y_j(t_i)\}_{i=1}^{\infty}$ *given the latent factors, assuming that* f_j^* *is the true link function,* $\boldsymbol{\Theta}_{xk}^*$ *for* $k = 1, \ldots, K$ *and* σ_*^2 *are the true model parameters,* $\boldsymbol{\Theta}_{fj}^*$ *for* $j = 1, \ldots, J$ *are the true hyperparameters. Let* f_j *have prior* Π *and* $\boldsymbol{\lambda}$ *be a* K*-dimensional Lebesgue measure. Then, under the true informative matrix* $\boldsymbol{Q}^*$*, i.e.* $\boldsymbol{R}_j$ *accurately describes which latent factors are associated with the manifest variable* $Y_j(t_i)$*, given the consistent estimation* $(\widehat{\boldsymbol{x}}(t_i), \widehat{\boldsymbol{\Theta}}_{xk}, \widehat{\sigma}^2, \widehat{\boldsymbol{\Theta}}_{fj})$ *of latent factors, unknown parameters, and hyperparameters,* *if Assumptions A1-A2 given in the online Supplementary Materials hold, for every* $\epsilon > 0$,

$$\Pi\left\{f_j \in W_\epsilon \mid \boldsymbol{Y}, \widehat{\boldsymbol{x}}(t_i), \boldsymbol{R}_j, \widehat{\boldsymbol{\Theta}}_{xk}, \widehat{\sigma}^2, \widehat{\boldsymbol{\Theta}}_{fj}\right\} \to 1 \ \textit{in } P_0 - \textit{probability},$$

where

$$W_\epsilon = \left\{f_j : \int \left| f_j\left(\widehat{\boldsymbol{x}}^\top(t_i)\boldsymbol{R}_j\right) - f_j^*\left(\boldsymbol{x}^{*\top}(t_i)\boldsymbol{R}_j\right)\right| d\boldsymbol{\lambda}(\boldsymbol{x}^*(t_i)) < \epsilon\right\}.$$

In other words, for each j*, we have* $f_j\left(\widehat{\boldsymbol{x}}^\top(t_i)\boldsymbol{R}_j\right) \xrightarrow{p} f_j^*\left(\boldsymbol{x}^{*\top}(t_i)\boldsymbol{R}_j\right)$ *as* $N \to \infty$.

The proofs of both theorems are given in the online Supplementary Materials.

Remark 5. For the consistency of the latent factors and the unknown link function, we have the following explanation. From Model (2.2), it can be seen that the information about the unknown link function f_j is mainly provided by observations of the j-th component Y_j at $i = 1, \ldots, N$, and is not related to other components of $\boldsymbol{Y}$, while the information about the latent factors primarily comes from all components $Y_j(t_i)$, $j = 1, \ldots, J$. For the link functions f_j , we show that it suffices to establish the consistency of f_j for a fixed j, which is not in conflict with letting $J \to \infty$. The dependence of the link function on j is introduced for model generality. In practice, multiple manifest variables are often linked to the same latent factor and therefore share the same link function for the corresponding indices j. See Equation (4.19) in the Simulation section for an example. For the consistency of the latent factors, we assume that N, J tend to infinity. Since $Y_j(t_i)$ is assumed to have a certain smooth-ness in the time dimension (i.e. different t_i), we also assume a certain smoothness for

$x_k(t_i)$, but its consistency mainly relies on the information from different j of $Y_j(t_i)$.

4. Simulation Studies

In this section, simulation studies are conducted to evaluate the finite-sample performance of both latent factor estimator and the unknown link function estimator. First, to assess the robustness of our model, under a general nonlinear structured latent factor model, we compared two scenarios: one where all factors satisfy the identifiability conditions in Theorem 1 of our paper but not those in Zhang et al. (2025), and another where some factors violate our Theorem 1 conditions. The results confirm that our algorithm correctly identifies factors meeting our theoretical conditions, while Zhang et al. (2025)'s model fails in these cases. Second, using multi-index structured data from Zhang et al. (2025), we demonstrate that our model matches their performance, proving its flexibility without sacrificing accuracy. Finally, to improve computational efficiency, we used NNGP in the latent factor estimator, and the average computation time of NNGP and GP was compared.

(i) Robustness Analysis with General Structured Data. The data are generated from a nonlinear structured latent factor model as follows

$$Y_j(t_i) = f_j(\boldsymbol{x}^\top(t_i)\boldsymbol{R}_j) + \varepsilon_j(t_i), \quad j = 1, \ldots, J, \quad i = 1, \ldots, N,$$

where $\boldsymbol{x}(t_i) = (x_1(t_i), \ldots, x_K(t_i))^\top$, $\varepsilon_j(t_i) \sim N(0, \sigma^2)$ with $\sigma^2 = 0.25$, and the specific expression for f_j will be provided below. In all settings, we consider $K = 3$, $J = \{6, 12, 21, 30, 99\}$ and $N = 5J$. The factor scores $\boldsymbol{x}_k = (x_k(t_1), \ldots, x_k(t_N))^\top \sim \mathcal{GP}$ which are dependent for $i = 1, \ldots, N$ but independent for $k = 1, 2, 3$ using kernel function (2.4) with $v_{xk} = 1$ and $w_{xk} = 100$ for $k = 1, 2, 3$. Compared to our previous work (Zhang et al., 2025), here we consider the dependence of $x_k(t_i)$ between different t_i's, and our model allows $f_j(\cdot)$ to be a multivariate function, which is more flexible in practice. To investigate the conditions of identifiability, we consider two scenarios, where the first one satisfies identifiability conditions, but the second one violates them.

Scenario 1. Here, we consider Model (2.1) with

$$f_j(\cdot) = \begin{cases} \frac{1}{1+e^{-a_{j1}x_1(t_i)}} + \frac{1}{1+e^{-a_{j2}x_2(t_i)}} & \text{for } j = 1, \ldots, J/3 \\ \frac{1}{1+e^{-a_{j1}x_1(t_i)}} + \frac{1}{1+e^{-a_{j3}x_3(t_i)}} & \text{for } j = J/3+1, \ldots, 2*J/3 \\ \frac{1}{1+e^{-a_{j2}x_2(t_i)}} + \frac{1}{1+e^{-a_{j3}x_3(t_i)}} & \text{for } j = 2*J/3+1, \ldots, J. \end{cases} \quad (4.19)$$

The corresponding $\boldsymbol{R}_j = \text{diag}\{1, 1, 0\}$ for $j = 1, \ldots, J/3; \boldsymbol{R}_j = \text{diag}\{1, 0, 1\}$ for $j = J/3 + 1, \ldots, 2 * J/3$ and $\boldsymbol{R}_j = \text{diag}\{0, 1, 1\}$ for $j = 2 * J/3 + 1, \ldots, J$. The coefficients $\boldsymbol{a}_j = (a_{j1}, a_{j2}, a_{j3})^\top$'s are generated iid from distributions over the ball

$$\left\{\boldsymbol{a} \in \mathbb{R}^2 : \|\boldsymbol{a}\| \leq 2.5, \text{each element of } \boldsymbol{a} \geq 0\right\}.$$

All latent factors are structurally identifiable even when there is no item measuring a single latent factor.

Scenario 2. Here only the values of $\boldsymbol{R}_j$ and f_j for $j = J/3+1, \ldots, 2*J/3$ from Scenario 1 are different, that is $\boldsymbol{R}_j = \text{diag}\{0, 0, 1\}$ for $j = J/3+1, \ldots, 2*J/3$, where $Y_j(t_i) = \frac{1}{1+e^{-a_{j3}x_3(t_i)}} + \varepsilon_j(t_i)$, all other elements remain unchanged. Then the first and the third latent factors are identifiable, while the second latent factor is not identifiable.

In this simulation study, we mainly investigate the finite sample performance of the estimators for the latent variables and the unknown link function. The former is measured by the correlation and sin value between the latent variables and their estimators. For the sake of simplicity and convenience of notation, we denote $\mathrm{Corr}_{\boldsymbol{x}_1} = \mathrm{corr}(\boldsymbol{x}_1, \widehat{\boldsymbol{x}}_1)$, $\mathrm{Sin}_{\boldsymbol{x}_1} = \sin(\boldsymbol{x}_1, \widehat{\boldsymbol{x}}_1)$ and similar notations for $\boldsymbol{x}_2$ and $\boldsymbol{x}_3$. To measure the convergence of the estimation of the unknown function f_j for $j = 1, \ldots, J$, we used the quantity $d_f = \frac{\sum_{j=1}^{J} \|\widehat{\boldsymbol{f}}_j - \boldsymbol{f}_{j*}\|^2}{NJ}$, where f_{j*} is the true value of link function $f_j(\cdot)$ given in Expression (4.19) or (4.20). For each sample size, 100 replications are conducted and we take the average of those measures. We use two gradient-based algorithms respectively in the optimization step. Both the scaled conjugate gradient algorithm and the gradient descent algorithm yield the same result, although they have different convergence speeds.

As a comparison to our proposed method (denoted by GNSLFM), we also use a linear latent factor model (denoted by LLF) and our previous work (Zhang et al., 2025, denoted by NSLFM). We apply the same variable-factor linkage relationships to both LLF and NSLFM. The results of the simulation studies are shown in Tables 1-2, as well as Figures 2-3.

Table 1 shows the results of latent factor estimation and unknown link function estimation under Setting 1 for different values of J ($N = 5 * J$). From Scenario 1 in Table 1, we have the following findings. First, we can see that the correlation between the latent variables and their estimates using our proposed method tends to 1, and the sin value tends to 0 as J increases, indicating the identifiability of the estimated factor scores. The convergence of the estimation of the factor scores is mainly dependent on J. The results of d_f show the good accuracy of the estimation for the unknown link function, which improves as N and J increases. Second, the accuracy of the factor score estimates derived from the LLF method is inferior to that of our proposed model, because the LLF method fails to capture the nonlinear characteristics of the data. The results of NSLFM also do not perform as well as our method, as the model GNSLFM offers enhanced flexibility while the NSLFM can only handle data that has a multi-index structure. Besides, GNSLFM takes into account the correlations among latent factors across different time points, which is not considered in both NSLFM and LLF.

From Scenario 2 in Table 1, when the latent factor is not identifiable, the $\mathrm{Corr}_{\boldsymbol{x}_2}$ is not close to 1, $\mathrm{Sin}_{\boldsymbol{x}_2}$ is not close to 0 even for large J, which reflects the impact of identifiability on latent factor estimation. The accuracy of the estimation of the unknown function improves with the increase in N and J. The accuracy of the unknown function estimators is unaffected by identifiability conditions and improves as the sample size grows. This robustness is due to the fact that, even in cases where the latent space lacks uniqueness, GNSLFM can still extract meaningful information and capture patterns within the data, resulting in effective prediction performance.

Figures 2-3 show the true values and estimations of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, evaluated from 99 observations in Scenario 1 for one replication, using our proposed model (GNSLFM) and LLF. The estimated values for $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ obtained from our model closely match their true values. In contrast, the estimates obtained from LLF do not smoothly replicate the true $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. This discrepancy arises because the LLF method

Table 1: The results of *GNSLFM*, *LLF* and *NSLFM* in Simulation Study(i)

		J	Corr$_{\boldsymbol{x}_1}$	Corr$_{\boldsymbol{x}_2}$	Corr$_{\boldsymbol{x}_3}$	Sin$_{\boldsymbol{x}_1}$	Sin$_{\boldsymbol{x}_2}$	Sin$_{\boldsymbol{x}_3}$	d_f
Scenario 1	*GNSLFM*	6	0.83	0.82	0.82	0.45	0.44	0.45	0.62
		12	0.86	0.88	0.86	0.40	0.39	0.41	0.52
		21	0.91	0.89	0.91	0.38	0.38	0.38	0.41
		30	0.95	0.92	0.94	0.34	0.34	0.32	0.40
		99	1.00	1.00	1.00	0.15	0.17	0.16	0.22
	LLF	6	0.74	0.74	0.74	0.75	0.80	0.77	
		12	0.81	0.81	0.82	0.65	0.68	0.67	
		21	0.87	0.88	0.87	0.51	0.56	0.49	
		30	0.91	0.92	0.91	0.49	0.45	0.42	
		99	0.92	0.92	0.92	0.35	0.35	0.36	
	NSLFM	6	0.76	0.76	0.77	0.55	0.58	0.59	0.58
		12	0.83	0.83	0.85	0.46	0.49	0.49	0.45
		21	0.90	0.89	0.90	0.31	0.36	0.39	0.34
		30	0.93	0.93	0.93	0.29	0.29	0.25	0.22
		99	0.93	0.94	0.94	0.27	0.28	0.26	0.12
Scenario 2	*GNSLFM*	6	0.82	0.45	0.82	0.45	0.85	0.45	0.82
		12	0.85	0.48	0.87	0.40	0.80	0.41	0.62
		21	0.90	0.53	0.90	0.38	0.76	0.38	0.41
		30	0.95	0.55	0.95	0.34	0.70	0.32	0.40
		99	1.00	0.55	1.00	0.05	0.61	0.06	0.25

assumes that the samples are independent and fails to account for the nonlinear characteristics of the data.

(ii) Performance on Multi-index Structured Data. In this setting, we use the multi-index structured data as discussed in Zhang et al. (2025). The data is generated as follows

$$Y_j(t_i) = \frac{1}{1 + e^{-\boldsymbol{a}_j^\top \boldsymbol{x}(t_i)}} + \varepsilon_j(t_i), \quad j = 1, \ldots, J, \quad i = 1, \ldots, N, \tag{4.20}$$

where $\boldsymbol{x}(t_i) = (x_1(t_i), x_2(t_i), x_3(t_i))^\top$ and $\varepsilon_j(t_i) \sim N(0, \sigma^2)$ with $\sigma^2 = 0.25$. The J and N follow the same setup as Setting 1. The confirmatory matrix definded in

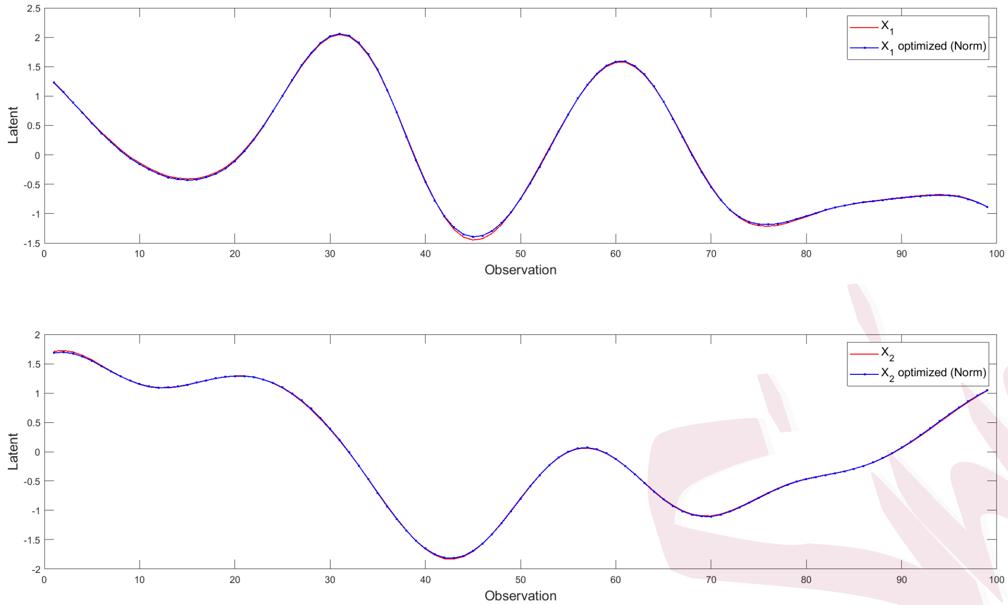

Figure 2: The true values and estimates of $x_1(t)$ and $x_2(t)$ by using GNSLFM with dependent data.

Zhang et al. (2025) is set as

$$\boldsymbol{Q}^\top = \begin{pmatrix} 1 & 1 & 0 & \cdots & 1 & 1 & 0 & \cdots \\ 1 & 0 & 1 & \cdots & 1 & 0 & 1 & \cdots \\ 0 & 1 & 1 & \cdots & 0 & 1 & 1 & \cdots \end{pmatrix}.$$

The factor scores $\boldsymbol{x}(t_i)'s$ and the loading coefficients $\boldsymbol{a}_j's$ are generated iid from distributions over the ball $\{\boldsymbol{x} \in \mathbb{R}^3 : \|\boldsymbol{x}\| \leq 2.5\}$. In this setting, all latent factors satisfy the identifiability conditions in Zhang et al. (2025) and in Theorem 1 of this paper.

Table 2 shows the results of latent factor estimation and unknown link function estimation. Even if the data has a multi-index type structure, the performance of GNSLFM is quite good and similar to the results achieved by the true model of

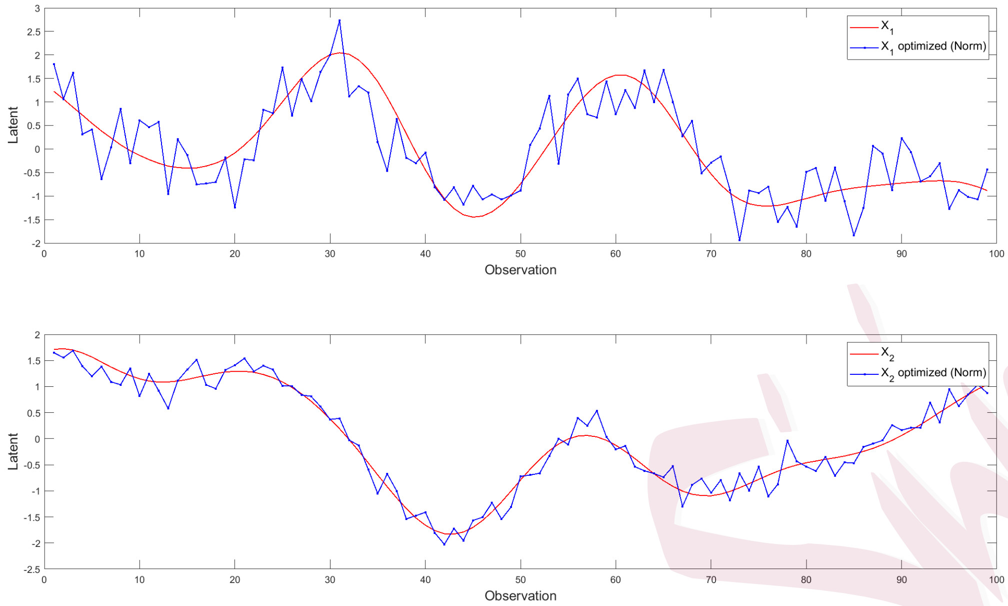

Figure 3: The true values and estimates of $x_1(t)$ and $x_2(t)$ by using linear FA with dependent data.

NSLFM. The model NSLFM, as described in Zhang et al. (2025), can be considered a special case of our proposed model, GNSLFM. Both GNSLFM and NSLFM outperform LLF, and GNSLFM is applicable to a wider range of data types. When handling data with a multi-index type structure, both GNSLFM and NSLFM are suitable; however, for the data types presented in Table 1, our proposed GNSLFM model is superior.

(iii) Computational Efficiency via NNGP. Table 3 summarizes the computation time between the full GP and NNGP methods. Under the setting where $K = 2, N = 5 * J$, and $m = 20$ nearest neighbors are taken as the conditional information for the density function $p(\boldsymbol{X})$, it can be seen that using the NNGP algorithm is much more efficient than the full GP.

Table 2: The result of *GNSLFM*, *LLF* and *NSLFM* in Simulation Study(ii)

	J	Corr$_{\boldsymbol{x}_1}$	Corr$_{\boldsymbol{x}_2}$	Corr$_{\boldsymbol{x}_3}$	Sin$_{\boldsymbol{x}_1}$	Sin$_{\boldsymbol{x}_2}$	Sin$_{\boldsymbol{x}_3}$	d_f
GNSLFM	6	0.76	0.76	0.77	0.55	0.58	0.59	0.58
	12	0.83	0.83	0.85	0.46	0.49	0.49	0.45
	21	0.90	0.89	0.90	0.31	0.36	0.39	0.34
	30	0.93	0.93	0.93	0.29	0.29	0.25	0.22
	99	0.97	0.97	0.98	0.17	0.18	0.16	0.12
LLF	6	0.74	0.74	0.74	0.75	0.80	0.77	
	12	0.81	0.81	0.82	0.65	0.68	0.67	
	21	0.87	0.88	0.87	0.51	0.56	0.49	
	30	0.91	0.92	0.91	0.49	0.45	0.42	
	99	0.94	0.94	0.94	0.07	0.50	0.06	
NSLFM	6	0.83	0.82	0.82	0.45	0.44	0.45	0.62
	12	0.86	0.88	0.86	0.40	0.39	0.41	0.52
	21	0.91	0.89	0.91	0.38	0.38	0.38	0.41
	30	0.95	0.92	0.94	0.34	0.34	0.32	0.40
	99	1.00	1.00	1.00	0.15	0.17	0.16	0.22

5. Analysis of Gait Data

Parkinson's disease (PD) is a complex neurodegenerative disorder that leads to challenges in disease management, reduced quality of life, and increased healthcare costs (Hoehn and Yahr, 1998). Gait, as an early diagnostic tool for PD, is also used to predict morbidity, mortality, fall risk and other neurological disorders (Buckley et al., 2019). Currently, this research still faces challenges due to numerous factors influencing the performance of early identification of PD, such as walking protocols, gait assessment systems, cohort size, disease severity stage of PD, and validation methods. In this section, we apply the proposed method to gait data collected in our laboratory. We collected one week of continuous steady-state gait data from Parkinson's patients using wearable devices. After data preprocessing, the gait features of each individual can be represented as functional data over time. It is important

Table 3: Comparison of the computation time between the full GP and NNGP methods

	Full GP		**NNGP**	
N	**Iter**	**CPU(s)**	**Iter**	**CPU(s)**
20	6	1.00	2	0.62
40	7	14.62	7	1.06
60	4	64.80	7	4.35
80	6	223.83	3	9.81
100	4	178.83	2	15.79
200	6	354.96	4	40.61
500	6	754.96	4	60.61

to note that there are many variables that could describe gait characteristics, i.e., the number of manifest variables J, could be very large. Moreover, gait data is a form of free living data, meaning it can be used to identify activities, detect diseases, and other applications. Given the complexity and high-dimensional nature of such data, there is a need for low-dimensional, interpretable latent variables to effectively capture and analyze the relationships among various gait characteristics that vary continuously over time. Our aim is to measure these relationships using a few latent factors in a nonlinear way.

We treat gait characteristics for one Parkinson's patient as manifest variables. For illustrative purposes, we focus on a 13-dimensional feature set, which includes gait speed, step length, stride time, degree of asymmetry in time and distance, and other relevant metrics. We collected observations at 300 time points, i.e., $N = 300$. Referring to the analysis of gait data using a linear latent factor model in Morris et al. (2017), we consider four latent factors: Pace, Rhythm, Asymmetry, and Variability (SD). The first factor is related to gait speed and step length. The second factor is associated with the duration of one stride, one step, one stance and one swing.

The third factor represents the degree of asymmetry in time and distance, measured in absolute value. The fourth factor is related to all the variables corresponding to the first and second factors, and can be regarded as a composite index of time, length, and velocity. Based on the above relationships and identifiability conditions, we formulated the design matrix $\boldsymbol{R}_j$. We obtained estimates of the latent factors under both models, LFA and GNSLFM.

Figure 4 shows the estimation curves of the second and fourth factor scores under the two models, respectively. From Figure 4, we can see that the factor scores obtained from GNSLFM exhibit smoother patterns. Figure 5 displays the first-order autocorrelation function of the second and fourth factors derived from both models. The autocorrelation function obtained from LLF remains close to 0, indicating that the factor scores are largely uncorrelated. This reveals that LLF usually cannot capture the correlations between different time points. In contrast, the autocorrelation function from the GNSLFM exhibits a distinct pattern and trend, showing characteristics more naturally. This indicates that GNSLFM may capture nonlinear structures within the data that LLF fails to represent. Furthermore, we performed a frequency domain analysis using power spectral density (Bansal and Dimri, 2021), which is a measure that describes the distribution of power contained within a signal as a function of frequency. As illustrated in Figure 6, the power spectral density of LLF is predominantly concentrated in the low-frequency region. In contrast, the power spectral density of GNSLFM is distributed across a broader frequency range, revealing a more diverse spectrum of frequency components. This indicates that nonlinear dimensionality reduction captures a wider range of frequency features and finer details within the data. For the other two factors, we obtained similar results, which are omitted here for simplicity. In summary, our proposed method (GNSLFM) captures richer information from the data, and the estimated factor scores are smoother, which is attributed to the model's ability to effectively capture nonlinear features.

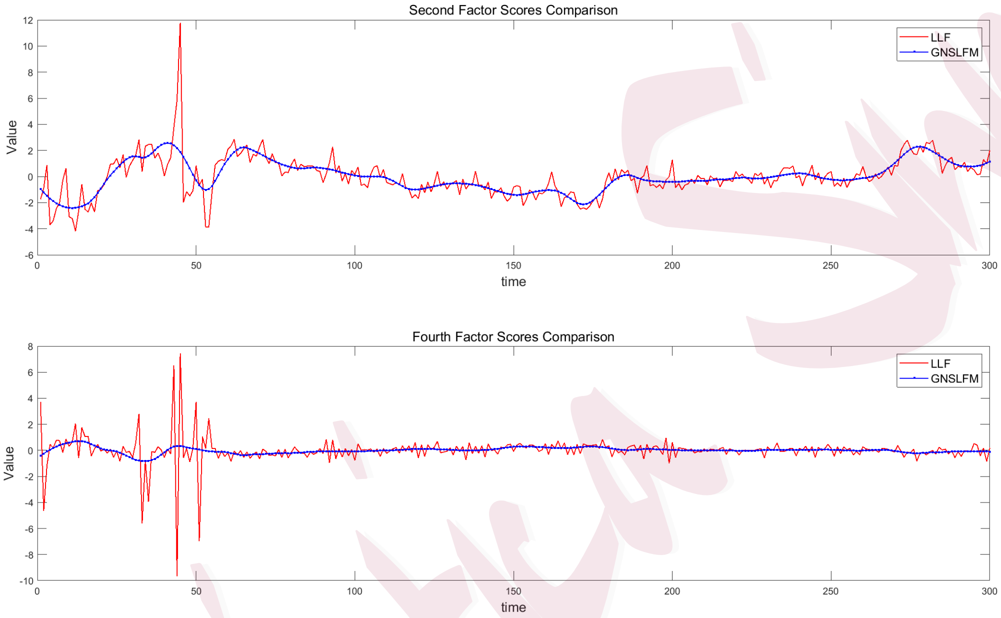

Figure 4: Factors scores of LLF and GNSLFM for the second and fourth factor

6. Discussion

We propose a general nonlinear structured latent factor model for functional data. It allows the nonlinear link functions to be multivariate and captures the correlations among observed variables using a small number of latent factors. First, identifiability of the latent factors is established by imposing certain constraints on the structured index matrix. Second, we estimate the unknown nonlinear functions by assuming Gaussian process priors, and then consider the correlation of latent factors across

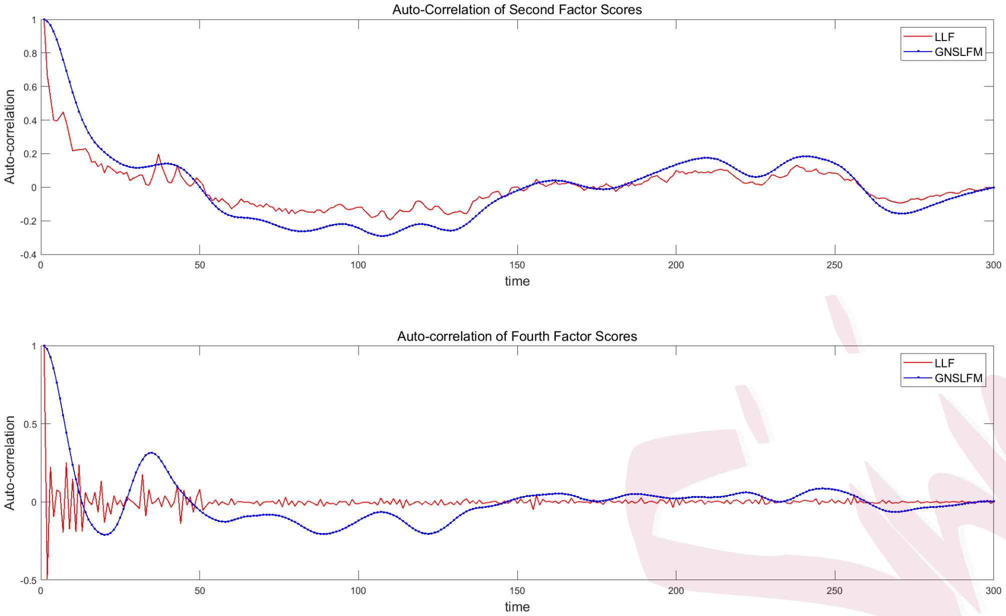

Figure 5: First-order autocorrelation function of of LLF and GNSLFM for the second and fourth factor

different t_i by assuming Gaussian process for the latent factors. We propose a two-step estimation procedure for the latent factors and unknown parameters. Finally, the posterior consistency of the nonlinear link functions, as well as the consistency of latent factors and unknown parameters, are established. Simulation studies and real-world data analysis further validate the finite-sample performance of the proposed method.

There are several extensions worth pursuing in future research. First, we assume that the factor loading structures $\boldsymbol{R}_j$ is pre-specified based on identifiability conditions. However, in practice, $\boldsymbol{R}_j$ is unknown. In Zhang (2025), we explored the automatic selection of $\boldsymbol{R}_j$ in settings with a small number of latent factors, using correlation-based clustering and residual diagnostics, and obtained promising empirical results. Developing a full general theory for this problem remains challenging

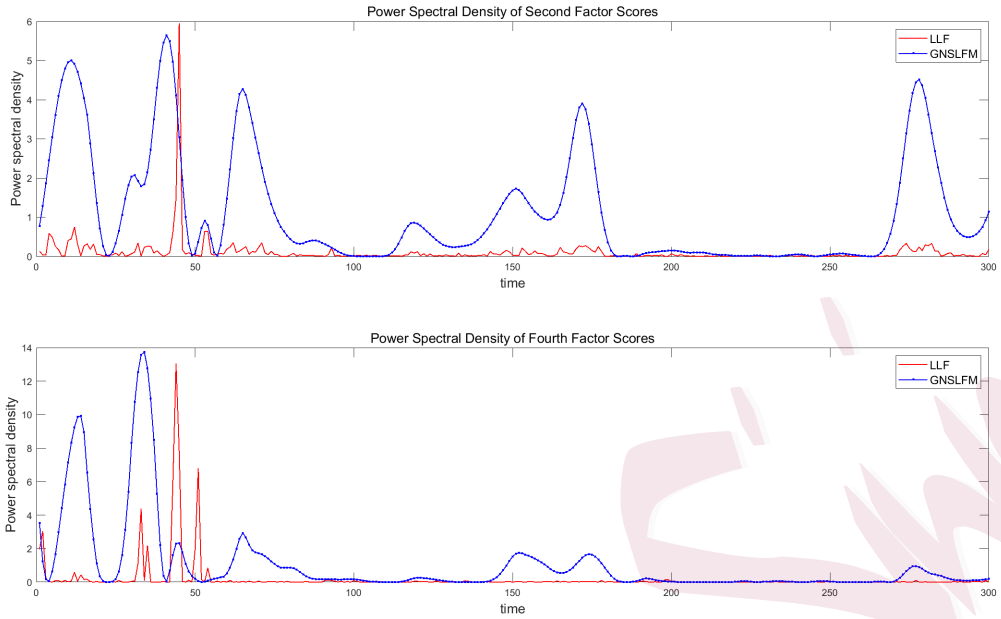

Figure 6: Power spectral density of LLF and GNSLFM for the second and fourth factor

and is left for future research. Second, theoretically, we have only proven the consistency of the unknown link functions and latent factors. We have not yet studied their asymptotic convergence rates or the convergence of the algorithm.

Supplementary Materials

The online Supplementary Materials include all the technical proofs.

Acknowledgements

The authors would like to thank two anonymous reviewers, an associate editor and the editor for constructive comments and helpful suggestions. This publication is supported by funds of the National Key R & D Program of China (2023YFA1011400), National Natural Science Foundation of China (No.12271239), Shenzhen Fundamental Research Program JCYJ20220818100602005 (No.20220111). Wang's work is also supported by Postdoctoral Fellowship Program of CPSF under Grant Number GZB20240292 and China Postdoctoral Science Foundation (No. 2024T170376).

References

Bansal, A. R. and Dimri, V. (2021). Power spectral density. *Encyclopedia of Mathematical Geosciences*, 1–3. Springer.

Basawa, I. and Rao, B. (1980). *Statistical Inference for Stochastic Processes*. Academic Press, London.

Buckley, C., Alcock, L., McArdle, R., Rehman, R. Z. U., Del Din, S., Mazza‘, C., Yarnall, A. J. and Rochester, L. (2019). The role of movement analysis in diagnosing and monitoring neurodegenerative conditions: Insights from gait and postural control. *Brain Sciences* **9**(2), 34.

Chen, Y., Li, X. and Zhang, S. (2020). Structured latent factor analysis for large-scale data: Identifiability, estimability, and their implications. *Journal of the American Statistical Association* **115**(532), 1756–1770.

Choi, T. and Schervish, M. J. (2007). On posterior consistency in nonparametric regression problems. *Journal of Multivariate Analysis* **98**(10), 1969–1987.

Choi, T., Shi, J. Q. and Wang, B. (2011). A gaussian process regression approach to a single-index model. *Journal of Nonparametric Statistics* **23**(1), 21–36.

Coube-Sisqueille, S. and Liquet, B. (2022). Improving performances of MCMC for nearest neighbor gaussian process models with full data augmentation. *Computational Statistics & Data Analysis* **168**, 107368.

Damianou, A., Titsias, M. and Lawrence, N. (2011). Variational gaussian process dynamical systems. *Advances in Neural Information Processing Systems* **24**.

Datta, A., Banerjee, S., Finley, A. O. and Gelfand, A. E. (2016). Hierarchical nearest-neighbor gaussian process models for large geostatistical datasets. *Journal of the American Statistical Association* **111**(514), 800–812.

Fang, G., Guo, J., Xu, X., Ying, Z. and Zhang, S. (2021). Identifiability of bifactor models. *Statistica Sinica*, **31**, 2309-2330.

Finley, A. O., Datta, A., Cook, B. D., Morton, D. C., Andersen, H. E. and Banerjee, S. (2019). Efficient algorithms for bayesian nearest neighbor gaussian processes. *Journal of Computational and Graphical Statistics* **28**(2), 401–414.

Gramacy, R. B. (2016). lagp: large-scale spatial modeling via local approximate gaussian processes in R. *Journal of Statistical Software* **72**, 1–46.

Gramacy, R. B. and Lee, H. K. H. (2008). Bayesian treed gaussian process models with an application to computer modeling. *Journal of the American Statistical Association* **103**(483), 1119–1130.

Gu, M., Lin, Y., Lee, V. C. and Qiu, D. Y. (2024). Probabilistic forecast of nonlinear dynamical systems with uncertainty quantification. *Physica D: Nonlinear Phenomena* **457**, 133938.

Gu, M. and Shen, W. (2020). Generalized probabilistic principal component analysis of correlated data. *Journal of Machine Learning Research* **21**(13), 1–41.

Guinness, J. (2018). Permutation and grouping methods for sharpening gaussian process approximations. *Technometrics* **60**(4), 415–429.

Hoehn, M. M. and Yahr, M. D. (1998). Parkinsonism: onset, progression, and mortality. *Neurology* **50**(2), 318–318.

Lawrence, N. (2005). Probabilistic non-linear principal component analysis with Gaussian process latent variable models. *Journal of Machine Learning Research* **6**, 1783–1816.

Leeb, W. (2021). A note on identifiability conditions in confirmatory factor analysis. *Statistics & Probability Letters* **178**, 109190.

Lin, Y., Liu, X., Segall, P. and Gu, M. (2025). Fast data inversion for high-dimensional dynamical systems from noisy measurements. *arXiv preprint* arXiv:2501.01324.

Liu, H., Cai, J., Wang, Y. and Ong, Y. S. (2018a). Generalized robust bayesian committee machine for large-scale gaussian process regression. *International Conference on Machine Learning (ICML)*, 3131–3140.

Liu, H., Ong, Y. S., Shen, X. and Cai, J. (2018b). When gaussian process meets big data: A review of scalable GPs. *IEEE transactions on neural networks and learning systems* **31**(11), 4405–4423.

McCabe, G. P. (1984). Principal variables. *Technometrics* **26**(2), 137–144.

Morris, R., Hickey, A., Del Din, S., Godfrey, A., Lord, S. and Rochester, L. (2017). A model of free-living gait: A factor analysis in parkinson's disease. *Gait & Posture* **52**, 68–71.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. The MIT Press, Cambridge, MA.

Saha, A., Datta, A. and Banerjee, S. (2022). Scalable predictions for spatial probit linear mixed models using nearest neighbor gaussian processes. *Journal of Data Science* **20**(4), 533.

Shi, J. Q. and Choi, T. (2011). *Gaussian Process Regression Analysis for Functional Data*. Chapman & Hall/CRC, London.

Spearman, C. (1904). "General Intelligence" Objectively determined and measured. *The American Journal of Psychology* **15**, 201–292.

Tu, J. H., Rowley, C. W., Luchtenburg, D. M., Brunton, S. L. and Kutz, J. N. (2014). On dynamic mode decomposition: Theory and applications. *Journal of Computational Dynamics* **1**(2), 391–421.

Villarraga, D. F. and Daziano, R. A. (2025). Hierarchical nearest neighbor gaussian process models for discrete choice: Mode choice in new york city. *Transportation Research Part B: Methodological* **191**, 103132.

Wang, B. and Shi, J. Q. (2014). Generalized gaussian process regression model for non-gaussian functional data. *Journal of the American Statistical Association* **109**(507), 1123–1133.

Wang, J., Hertzmann, A. and Fleet, D. J. (2005). Gaussian process dynamical models. *Advances in Neural Information Processing Systems* **18**.

Wang, Z., Noh, M., Lee, Y. and Shi, J. Q. (2021). A general robust t-process regression model. *Computational Statistics & Data Analysis* **154**, 107093.

Wang, Z., Shi, J. Q. and Lee, Y. (2017). Extended t-process regression models. *Journal of Statistical Planning and Inference* **189**, 38–60.

Wu, L., Pleiss, G. and Cunningham, J. P. (2022). Variational nearest neighbor gaussian process. *International Conference on Machine Learning*, 24114–24130.

Zhang, Y. (2025). Advancing latent factor analysis: Bayesian approaches for nonlinear and functional models. *PhD thesis*, Southern University of Science and Technology.

Zhang, Y., Wang, X. and Shi, J. Q. (2025). Bayesian analysis of nonlinear structured latent factor models using a gaussian process prior. *Journal of Multivariate Analysis*, In Press.

School of Mathematics and Statistics, Nanjing University of Information Science and Technology, Nanjing, 210044, China

E-mail: (sjxywxr@163.com)

Department of Statistics and Data Science, Southern University of Science and Technology, Shenzhen, 518055, China

E-mail: (12031339@mail.sustech.edu.cn)

National Center for Applied Mathematics Shenzhen & Department of Statistics and Data Science, Southern University of Science and Technology, Shenzhen, 518055, China

E-mail: (shijq@sustech.edu.cn)