Journal of Machine Learning Research 22 (2021) 1-27 Submitted 1/20; Revised 10/21; Published 11/21

Bayesian time-aligned factor analysis of paired multivariate time series

Arkaprava Roy ark007@ufl.edu
Department of Biostatistics
University of Florida
Gainnesville, FL 32611, USA

Jana Schaich Borg js524@duke.edu
Social Science Research Institute
Duke University
Durham, NC 27708-0251, USA

David B Dunson dunson@duke.edu
Department of Statistical Science
Duke University
Durham, NC 27708-0251, USA

Editor: Barbara Engelhardt

Abstract

Many modern data sets require inference methods that can estimate the shared and individual-specific components of variability in collections of matrices that change over time. Promising methods have been developed to analyze these types of data in static cases, but only a few approaches are available for dynamic settings. To address this gap, we consider novel models and inference methods for pairs of matrices in which the columns correspond to multivariate observations at different time points. In order to characterize common and individual features, we propose a Bayesian dynamic factor modeling framework called Time Aligned Common and Individual Factor Analysis (TACIFA) that includes uncertainty in time alignment through an unknown warping function. We provide theoretical support for the proposed model, showing identifiability and posterior concentration. The structure enables efficient computation through a Hamiltonian Monte Carlo (HMC) algorithm. We show excellent performance in simulations, and illustrate the method through application to a social mimicry experiment.

Keywords: CIFA; Dynamic factor model; Hamiltonian Monte Carlo; JIVE; Monotonicity; Paired time series; Social mimicry; Time alignment; Warping.

1. Introduction

Many fields are routinely collecting matrix-variate data and asking questions about the similarity between subsets of those data. As the collection of these types of data expands, so does the need for new statistical methods that can capture the shared and individual-specific structure in multiple matrices, especially when matrices in a collection consist of multivariate observations collected over time. Here, we are motivated by the particular challenge of measuring the coordination between two people interacting dynamically. Many scientific questions require measurements of how similar the movements and expressions of two people are in these cases, because such similarity has been shown to be related to many interested phenomena and behaviors, including much people like each other or cooperate (Lakin and Chartrand, 2003; Johnston, 2002; Marsh et al., 2016). To address these questions, videos of social interactions are typically recorded, and the coordinates of different facial and body features from each individual in the pair are extracted over time. The data for each individual form a matrix, with the columns corresponding to different time points. One component of the variability in the two matrices will be attributable to shared structure, such as the patterns in which lips tend to move during conversation. Another component will be attributable to variability specific to each individual, such as differences in smile shapes, camera placements, sitting postures, and head sizes. When people interact, they often subconsciously imitate each other, but who initiates the imitation and the speed at which the imitation occurs varies over time. Thus, modeling the similarity in these paired dynamic matrix-variate data requires a strategy that can accommodate: 1) complex multivariate dependence among variables, and 2) dynamic time-varying lags between the two multivariate time series. Although our motivating example is from human social interactions, similar challenges are posed by other types of paired multivariate data, such as that collected in animal behavior studies, cellular imaging studies, finance, or handwriting recognition where there is interest in how similar the behaviors of two mice, the spiking of two cells, the rates of two stocks, or samples of two signatures are.

The individual-specific spaces will account for the variations due to camera placements, sitting postures, head size/shape, etc. Likewise, the time lag between the two participants may also change depending on the change in the direction of mimicry, complexity of the gesture, etc. In one of our real data illustrations, we have participants switching their roles as leader and follower in the middle of their mimicry session. Thus, analyzing these paired dynamic matrix-variate data requires a strategy that can accommodate two significant challenges: 1) complex multivariate dependence among variables, and 2) dynamic time-varying lags between the two multivariate time series. Here, dynamic time-varying lag refers to the situation when the lag dependency order between the two multivariate time series changes over time. Although our motivating example is from human social interactions, similar challenges are posed by other types of paired multivariate data, such as that collected in animal behavior studies, cellular imaging studies, finance, or speech, gesture, and handwriting recognition.

Joint and Individual Variation Explained (JIVE) (Lock et al., 2013) and Common and Individual Feature Analysis (CIFA) (Zhou et al., 2016) were developed to capture shared and individual-specific features in pairs of multivariate matrices. In the case of JIVE, the data X_i's are decomposed into three parts: a low-rank approximation of joint structure J_i, a low-rank approximation of individual variation S_i, and an error E_i under the restriction $JS_i^T = 0$ for all i. Here J is the matrix stacking J_i's on top of each other. The CIFA decomposition defines a matrix factorization problem: $\min_{A,A_i,B_i,\tilde{B}_i} \|Y_i - (A, A_i)^T(B_i, \tilde{B}_i)\|_F^2$ under the restriction that $A^T A_i = 0$ for all i, with $\| \cdot \|_F$ denoting the Frobenius norm. Thus, the shared subspace of the data matrix Y_i in the CIFA decomposition is AB_i and the individual specific subspace is characterized by $A_i\tilde{B}_i$. Due to the assumed orthogonality between the columns of A and A_i, the shared and individual-specific spaces become orthogonal. Extensions of these methods are proposed in Li and Gaynanova (2018) and Feng et al. (2018). Related approaches have been used in behavioral research (Schouteden et al., 2014), genomic clustering (Lock and Dunson, 2013; Ray et al., 2014), railway network analysis (Jere et al., 2014), etc. In most cases, frequentist frameworks are used for inference, the methods are not likelihood-based, and the focus is on static data. De Vito et al. (2021) developed a method for multigroup factor analysis in a Bayesian framework, which has some commonalities with these approaches but does not impose orthogonality.

One way to accommodate time-varying lags is to temporally align the features in a shared space, avoiding the need to develop a complex model of lagged dependence across the series. However, time alignment is a hard problem. Typically, alignment is done in a first stage, and then an inferential model is applied to the aligned data (Vial et al., 2009). However, such two-stage approaches do not provide adequate uncertainty quantification. Trigeorgis et al. (2017) also considered a problem of time aligned image analysis. Their proposed loss function combines costs for non-linear discriminant analysis and dynamic time warping. They further modelled the unknown non-linear functions using deep neural-nets. Unlike our approach, their method does not adjust for individual-specific variations.

Several approaches have been proposed to model warping functions. Tsai et al. (2013) used basis functions similar to B-splines with varying knot positions, using stochastic search variable selection for the knots. This makes the model more flexible, but at the cost of very high computational demand. Kurtek (2017) put a prior on the warping function based on a geometric condition and developed importance sampling methods. Extending their geometric characterization to the multivariate case is not straightforward; hence it is difficult to extend their method to our setting. Lu et al. (2017) use a similar structure in placing a prior on the warping function.

Bharath and Kurtek (2017); Cheng et al. (2016) put a Dirichlet prior on the increments of the warping function over a grid of time points. Thus, the estimated warping function is not smooth. Also, when the warping function is convolved with an unknown function, computation becomes inefficient due to poor mixing. The concept of warplets of Claeskens et al. (2010) is very interesting. Nevertheless, this method also suffers from a similar computational problem.

For multivariate time warping, Listgarten et al. (2005) proposed a method based on a hidden Markov model. Other works propose to use a warping based distance to cluster similar time series (Orsenigo and Vercellis, 2010; Che et al., 2017). Unfortunately, these algorithms require the two time series to be collected at the same time points. In addition, it is difficult to avoid a two-stage procedure, since there is no straightforward way to combine a statistical model with the warping algorithms.

Gervini and Gasser (2004) modeled the warping function as $M(t) = t + \sum_j s_j f_j(t)$, where $f_j(t)$'s are characterized using B-splines with the sum of the s_j's equal to zero. For identifiability, they assumed restrictive conditions on the spline coefficients and did not accommodate multivariate data. Telesca and Inoue (2008) developed a related Bayesian approach, but their structure makes it difficult to apply gradient-based MCMC, and finding a good proposal for efficient sampling is problematic.

We propose to estimate the similarity between two multivariate time series with time-varying lags using a Bayesian dynamic factor model that incorporates time warping and parameter estimation in a single step. Our proposed dynamic factor model is different from traditional state-space models (Aguilar and West, 2000). Instead of assuming any Markovian propagation of the latent factors, we assume the latent factors to vary smoothly over time t. We further assume the multivariate time series have both time-aligned shared factors and individual-specific factors. Estimating the shared factors is to assess similarity between the time series, while the main goal of the individual factors is to ensure the inference is robust. The resulting model reduces to a CIFA-style dependence structure, but unlike previous work, we accommodate time dependence and take a Bayesian approach to inference. Key aspects of our Bayesian implementation include likelihood-based estimation of shared and individual-specific subspaces, incorporation of a monotonicity constraint on the warping function for identifiability, and development of an efficient gradient-based Markov chain Monte Carlo (MCMC) algorithm for posterior sampling.

We align the two time series by mapping the features of the shared space using a monotone increasing warping function $M : [0,1] \to [0,1]$. If we have two univariate time-varying processes $a(t)$ and $b(t)$, then the warping function M is generally computed as the minimizer of $d(a(t), b(M(t)))$ for some distance metric d. To ensure identifiability of M in this minimization problem, we need to further assume that $M(0) = 0, M(1) = 1$ and $M(t)$ is monotone increasing. This flexible function $M(t)$ can accommodate situations where the time lags between the multivariate time series change sign and direction. Our monotone function construction differs from previous Bayesian approaches (Ramsay et al., 1988; He and Shi, 1998; Neelon and Dunson, 2004; Shively et al., 2009; Lin and Dunson, 2014), motivated by tractability in obtaining a nonparametric specification amenable to Hamiltonian Monte Carlo (HMC) sampling.

In general, posterior samples of the loading matrices are not interpretable without identifiability restrictions (Seber, 2009; Lopes and West, 2004; Ročková and George, 2016; Fruehwirth-Schnatter and Lopes, 2018). To avoid arbitrary constraints, which complicate computation, one technique is to post-process an unconstrained MCMC chain. Aßmann et al. (2016) post-process by solving an Orthogonal Procrustes problem to produce a point estimate of the loading matrix, but without uncertainty quantification. We consider to post-process the MCMC chain iteratively so that it becomes possible to draw inference based on the whole chain. Apart from the computational advantages, we also show identifiability of the warping function in our factor modeling setup both in theory and simulations. Moreover, our identifiability result is more general than the result in Gervini and Gasser (2004) as we do not assume any particular form of the warping function other than monotonicity and also it has been derived in a multivariate setting.

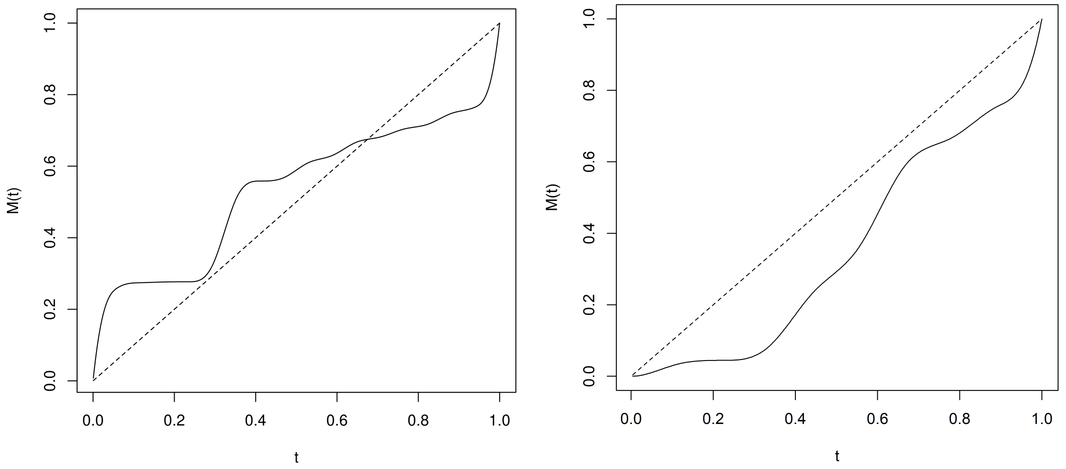
In section 2 we discuss our model in detail. Prior specifications are described in Section 3. Our computational scheme is outlined in Section 4. Section 5 discusses theoretical properties such as identifiability of the warping function and posterior concentration. We study the performance of our method in two simulation setups in Section 6. Section 7 considers applications to human social interaction datasets. We end with some concluding remarks in Section 8. Supplementary Materials have all the proofs, additional algorithmic details, and additional results.

2. Modeling

We have a pair of p dimensional time varying random variables $\mathbf{x}_t$ and $\mathbf{y}_t$. We propose to model the data as a function of time varying shared latent factors, $\boldsymbol{\eta}(t) = \{\eta_1(t), \ldots, \eta_r(t)\}$, and individual-specific factors, $\boldsymbol{\zeta}_1(t) = \{\zeta_{11}(t), \ldots, \zeta_{1r_1}(t)\}$ and $\boldsymbol{\zeta}_2(t) = \{\zeta_{21}(t), \ldots, \zeta_{2r_2}(t)\}$. We do time alignment through the shared factors in $\eta(t)$ using warping functions $M_1(t), \ldots, M_r(t)$. Here M_i is the warping function for the latent variable η_i.

Latent factor modeling is natural in this setting in relating the measured multivariate time series to lower-dimensional characteristics, while reducing the number of parameters needed to represent the covariance. Since we are using the warping function to align the time-varying factors of the shared space, to ensure identifiability, the individual-specific space and the shared space are required to be orthogonal. Thus, the corresponding loading matrices of the two orthogonal subspaces are assumed to have orthogonal column spaces. Let $\boldsymbol{\Lambda}$ be the loading matrix of the shared space. Then the shared space signal belongs to the span of the columns of $\boldsymbol{\Lambda}$ with weights as some multiple of the shared factors $\boldsymbol{\eta}(t) = \{\eta_1(t), \ldots, \eta_r(t)\}$. An element from the time-varying shared space can be represented as $\sum_{j=1}^{r} a_j \boldsymbol{\Lambda}_{.j}\eta_j(t)$ for some constant $(a_1, \ldots, a_r) \in \mathbb{R}^r$ where $\boldsymbol{\Lambda}_{.j}$ is the *j-th* column of $\boldsymbol{\Lambda}$. Alternatively it can also be written as $\boldsymbol{\Lambda}\boldsymbol{\Xi}_1\beta(t)$, where $\boldsymbol{\Xi}_1$ is a diagonal matrix with entries $(a_1, \ldots, a_r)$. The individual-specific space is assumed to be in the orthogonal subspace of the column space of $\boldsymbol{\Lambda}$. Thus we use the orthogonal projection matrix $\boldsymbol{\Psi} = \mathbf{1} - \boldsymbol{\Lambda}(\boldsymbol{\Lambda}^T\boldsymbol{\Lambda})^{-1}\boldsymbol{\Lambda}^T$ to construct the loading matrix of the individual-specific part of each signal. The loading matrix for the individual-specific space $\mathbf{x}_t$ is assumed to be $\boldsymbol{\Psi}\boldsymbol{\Gamma}_1$ for some matrix $\boldsymbol{\Gamma}_1$ of dimension $p \times r_1$, where r_1 is the rank. The corresponding loading matrix for the individual-specific space of $\mathbf{y}_t$ is $\boldsymbol{\Psi}\boldsymbol{\Gamma}_2$, with $\boldsymbol{\Gamma}_2$ being a $p \times r_2$ matrix with r_2 the rank. The shared signals of $\mathbf{x}_t$ and $\mathbf{y}_t$ are $\boldsymbol{\Lambda}\boldsymbol{\eta}(t)$ and $\boldsymbol{\Lambda}\boldsymbol{\eta}_1(t)$. In order to align the two shared spaces, we further assume that the factors in $\boldsymbol{\eta}_1(t)$ are a warped version of the factors in $\boldsymbol{\eta}(t)$. For simplicity, we assume that there is a single warping function that holds for all the latent factors.

The warping function $M : [0, 1] \to [0, 1]$ is assumed to be monotone increasing, which is important for identifiability. As motivation, consider the case of social interactions. People often imitate each other subconsciously. In a normal conversation, people take turns mimicking each other without knowing it. Let us assume that A and B are playing a game where they take turns mimicking each other so that sometimes A mimics B and sometimes B mimics A. This motivates us to model this mimicry to assess how similar A and B's gestures are. By the definition of a warping function, if person A makes a gesture at time t, person B does the same gesture at $M(t)$. If one person mimics the other almost instantly, we must have $t = M(t)$. Hence, in Figure 1, the dashed line through the origin with slope one corresponds to the case when there is no lag among the participants. However such instantaneous mimicry is often unrealistic. Thus it might be either $t < M(t)$ or $t > M(t)$ depending on whether individual A or B makes the gesture for the first time. A method that models this mimicry would need to be able to account for the fact that the roles change dynamically over time. In Figure 1, we illustrate behavior of the warping function in two possible experimental situations that we consider in our real data illustration. Hence, panel (a) shows the warping function when one individual is mimicking the other for the first part of the experiment, and then the leader shifts. In the panel (b) experiment, the leader remains the same throughout. Both of these functions are estimated based on real data.



(a) The direction of mimicry changes

(b) The direction of mimicry does not change

Figure 1: Estimated warping functions for two social mimicry experiments (solid lines). The dashed line is when individual 1 has perfectly aligned behaviors as individual 2.

To model a smooth monotone increasing warping function bounded in [0, 1] such that $M(0) = 0$ and $M(1) = 1$ we use a B-spline expansion with J many bases as follows,

$$M(t) = \sum_{j=1}^{J} \gamma_j B_j(t), \gamma_{ij} = \frac{\sum_{\ell=2}^{j} \exp(\kappa_\ell)}{\sum_{k=2}^{J} \exp(\kappa_k)}, \quad \gamma_1 = 0,$$

where $B_j(\cdot)$'s are B-spline basis functions and $\kappa_k \in (-\infty, \infty)$. To restrict $M(t)$ to be monotone increasing and bounded between [0, 1], it is sufficient to have the B-spline coefficients $\{\gamma_j\}_{j=1}^{J}$ be monotone increasing in index j and bounded between [0, 1] (De Boor, 1978). This construction restricts M to be a smooth monotone increasing function such that $M(0) = 0$ and $M(1) = 1$. These are the desired properties of a warping function. A short review on B-splines is provided in Section 1 of the supplementary materials.

For simplicity, we consider a single warping function for all the shared latent variables. The complete model that we consider is

$$\begin{aligned}
\mathbf{x}_t =& \boldsymbol{\Psi}\boldsymbol{\Gamma}_1\boldsymbol{\zeta}_1(t) + \boldsymbol{\Lambda}\boldsymbol{\Xi}_1\boldsymbol{\eta}(t) + \boldsymbol{\epsilon}_{1t}, & (1)\\
\mathbf{y}_t =& \boldsymbol{\Psi}\boldsymbol{\Gamma}_2\boldsymbol{\zeta}_2(t) + \boldsymbol{\Lambda}\boldsymbol{\Xi}_2\boldsymbol{\eta}(M(t)) + \boldsymbol{\epsilon}_{2t}, & (2)\\
\zeta_{ij}(t) =& \sum_{j=1}^{K_i} \beta_{ilj}B_j(t), \quad i = 1,2; j = 1,\dots r_i, & (3)\\
\eta_i(t) =& \sum_{j=1}^{K} \beta_{ij}B_j(t), & (4)\\
M(t) =& \sum_{j=1}^{J} \gamma_j B_j(t), & (5)\\
\gamma_j =& \frac{\sum_{l=2}^{j}\exp(\kappa_l)}{\sum_{k=2}^{J}\exp(\kappa_k)}, \quad \gamma_1 = 0, & (6)\\
\epsilon_{it} \sim& \mathrm{N}(0,\boldsymbol{\Sigma}_i), \quad \boldsymbol{\Sigma}_i = \text{diagonal}(\sigma_{i1}^2,\dots,\sigma_{ip}^2), & (7)
\end{aligned}$$

where $\boldsymbol{\Lambda}$, $\boldsymbol{\Gamma}_1, \boldsymbol{\Gamma}_2$ are static factor loading matrices of dimension $p \times r$,$p \times r_1$ and $p \times r_2$, respectively, with $\boldsymbol{\Psi} = \mathbf{I}_p - \boldsymbol{\Lambda}(\boldsymbol{\Lambda}^T\boldsymbol{\Lambda})^{-1}\boldsymbol{\Lambda}^T$; $\boldsymbol{\Xi}_1$ and $\boldsymbol{\Xi}_2$ are $r \times r$ diagonal matrices; r is the number of shared time varying latent factors and r_1, r_2 are the number of individual-specific latent factors for the 1st and 2nd individual, respectively; the error variances are given by $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$. In (1) and (2), we define $\boldsymbol{\eta}(t) = \{\eta_i(\cdot) : 1 \le i \le r\}$ as the vector of shared time-varying factors. Similarly, we define the individual-specific array of time-varying factors $\boldsymbol{\zeta}_1(t) = \{\zeta_{1j}(\cdot) : 1 \le j \le r_1\}$ and $\boldsymbol{\zeta}_2(t) = \{\zeta_{2j}(\cdot) : 1 \le j \le r_2\}$. In (3), we denote the number of B-spline bases to model individual-specific factors of the *i-th* individual by K_i. To model the shared time-varying latent factors, $\eta_i(\cdot)$'s, we use K B-spline bases in (4). The number of bases to model the warping function in (5) is J. The constraint $\gamma_1 = 0$ ensures $M(0) = 0$ and the softmax type reparametrization ensures monotonicity. Under the above characterization, we have $\eta_i(M(t)) = \sum_{j=1}^{K}\beta_{ij}B_j\{\sum_{\ell=1}^{J}\gamma_\ell B_\ell(t)\}$.

A schematic representation of our proposed model is shown in Figure 2. We project the individual-specific loading matrices on the orthogonal space of the shared space spanned by columns of $\boldsymbol{\Lambda}$ using $\boldsymbol{\Psi}$. The data are collected over T time points longitudinally for individual 1 and 2 respectively, and X and Y are $p \times T$ and $p \times T$ dimensional data matrices. Correspondingly, $\boldsymbol{\Psi}\boldsymbol{\Gamma}_1\boldsymbol{\zeta}$ and $\boldsymbol{\Lambda}\boldsymbol{\Xi}_1\beta$ are the individual-specific mean and shared space mean of X, respectively. The columns of these two matrices are orthogonal due to the orthogonality of $\boldsymbol{\Psi}$ and $\boldsymbol{\Lambda}$. Since $\boldsymbol{\zeta}_1(t)$ and $\boldsymbol{\eta}(t)$ are modeled independently, the rows of the two means are also independent in probability. A similar result holds for Y. Thus, this model conveniently explains both joint and individual variations.

The loading matrix $\boldsymbol{\Lambda}$ identifies the shared space of the two signals. We assume a single shared set of latent factors $\boldsymbol{\eta}(t)$ for both X_t and Y_t. The warping function $M(t)$ aligns those for the Y_t series relative to the $\mathbf{x}_t$ series. Then we have individual-specific factors $\boldsymbol{\zeta}_1(t), \boldsymbol{\zeta}_2(t)$ and factor loading matrices $\boldsymbol{\Psi}\boldsymbol{\Gamma}_1, \boldsymbol{\Psi}\boldsymbol{\Gamma}_2$ that can accommodate within series covariances in $\mathbf{x}(t)$ and $\mathbf{y}(t)$. We call our proposed method Time Aligned Common and Individual Factor Analysis (TACIFA).

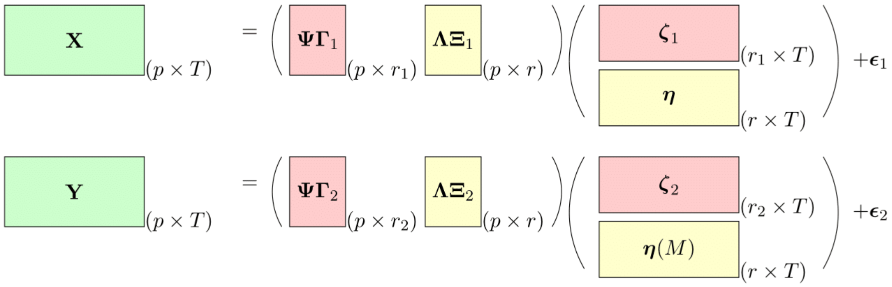


Figure 2: A schematic representation of our proposed model where the dimensions of the matrices are illustrated at the bottom right corner, and $\boldsymbol{\eta}(M)$ stands for time aligned factors from $\boldsymbol{\eta}$ using the warping function $M(t)$ and $\boldsymbol{\Psi} = \mathbf{I}_p - \boldsymbol{\Lambda}(\boldsymbol{\Lambda}^T\boldsymbol{\Lambda})^{-1}\boldsymbol{\Lambda}^T$. The dimensions of the individual matrices, $\boldsymbol{\Lambda}$, $\boldsymbol{\Gamma}_1, \boldsymbol{\Gamma}_2$ are $p \times r$,$p \times r_1$ and $p \times r_2$ respectively. The other two matrices $\boldsymbol{\Xi}_1$ and $\boldsymbol{\Xi}_2$ are $r \times r$ diagonal matrices. Additionally, in the Figure, $\mathbf{X} = [\mathbf{x}_1; \cdots; \mathbf{x}_T]$,$\mathbf{Y} = [\mathbf{y}_1; \cdots; \mathbf{y}_T]$, $\boldsymbol{\zeta}_1 = [\boldsymbol{\zeta}_1(1); \cdots; \boldsymbol{\zeta}_1(T)]$,$\boldsymbol{\zeta}_2 = [\boldsymbol{\zeta}_2(1); \cdots; \boldsymbol{\zeta}_2(T)]$ and $\boldsymbol{\eta} = [\boldsymbol{\eta}(1); \cdots; \boldsymbol{\eta}(T)]$

3. Prior specification

We use priors similar to those in Bhattacharya and Dunson (2011) for $\boldsymbol{\Lambda}$, $\boldsymbol{\Gamma}_1$ and $\boldsymbol{\Gamma}_2$ to allow for automatic selection of rank. We try to maintain conjugacy as much as possible for easier posterior sampling. For clarity, we define $\boldsymbol{\kappa} = \{\kappa_j : 2 \le j \le J\}$ and $\boldsymbol{\beta} = \{\beta_{ij} : 1 \le j \le r_i, 1 \le i \le 2, \}$ The detailed prior description for $\boldsymbol{\kappa}, \boldsymbol{\beta}, \boldsymbol{\Lambda}, \boldsymbol{\Xi}_1, \boldsymbol{\Xi}_2, \boldsymbol{\Gamma}_1, \boldsymbol{\Gamma}_2, \boldsymbol{\sigma}_1$ and $\boldsymbol{\sigma}_2$ is described below,

$$\Lambda_{lk}|\phi_{1,lk}, \tau_{1k} \sim \mathrm{N}(0, \phi_{1,lk}^{-1}\tau_{1k}^{-1}), \quad 1 \le l \le p, 1 \le k \le r, \tag{8}$$

$$\phi_{1,lk} \sim \mathrm{Gamma}(\nu_1, \nu_1), \quad \tau_{1k} = \prod_{i=1}^{k} \delta_{mi}, \quad 1 \le l \le p, 1 \le k \le r, \tag{9}$$

$$\delta_{1,1} \sim \mathrm{Gamma}(\alpha_1, 1), \quad \delta_{1,i} \sim \mathrm{Gamma}(\alpha_2, 1), 1 \le i \le r, \tag{10}$$

$$\Gamma_{1,lk}|\phi_{11,lk}, \tau_{11k} \sim \mathrm{N}(0, \phi_{11,lk}^{-1}\tau_{11k}^{-1}), \quad 1 \le l \le p, 1 \le k \le r_1, \tag{11}$$

$$\phi_{11,lk} \sim \mathrm{Gamma}(\nu_1, \nu_1), \quad \tau_{11k} = \prod_{i=1}^{k} \delta_{mi} \tag{12}$$

$$\delta_{11,1} \sim \mathrm{Gamma}(\alpha_{111}, 1), \quad \delta_{11,i} \sim \mathrm{Gamma}(\alpha_{112}, 1), \tag{13}$$

$$\Gamma_{2,lk}|\phi_{12,lk}, \tau_{12k} \sim \mathrm{N}(0, \phi_{12,lk}^{-1}\tau_{12k}^{-1}), \quad 1 \le l \le p, 1 \le k \le r_2, \tag{14}$$

$$\phi_{12,lk} \sim \mathrm{Gamma}(\nu_1, \nu_1), \quad \tau_{12k} = \prod_{i=1}^{k} \delta_{mi}, \tag{15}$$

$$\delta_{12,1} \sim \mathrm{Gamma}(\alpha_{121}, 1), \quad \delta_{12,i} \sim \mathrm{Gamma}(\alpha_{122}, 1), \tag{16}$$

$$\sigma_{1l}^{-2} \sim \mathrm{Gamma}(\alpha_1, \alpha_1), \quad \sigma_{2l}^{-2} \sim \mathrm{Gamma}(\alpha_2, \alpha_2), \quad 1 \le l \le p \tag{17}$$

$$\Xi_{1,ll}, \Xi_{2,ll}, \kappa_j, \beta_{qk}\beta_{siK_s} \sim N(0, \omega), \tag{18}$$

for $1 \leq k \leq K$,$q = 1, \ldots, r$ $1 \leq j \leq J$, $i = 1, \ldots, r_s$,$s = 1, 2$ and $l = 1, \ldots, r$. Higher values of α_{m2} ensure increasing shrinkage as we increase rank.

We initially set the number of factors to a conservative upper bound. Then the multiplicative gamma prior will tend to induce posteriors for τ_k^{-1} in the later columns that are concentrated near zero. Those columns in $\boldsymbol{\Lambda}$ will tend to zero. Thus, the corresponding factors are then effectively deleted. The extra factors in the model may either be left, as they will have essentially no impact, or may be removed via a factor selection procedure which will remove the columns having entries within $\pm\zeta$ of zero. We follow the second strategy, motivated by our goal of obtaining a few interpretable factors. In particular, we apply the adaptive MCMC procedure of Bhattacharya and Dunson (2011) with $\zeta = 1 \times 10^{-3}$.

4. Computation

We use Gibbs updates for all the parameters except for $\boldsymbol{\Lambda}$ and $\boldsymbol{\kappa}$; details are provided in Section 2 of Supplementary Materials. For $\boldsymbol{\Lambda}$ and $\boldsymbol{\kappa}$, we propose an efficient gradient-based MCMC algorithm. For our proposed model, we can easily calculate the derivative of the log-likelihood with respect to $\boldsymbol{\kappa}$ using derivatives of B-splines (De Boor, 1978). This parameter $\boldsymbol{\kappa}$ is only involved in the model of $\mathbf{y}_t$. The negative of that log-likelihood function including the prior on $\boldsymbol{\kappa}$ is

$$L(\boldsymbol{\kappa}) = \sum_{t=1}^{T}\sum_{i=1}^{p}\frac{1}{\sigma_{2i}^2}\left[Y_{it} - \Psi_{2i}\zeta_2(t) - \Lambda_{2i}\eta\Big\{\sum_{j=1}^{J}\frac{\sum_{l=2}^{j}\exp(\kappa_l)}{\sum_{k=2}^{J}\exp(\kappa_k)}B_j(t)\Big\}\right]^2 + \frac{\sum_{j=2}^{J}\kappa_j^2}{2\omega^2}.$$

For simplicity in expression of the derivative, let us denote A_{it} $= \Lambda_{2i}\eta\big(\sum_{j=1}^{J}\frac{\sum_{l=2}^{j}\exp(\kappa_l)}{\sum_{k=2}^{J}\exp(\kappa_k)}B_j(t)\big)$ and $M(t) = \sum_{j=1}^{J}\frac{\sum_{l=2}^{j}\exp(\kappa_l)}{\sum_{k=2}^{J}\exp(\kappa_k)}B_j(t)$, as defined earlier. Then the derivative is given by

$$\begin{aligned} L'(\kappa_j) = &- \sum_{t=1}^{T}\sum_{i=1}^{p}\frac{1}{\sigma_{2i}^2}(Y_{it} - \Psi_{2i}\zeta_2(t) - A_{it})A_{it}\Bigg[\sum_{l=j}^{J}B_l(t) \\ &- M(t)\Bigg]\exp(\kappa_j)/\sum_{k=2}^{J}\exp(\kappa_k) + \kappa_j/\omega^2. \end{aligned}$$

Let us denote $L'(\boldsymbol{\kappa}) = (L'(\kappa_2), \ldots, L'(\kappa_J))'$.

Now, we discuss the sampling for Λ. To update the *j-th* column of Λ, we first rewrite the orthogonal projection matrix using the matrix inverse result of block matrices as

$$\boldsymbol{\Psi} = (\mathbf{1} - \mathbf{P}_1)(\mathbf{1} - \mathbf{P}_2)(\mathbf{1} - \mathbf{P}_1)$$

where $\mathbf{P}_1 = \boldsymbol{\Lambda}_{.-j}(\boldsymbol{\Lambda}_{-j}^T\boldsymbol{\Lambda}_{-j})^{-1}\boldsymbol{\Lambda}_{.-j}^T$ and $\mathbf{P}_2 = \boldsymbol{\Lambda}_{.j}(\boldsymbol{\Lambda}_j^T(\mathbf{1} - \mathbf{P}_1)\boldsymbol{\Lambda}_j)^{-1}\boldsymbol{\Lambda}_{.j}^T$. Here $\boldsymbol{\Lambda}_{.-j}$ is the reduced matrix after removing the *j-th* column of $\boldsymbol{\Lambda}$ and $\boldsymbol{\Lambda}_{.j}$ is the *j-th* column. The negative log-likelihood with respect to $\boldsymbol{\Lambda}_{.j}$ is

$$
\begin{aligned}
L_1(\mathbf{\Lambda}_{.j}) = &\sum_t \sum_{i=1}^{p} (\mathbf{X} - \mathbf{\Psi}\mathbf{\Gamma}_1\boldsymbol{\zeta}_1(t) - \mathbf{\Lambda}\mathbf{\Xi}_1\boldsymbol{\eta}(t))^2/(2\boldsymbol{\sigma}_1^2) \\
&+ \sum_t \sum_{i=1}^{p} (\mathbf{Y} - \mathbf{\Psi}\mathbf{\Gamma}_2\boldsymbol{\zeta}_1(t) - \boldsymbol{\lambda}\Xi_2\boldsymbol{\eta}(t))^2/(2\boldsymbol{\sigma}_2^2) + \sum_k \Lambda_{kj}^2/(2\phi_{1,kj}\tau_j),
\end{aligned}
$$

and the derivative is

$$
\begin{aligned}
L_1'(\Lambda_{kj}) = \sum_t \sum_{i=1}^{p} (X_{ti} - \mathbf{\Psi}\mathbf{\Gamma}_1\boldsymbol{\zeta}_1(t) - \mathbf{\Lambda}\mathbf{\Xi}_1\boldsymbol{\eta}(t))(B_{ti} - \boldsymbol{\eta}(t))/(\boldsymbol{\sigma}_1^2) + \sum_t \sum_{i=1}^{p} (Y_{ti} - \\
\mathbf{\Psi}\mathbf{\Gamma}_2\boldsymbol{\zeta}_2(t) - \mathbf{\Lambda}\mathbf{\Xi}_2\boldsymbol{\eta}(M(t)))(B_{ti} - \boldsymbol{\eta}(t))/(\boldsymbol{\sigma}_2^2) + \Lambda_{kj}/(\phi_{1,kj}\tau_j),
\end{aligned}
$$

where

$$
\begin{aligned}
\mathbf{B} =& - (\mathbf{1} - \mathbf{P}_1)\mathbf{Q}(\mathbf{1} - \mathbf{P}_1) \\
\mathbf{Q} =& \Big\{ \big((\mathbf{\Lambda}_{.j}\mathbf{e}_k^T + \mathbf{e}_k\mathbf{\Lambda}_{.j}^T)\mathbf{\Lambda}_{.j}^T(\mathbf{1} - \mathbf{P}_1)\mathbf{\Lambda}_{.j} \\
&\quad - 2\mathbf{e}_k(\mathbf{1} - \mathbf{P}_1)\mathbf{\Lambda}_{.j}^T\mathbf{\Lambda}_{.j}(\mathbf{1} - \mathbf{P}_1)\mathbf{\Lambda}_{.j}^T \big) / (\mathbf{\Lambda}_{.j}^T(\mathbf{1} - \mathbf{P}_1)\mathbf{\Lambda}_{.j})^2 \Big\},
\end{aligned}
$$

with $\mathbf{e}_k$ a vector of length p having 1 at the *k-th* position and zero elsewhere.

Relying on the above gradient calculations we use HMC (Duane et al., 1987; Neal et al., 2011). We keep the leapfrog step fixed at 30. We tune the step size parameter to maintain an acceptance rate within the range of 0.6 to 0.8. If the acceptance rate is less than 0.6, we reduce the step length and increase it if the acceptance rate is more than 0.8. We do this adjustment after every 100 iterations. We also incorporate removal of columns of $\mathbf{\Lambda}$, $\mathbf{\Gamma}_1$ and $\mathbf{\Gamma}_2$ if the contributions are below a certain threshold as described in Section 3.2 of Bhattacharya and Dunson (2011).

4.1 Post-MCMC inference

Here we discuss the strategy to infer the loading matrix $\mathbf{\Lambda}_1 = \mathbf{\Lambda}\mathbf{\Xi}_1$. The loading matrices are identifiable up to an orthogonal right rotation. This implies that $(\mathbf{\Lambda}_1, \boldsymbol{\eta}(t))$ and $(\mathbf{\Lambda}_1\mathbf{R}, \mathbf{R}^T\boldsymbol{\eta}(t))$ for some orthonormal matrix R have equivalent likelihood. In our modeling framework, we may write $\boldsymbol{\eta}(t) = \boldsymbol{\beta}\mathbf{B}_t$, where $\boldsymbol{\beta} = (\!(\beta_{ij})\!)_{1\leq i\leq r, 1\leq j\leq K}$ is the coefficient matrix and $\mathbf{B}_t = (B_1(t), \ldots, B_K(t))$ is the array of K B-spline bases evaluated at t. Thus, $\mathbf{R}^T\boldsymbol{\eta}$ gives us a new array of latent factors with coefficient matrix $\mathbf{R}^T\boldsymbol{\beta}$. However, the same likelihood is obtained for values of $(\mathbf{\Lambda}_1, \boldsymbol{\eta}(t))$ or $(\mathbf{\Lambda}_1\mathbf{R}, \mathbf{R}^T\boldsymbol{\eta}(t))$, implying non-identifiability.

Let $\mathbf{\Lambda}_1^{(1)}, \ldots, \mathbf{\Lambda}_1^{(m)}$ be m post burn-in samples of Λ_1. To address the non-identifiability problem, we post-process the chain successively moving from the first sample to the last. First $\mathbf{\Lambda}_1^{(2)}$ is rotated with respect to $\mathbf{\Lambda}_1^{(1)}$ using some orthonormal matrix $\mathbf{R}_1$ such that $\|\mathbf{\Lambda}_1^{(1)} - \mathbf{\Lambda}_1^{(2)}\mathbf{R}_1\|_F^2$ is minimized, where $\|\|_F^2$ denotes the Frobenius norm. This minimization criterion rotates $\mathbf{\Lambda}_1^{(2)}$ to make it as close as possible to $\mathbf{\Lambda}_1^{(1)}$. The solution of $\mathbf{R}_1$ is obtained in Theorem 1. Then we post-process $\mathbf{\Lambda}_1^{(3)}$ with respect to $\mathbf{\Lambda}_1^{(2)}\mathbf{R}_1$ and so on.

Theorem 1 *The minimizer* $\mathbf{R}_1$ *of the objective function* $\|\mathbf{\Lambda}_1^{(1)} - \mathbf{\Lambda}_1^{(2)} R_1\|_F^2$ *is given by* $\mathbf{R}_1 = \mathbf{Q}_2\mathbf{Q}_1^T$*, where* $\mathbf{Q}_1\mathbf{D}\mathbf{Q}_2^T$ *is the singular value decomposition (SVD) of* $(\mathbf{\Lambda}_1^{(1)})^T\mathbf{\Lambda}_1^{(2)}$*.*

The proof of the theorem is in the Section 1.1 of Supplementary Materials. Intuitively, the columns of $\mathbf{Q}_1$ and $\mathbf{Q}_2$ are the canonical correlation components of $\mathbf{\Lambda}_1^{(1)}$ and $\mathbf{\Lambda}_1^{(2)}$, respectively. Thus the rotation matrix $\mathbf{R}_1$ rotates $\mathbf{\Lambda}_1^{(2)}$ towards the least principal angle between $\mathbf{\Lambda}_1^{(2)}$ and $\mathbf{\Lambda}_1^{(1)}$. For instance, $\mathbf{\Lambda}_1^{(2)}$ could be an exact right rotation of $\mathbf{\Lambda}_1^{(1)}$. Thus before starting to post-process the MCMC chain, we transform $\mathbf{\Lambda}_1^{(1)}$ as $\mathbf{\Lambda}_1^{(1)}\mathbf{U}_2$ such that $\mathbf{U}_1\mathbf{E}\mathbf{U}_2^T$ is the SVD of the residual $(\mathbf{x}_t - \mathbf{\Psi}^{(1)}\mathbf{\Gamma}_1^{(1)}\boldsymbol{\zeta}(t)^{(1)})^T\mathbf{\Lambda}_1^{(1)}$ in the same way and here E is the diagonal matrix with elements in decreasing order. This initial transformation ensures that the higher order columns of the loading matrix are lower in significance in explaining the data. Then following the above result, we post-process the rest of the MCMC chain of the loading matrix on the post burn-in samples successively. In general, SVD computation is expensive. However, in most applications, the estimated rank is very small. Thus the computation becomes manageable. After the post-processing, we can construct credible bands for the parameters. We apply this post-processing step for all the loading matrices.

4.2 Measure of similarity

It is of interest to quantify similarity between paired time series. We propose the following measure of similarity,

$$\begin{aligned}&\mathrm{Syn}(\mathbf{X},\mathbf{Y})\\ &= 1-\frac{1}{pT}\sum_l \left|\sum_t \left[\frac{(\mathbf{\Lambda}_l\mathbf{\Xi}_1\boldsymbol{\eta}(t))^2}{(\mathbf{\Psi}_l\mathbf{\Gamma}_1\boldsymbol{\zeta}_1(t))^2+(\mathbf{\Lambda}_l\mathbf{\Xi}_1\boldsymbol{\eta}(t))^2+\sigma_{1l}^2} - \frac{(\mathbf{\Lambda}_l\mathbf{\Xi}_2\boldsymbol{\eta}(M(t)))^2}{(\mathbf{\Psi}_l\mathbf{\Gamma}_2\boldsymbol{\zeta}_2(t))^2+(\mathbf{\Lambda}_l\mathbf{\Xi}_2\boldsymbol{\eta}(M(t)))^2+\sigma_{2l}^2}\right]\right|,\end{aligned}$$

where $\mathbf{\Lambda}_l$, $\mathbf{\Psi}_l$ denote the l^{th} row of the corresponding matrices and p,T denote number of features and time points respectively. The measure 'Syn' is bounded between $[0,1]$. Here, the difference in relative contribution of each feature on the two shared spaces is considered as a measure of dissimilarity. Then as a measure of similarity, we consider the difference of the average dissimilarity from one. Smaller Syn-value would suggest that the warping function is not able to align the shared space perfectly.

5. Theoretical support

In this section, we provide some theoretical justification for our model. Identifiability of the warping function is a desirable property as well as posterior consistency.

5.1 Identifiability of the warping function

The following result shows that the warping function $M(t)$ is identifiable for model (2).

Theorem 2 *The warping function* $M(t)$ *is identifiable if* $\eta(t)$ *is continuous and not constant at any interval of time.*

The proof is by contradiction. Details of the proof are in Section 1.2 of Supplementary Materials. The assumptions on $\eta(t)$ are very similar to those assumed for the 'structural mean' in Gervini and Gasser (2004). The continuity assumption of $\eta(t)$ can be replaced with a 'piecewise monotone without flat parts' assumption (Gervini and Gasser, 2004). The proof is still valid with minor modifications for this alternative assumption. In our model $\eta(t)$ is varying with time smoothly. Thus $M(t)$ is identifiable.

5.2 Asymptotic result

We study the posterior consistency of our proposed model. Our original model is

$$\begin{aligned}\mathbf{x}_t =&\boldsymbol{\Psi}\boldsymbol{\Gamma}_1\boldsymbol{\zeta}_1(t) + \boldsymbol{\Lambda}\boldsymbol{\Xi}_1\boldsymbol{\eta}(t) + \boldsymbol{\epsilon}_{1t}, \quad \boldsymbol{\epsilon}_{1t} \sim \mathrm{N}(0, \boldsymbol{\sigma}_1^2),\\ \mathbf{y}_t =&\boldsymbol{\Psi}\boldsymbol{\Gamma}_2\boldsymbol{\zeta}_2(t) + \boldsymbol{\Lambda}\boldsymbol{\Xi}_2\boldsymbol{\eta}(M(t)) + \boldsymbol{\epsilon}_{2t}, \quad \boldsymbol{\epsilon}_{2t} \sim \mathrm{N}(0, \boldsymbol{\sigma}_2^2).\end{aligned} \tag{19}$$

We first show posterior concentration of a simplified model that drops $\boldsymbol{\Xi}_1$ and $\boldsymbol{\Xi}_2$. Then using that result we show posterior concentration of model (19) in Corollary 4. We rewrite $\boldsymbol{\zeta}_1(t) = \boldsymbol{\Psi}\boldsymbol{\Gamma}_1\boldsymbol{\zeta}_1(t)$, $\boldsymbol{\zeta}_2(t) = \boldsymbol{\Psi}\boldsymbol{\Gamma}_2\boldsymbol{\zeta}_2(t)$ and $\boldsymbol{\eta}(t) = \boldsymbol{\Lambda}\boldsymbol{\eta}(t)$. Based on the constructions, $\boldsymbol{\zeta}_i(t)$ and $\boldsymbol{\eta}(t)$ are orthogonal for $i = 1, 2$. We consider the following simplified model,

$$\begin{aligned}\mathbf{x}_{t_i} =&\boldsymbol{\zeta}_1(t_i) + \boldsymbol{\eta}(t_i) + \boldsymbol{\epsilon}_{1t_i}, \quad \boldsymbol{\epsilon}_{1t} \sim \mathrm{N}(0, \boldsymbol{\sigma}_{1t_i}^2),\\ \mathbf{y}_{t_i} =&\boldsymbol{\zeta}_2(t_i) + \boldsymbol{\eta}(M(t_i)) + \boldsymbol{\epsilon}_2, \quad \boldsymbol{\epsilon}_{2t} \sim \mathrm{N}(0, \boldsymbol{\sigma}_2^2),\end{aligned}$$

for $0 \leq t_i \leq 1$ and $i = 1, \ldots, n$. We study asymptotic properties in the increasing n and fixed p regime. We need to truncate the B-spline series after a certain level or place a shrinkage prior on the number of B-splines as $\Pi[K = k] = b'_1 \exp[-b'_2 k (\log k)^{b'_3}], \Pi[J = j] = b_1 \exp[-b_2 j (\log j)^{b_3}]$, $\Pi[K_i = j] = b_{i1} \exp[-b_{i2} j (\log j)^{b_{i3}}]$ for $i = 1, 2$, with $b_1, b_2, b_{12}, b_{22} b'_1$, $b'_2, b_{11}, b_{21} > 0$ and $0 \leq b_3, b'_3, b_{13}, b_{23} \leq 1$. For $b_3 = 0$ we obtain a geometric distribution and for $b_3 = 1$, a Poisson distribution.

To study posterior contraction rates, we consider the empirical ℓ_2-distance on the regression functions. The empirical ℓ_2-distance for the two sets of parameters $(\boldsymbol{\zeta}_{11}, \boldsymbol{\zeta}_{21}, \boldsymbol{\eta}_1, M_1)$ and $(\boldsymbol{\zeta}_{12}, \boldsymbol{\zeta}_{22}, \boldsymbol{\eta}_2, M_2)$ is given by

$$\begin{aligned}&d^2((\boldsymbol{\zeta}_{11}, \boldsymbol{\zeta}_{21}, \boldsymbol{\eta}_1, M_1), (\boldsymbol{\zeta}_{12}, \boldsymbol{\zeta}_{22}, \boldsymbol{\eta}_2, M_2))\\ &\quad = \frac{1}{n}\sum_{i=1}^{n} \big[\|\boldsymbol{\zeta}_{11}(t_i) - \boldsymbol{\zeta}_{12}(t_i)\|_2^2 + \|\boldsymbol{\zeta}_{21}(t_i) - \boldsymbol{\zeta}_{22}(t_i)\|_2^2 + \|\boldsymbol{\eta}_1(t_i) - \boldsymbol{\eta}_2(t_i)\|_2^2\\ &\qquad + \|\boldsymbol{\eta}_1(M_1(t_i)) - \boldsymbol{\eta}_2(M_2(t_i))\|_2^2\big].\end{aligned}$$

The smoothness of the underlying true functions $\boldsymbol{\zeta}_{10}, \boldsymbol{\zeta}_{20}, \beta_0$ and M_0 plays the most significant role in determining the contraction rate. The fixed dimensional parameters $\boldsymbol{\sigma}_1$ and $\boldsymbol{\sigma}_2$ do not have much impact on the rate. The constants b_{13}, b_{23}, b_3 and b'_3 appearing in the prior for the number of B-spline coefficients K_1, K_2, K, J have a mild effect.

Theorem 3 *Assume that the true functions $\boldsymbol{\zeta}_{10}, \boldsymbol{\zeta}_{20}, \boldsymbol{\eta}_0$ and M_0 belong to Hölder classes of smooth functions and are of regularity levels ι_1, ι_2, ι and ι' on $[0, 1]$. Then the posterior contraction rate is given by*

$$n^{-\bar{\iota}/(2\bar{\iota}+1)} (\log n)^{\bar{\iota}/(2\bar{\iota}+1) + (1-\bar{b}_3)/2},$$

where $\bar{\iota} = \min\{\iota, \iota_1, \iota_2, \iota'\}$ and $\bar{b}_3 = \min\{b_3, b'_3, b_{13}, b_{23}\}$.

The proof is based on the general theory of posterior contraction as in Ghosal and Van der Vaart (2017) for non-identically distributed independent observations and results for finite random series priors (Shen and Ghosal, 2015). Details of the proof are in Section 1.3 of Supplementary Materials.

Let the parameter space for dynamic latent factors $\boldsymbol{\zeta}_1, \boldsymbol{\zeta}_2, \boldsymbol{\eta}$ be $\mathcal{F}$, which is the class of real-valued smooth continuous functions on [0,1], and for the warping function $\mathcal{M}$ be the class of [0, 1] bounded smooth monotone continuous functions on [0,1]. Let $\tilde{X}, \tilde{X}, \tilde{L}, \tilde{G}_1, \tilde{G}_2$ be the priors for the matrices $\boldsymbol{\Xi}_1, \boldsymbol{\Xi}_2, \boldsymbol{\Lambda}, \boldsymbol{\Gamma}_1, \boldsymbol{\Gamma}_2$, respectively, and $\mathcal{X}, \mathcal{L}, \mathcal{G}_1, \mathcal{G}_2$ are the parameter spaces of $\tilde{X}, \tilde{L}, \tilde{G}_1, \tilde{G}_2$, respectively.

Assumption 1: For the true loading matrices and functions, we have
$\{\boldsymbol{\Xi}_{10}, \boldsymbol{\Xi}_{20}, \boldsymbol{\Lambda}_0, \boldsymbol{\Gamma}_{10}, \boldsymbol{\Gamma}_{20}, \boldsymbol{\zeta}_{10}, \boldsymbol{\zeta}_{20}, \beta_0, M_0\} \in \mathcal{X}^2 \times \mathcal{L} \times \mathcal{G}_1 \times \mathcal{G}_2 \times \mathcal{F}^3 \times \mathcal{M}$.

Similarly we can define empirical ℓ_2-distance $d_1^2((\boldsymbol{\Psi}_1, \boldsymbol{\Lambda}_1, \boldsymbol{\Gamma}_{11}, \boldsymbol{\Gamma}_{12}, \boldsymbol{\Xi}_{11}, \boldsymbol{\Xi}_{12}, \boldsymbol{\zeta}_{11}, \boldsymbol{\zeta}_{21}, \boldsymbol{\eta}_1, M_1),$ $(\boldsymbol{\Psi}_2, \boldsymbol{\Lambda}_2, \boldsymbol{\Gamma}_{21}, \boldsymbol{\Gamma}_{22}, \boldsymbol{\Xi}_{21}, \boldsymbol{\Xi}_{22}, \boldsymbol{\zeta}_{12}, \boldsymbol{\zeta}_{22}, \boldsymbol{\eta}_2, M_2))$ as d^2 for the full model and we have following consistency result.

Corollary 4 *Under the above assumption, the posterior for parameters in the model* (19) *is consistent with respect to the distance d_1.*

For the full model in (19), the test constructions will remain the same as in the proof of Theorem 3. We only need to verify the Kullback-Leibler prior positivity condition. Within our modeling framework, Assumption 1 trivially holds. Details of the proof are in Section 1.4 of Supplementary Materials. The posterior contraction rate of this full model will be the same as the given rate of Theorem 3 as the loading matrices can at most be $p \times p$-dimensional and we assume p is fixed.

6. Simulation Study

We run two simulations to evaluate the performance of TACIFA on pairs of multivariate time series. We evaluate TACIFA by: (1) ability to retrieve the appropriate number of shared and individual factors, (2) accuracy of the estimated warping functions and accompanying uncertainty quantification, (3) out of sample prediction errors, and (4) performance relative to two-stage approaches for estimating shared and individual-specific dynamic factors. In the first simulation, we generate data from the proposed model. In the second simulation, we analyze two shapes changing over time, data that does not have any inherent connection to our proposed model. We add two more simulations in Section 4 of Supplementary Materials. One of these two simulations focus on the case where direction of mimicry is changed. The other one corresponds to the case where there is no mimicry.

To assess out of sample prediction error, we randomly assign 90% of the time-points to the training set and the remaining 10% to the test set. Thus, the training set contains a randomly selected 90% of the columns of the data and the remaining 10% columns will be in the test set. The two-stage approaches we compare our method to apply JIVE on the training set in the first stage to estimate the shared space and warp the shared matrices, and then apply multivariate imputation algorithms (`missForest, MICE, mtsdi`) in the second stage to make predictions on the testing data set. We evaluate the performance of naive dynamic time warping (based solely on minimization of Euclidean distance), derivative dynamic time warping (based on local derivatives of the time data to avoid singularity points), and sliding window based dynamic time warping. Since our model is the only approach with a mechanism for uncertainty quantification, we can compare the prediction performance of TACIFA to two-stage approaches, but we cannot compare uncertainty estimation.

The individual-specific loading matrices are $\mathbf{\Psi\Gamma}_1$ and $\mathbf{\Psi\Gamma}_2$. The shared space loading matrices are $\mathbf{\Lambda\Xi}_1$ and $\mathbf{\Lambda\Xi}_2$. For the (i, j)-*th* coordinate of a loading matrix A, we define a summary measure $SP_{i,j}(A) = \big(|0.5 - P(A[i, j] > 0)|\big)/0.5$ quantifying the "importance" of the element. Here $P(A[i, j] > 0)$ is the posterior probability estimated from the MCMC samples of A after performing the post-processing steps defined in Section 4.1. These scores help to quantify the importance of the factors and to estimate the number of important factors retrieved by the model.

6.1 Simulation case 1

We generate data from a factor model with the following specifications: $\zeta_{1k}(t) = \sin(kt)$, $\zeta_{2k}(t) = \cos(kt)$ and $M_0(t) = t^{0.5}$, with k varying from 1 to 10. The shared latent factors $\eta_k(t)$'s are set to k-*th* degree orthogonal polynomials using the R function `poly`. The factor loading matrices are of dimension 15×3, with the elements of Γ_1,Γ_2 generated independently from $\mathrm{N}(0, 0.1^2)$. The entries in the true Λ are structured as a block diagonal matrix as shown in the first image of Figure 4, where the non-zero entries are generated from $\mathrm{N}(15, 0.1^2)$. We vary t from 1/500 to 1 with an increment of 1/500. The data X_t and Y_t are generated from $\mathrm{N}(\mathbf{\Psi}\boldsymbol{\zeta}_1+\mathbf{\Lambda}\boldsymbol{\eta}(t), 1)$ and $\mathrm{N}(\mathbf{\Psi}\boldsymbol{\zeta}_2+\mathbf{\Lambda}\boldsymbol{\eta}(M(t)), 1)$, respectively, where $\beta(t) = (\eta_1(t), \eta_2(t), \eta_3(t))$ and $\mathbf{\Psi} = \mathbf{1} - \mathbf{\Lambda}(\mathbf{\Lambda}^T\mathbf{\Lambda})^{-1}\mathbf{\Lambda}^T$.

The choices of hyper parameters are $\omega = 100$, $\alpha_{i1} = \alpha_{i2} = 5$ for $i = 1, 2$. We set $K_1 = K_2 = J = K$ and fit the model for 4 different choices of $K = 6, 8, 10, 12$. The choice $K = 10$ yields the best results among all the candidates. The hyperparameters of the inverse gamma priors for the variance components are all 0.1 which is weakly informative. We collect 6000 MCMC samples and consider the last 3000 as post burn-in samples for inferences. We start the MCMC chain setting the number of shared latent factors $r = p$ as a very conservative upper bound.

First, we evaluate whether our model retrieves the appropriate number of factors. The true dimension of Λ is 15×3. Figure 3 suggests that TACIFA retrieves 3 important shared space factors, as expected. The individual-specific loading matrices in Figure 3 also suggest approximately three important factors.

Figure 4 illustrates estimated shared loading matrices along with the true loading matrix. The estimated loading matrices roughly match with the true loading structure. The individual specific loadings, however, are not reliably distinguishable as they are constructed as $(\mathbf{I}_p - \mathbf{\Lambda}(\mathbf{\Lambda}^T\mathbf{\Lambda})^{-1}\mathbf{\Lambda}^T)\mathbf{\Gamma}_i$. Thus, we only present our results for the shared loading matrix. Figure 3, however, shows that the ranks of the individual specific loading matrices and the shared loading matrices are all roughly accurate using the proposed importance measures. Next, we evaluate the accuracy of our estimated warping function and accompanying uncertainty quantification. The estimated warping function in Figure 5 is for the training set. The estimate by TACIFA is clearly the best among all methods tested. In Table 1, we compare the prediction MSE results of our method with two-stage methods, and show that

TACIFA has the best performance. Furthermore, Figure 6 illustrates estimated warping curves for a different true warping function $M_0(t) = \{(0.33\sin(2\pi t))^2 + t^2\}^{0.5}$ which incorporates change in direct of mimicry. The TACIFA based estimate is again the best among all the other competing methods.

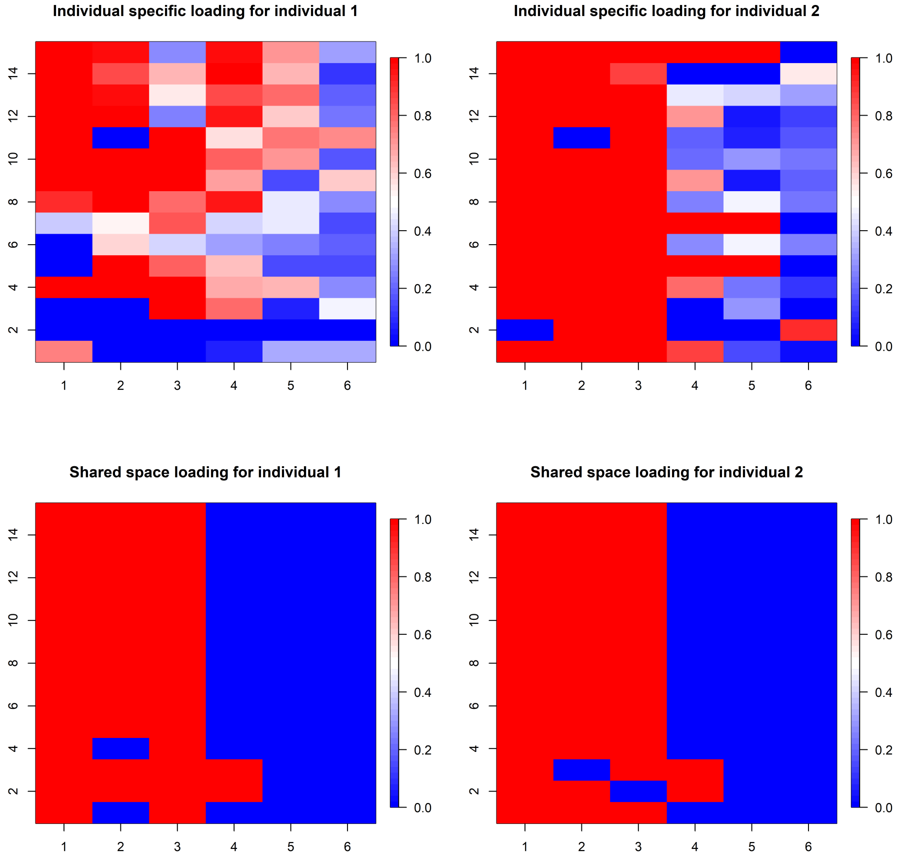


Figure 3: Estimated importance measures SP for loading matrices of shared and individual spaces of Series 1 and 2 in Simulation Case 1. Each column represents each factor. The columns with higher proportion of red correspond to the factors with higher importance.

Finally we measure the similarity of the simulated data using the measure described in Section 4.2. If $\zeta_{1k}(t) = \sin(kt)$ as above, the similarity is 0.95. To confirm that this measure is sensitive to the similarity between two time series, as intended, we change the first multivariate time series relative to the other multivariate time series by changing the first individual specific latent factors $\zeta_{1k}(t)$ systematically, and recalculating the similarity. When $\zeta_{1k}(t) = kt$, similarity drops from 0.95 to 0.89. When $\zeta_{1k}(t) = (kt)^2$, similarity further reduces to 0.79. The warping function estimated for each of these pairs of time

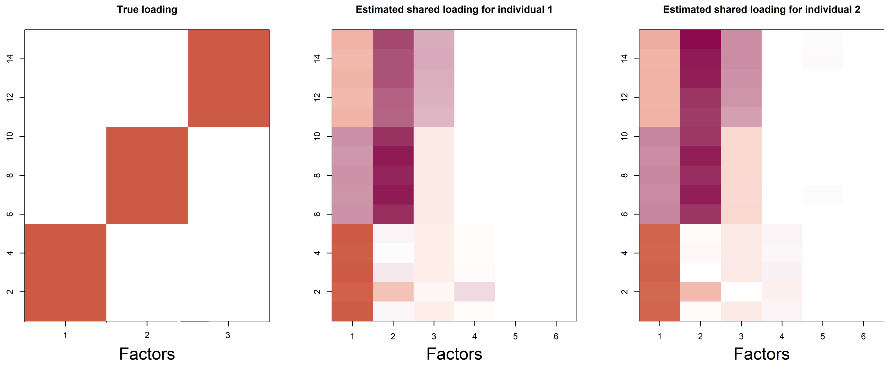


Figure 4: Estimated shared loading matrices along with the true loading structure in Simulation Case 1.

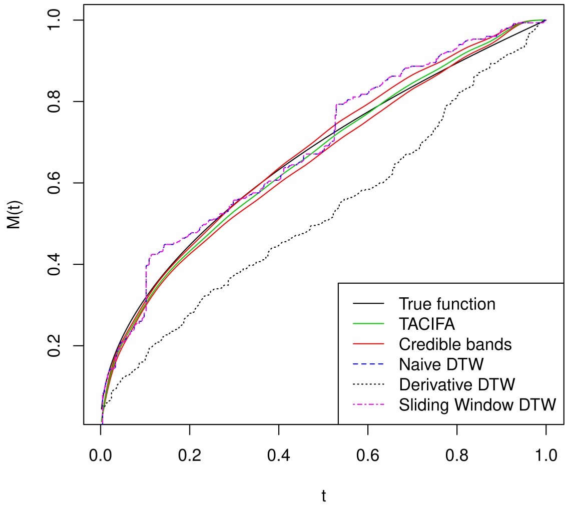


Figure 5: Estimated warping function for simulated data in Simulation Case 1. The black curve is the true warping function $M_0(t) = t^{0.5}$, the green curve is the estimated function, 95% credible bands are shown in red. Naive DTW and Sliding window DTW curves are indistinguishable. Of all the methods tested, the TACIFA estimated warping function is closest to the true warping function.

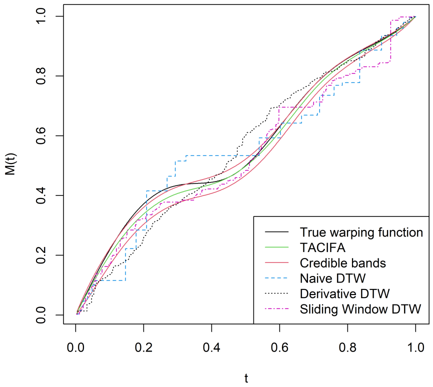


Figure 6: Estimated warping function for simulated data in a setting similar to Simulation Case 1, but with different true warping function. The black curve is the true warping function $M_0(t) = \{(0.33 \sin(2\pi t))^2 + t^2\}^{0.5}$, the green curve is the estimated function, 95% credible bands are shown in red. Naive DTW and Sliding window DTW curves are indistinguishable. Of all the methods tested, the TACIFA estimated warping function is closest to the true warping function.

Table 1: Prediction MSEs of the first and second time series in Simulation 1. using two-stage methods. The top row indicates the R package used to impute, and the first column indicates the warping method. The two-stage prediction MSEs are all greater than the TACIFA prediction MSEs (1.01, 1.02).

	`missForest`	`MICE`	`mtsdi`
Naive DTW	(6.12, 9.66)	(8.65,9.70)	(1.03,1.03)
Derivative DTW	(6.37, 9.49)	(8.06,9.80)	(1.03,1.03)
Sliding DTW	(7.15, 10.55)	(9.61,10.39)	(1.03,1.03)

series deteriorates as the two multivariate time series become more distinct as expected. Two stage methods do much worse in these cases (Figure 5 of the Supplementary Materials).

6.2 Simulation case 2

In Simulation Case 2, each series reflects a circle changing into an ellipse over time, similar to a mouth gaping and subsequently closing. The area of the shape is kept fixed by modifying the major and minor axis appropriately. The area of an ellipse, with a and b as the lengths of the major and minor axes, is given by πab. Thus to have the area remain fixed we need ab=constant. We maintain the constant to be 2. With the same true warping function $M_0(t)$ as in the previous simulation, the values for major and minor axes are linked over time across the two individuals. We let $ax(t) = 2(t+1)$ where t's are 500 equidistant values between 1/500 and 1 and $bx(t) = 2/(t+1)$; here $ax(t)$ and $bx(t)$ are major and minor axes of the ellipse at time t corresponding to X_t. At $t = 0$, it is a circle. For the second series we then have $ay(t) = 2(t^{0.5}+1)$ and $by(t) = 2/(t^{0.5}+1)$. We consider the pair of Cartesian coordinates of 12 equidistant points across the perimeter of the ellipse as features (yielding 24 features in total). The features correspond to 12 equidistant angles in $[0, 2\pi)$. Let $\theta_1, \ldots, \theta_{12}$ be those angles. Then $X_{it} = (ax(t)\sin(\theta_i), bx(t)\cos(\theta_i))$ and $Y_{it} = (ay(t)\sin(\theta_i), by(t)\cos(\theta_i))$.

The choices of hyperparameters and the number of MCMC iterations are all the same as in the previous simulation case. We again set $K_1 = K_2 = J = K$ and fit the model for 4 different choices as before. The best choice based on the out of sample prediction for this case is $K = 8$. We have a pair of 24 dimensional time series. The X or Y coordinate is zero for the following four features $\theta_i = 0, \pi$ and $\theta_i = \pi/2, 3\pi/2$. Thus, the warping should not have any effect on these features and should not contribute to the individual-specific space. The remaining 20 features represent 10 features and their mirror images with respect to either the major or minor axis. Thus, we might predict that the shared space should have 10 independent factors, which is consistent with the results displayed in Figure 6 of the Supplementary Materials. As there are 12 features, the individual-specific space should ideally have around two important factors. This is the case for one of the two individual-specific plots in Figure 6 of the Supplementary Materials. For the other individual, there is one more moderately important factor if we set a threshold of 0.9 on the importance measure SP. Figure 8 compares the estimates of the warping function when signal-to-noise ratio is low. Although our estimates perform much better than the rest, the width of credible bands expands with increasing error variance. Since the magnitudes of the features are very small, even noise with variance 1 or 1.5^2 is large.

We plot the estimated warping functions in Figure 7, and plot the estimated shapes in Figure 9. Figure 7 illustrates that the TACIFA-estimated warping function is once again the most accurate of the tested approaches. The TACIFA-estimated warping function is almost identical to the true curve, and has tightly concentrated credible bands. Figure 9 confirms that the TACIFA-estimated Cartesian coordinates of the 12 equidistant features are almost perfectly aligned with the true Cartesian coordinates. Quantifying these accuracies, we calculate the prediction TACIFA MSEs, which are 1.34×10^{-6} and 4.99×10^{-6} with 95% and 96% frequentist coverage within 95% posterior predictive credible bands for X and Y coordinates, respectively. In Table 2, we compare the results of our method with two-stage methods, and show that TACIFA again has the best performance, this time much more dramatically than in the first simulation. The method mtsdi gives similar prediction error to our method in the first simulation setup but fails to impute at any of the missing time points for the second simulation. MICE could impute in the first simulation, but only partially for the second simulation. Only missForest could produce results for both of the two simulations. Nonetheless, its prediction MSEs are much higher than those of our method.

Table 2: Prediction MSEs of the first and second time series in Simulation 2 using two-stage methods. The top row indicates the R package used to impute, and the first column indicates the method used to warp. mtsdi could not impute at any of the testing time points in this simulation. The two-stage prediction MSEs are all greater than the TACIFA prediction MSEs (1.34×10^{-6}, 4.99×10^{-6}).

	`missForest`	`MICE`	`mtsdi`
Naive DTW	(0.12,0.07)	(0.18,0.09)	(-,-)
Derivative DTW	(0.12,0.07)	(0.15,0.07)	(-,-)
Sliding DTW	(0.12,0.07)	(0.14,0.05)	(-,-)

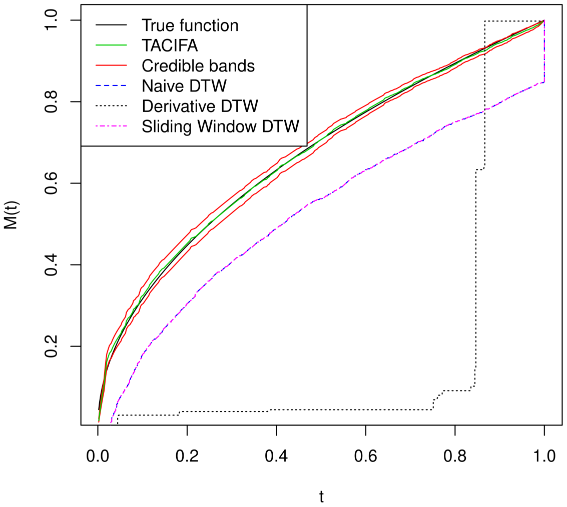
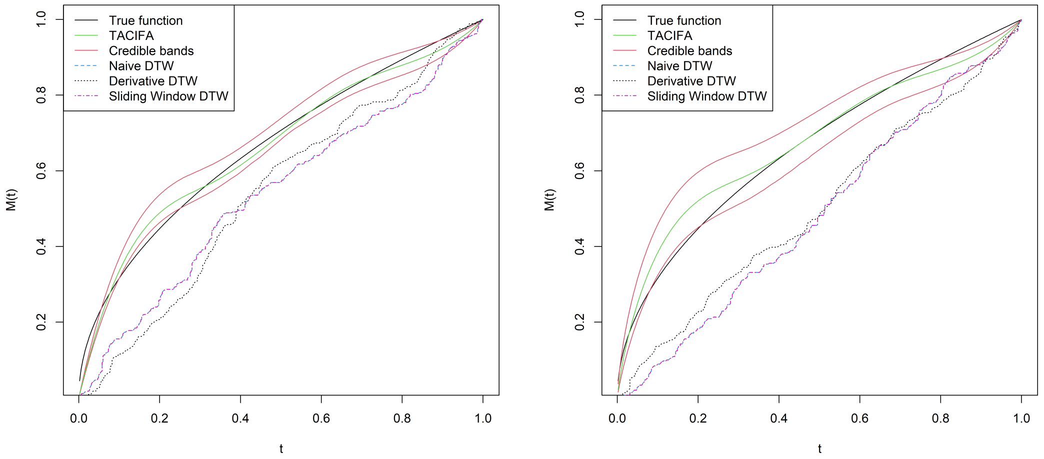


Figure 7: Estimated warping functions for Simulation case 2. The black curve is the true warping function $M_0(t) = t^{0.5}$. The green curve is the TACIFA estimated function, with the 95% credible bands shown in red. Naive DTW and Sliding window DTW curves are indistinguishable. Of all the methods tested, the TACIFA estimated warping function is closest to the true warping function.



(a) The noise follows N(0,1^2)

(b) The noise follows N(0,1.5^2)

Figure 8: Estimated warping functions for Simulation case 2 with more noise added to the data.

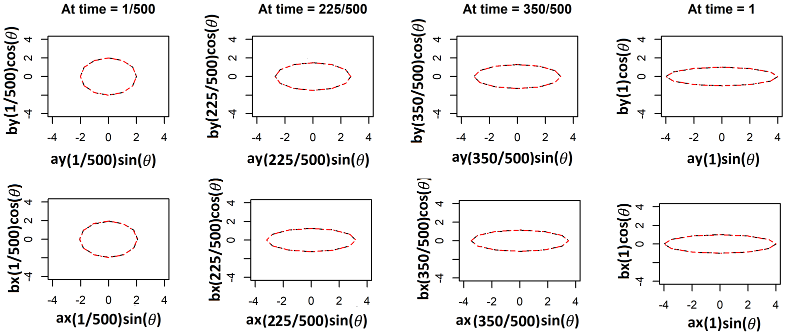


Figure 9: Results for simulation case 2. The first row corresponds to the co-ordinates $(ax(t)\sin(\theta), bx(t)\cos(\theta))$ for four choices of t, evaluated on a grid of θ. Likewise, the second row shows the co-ordinates of $(ay(t)\sin(\theta), by(t)\cos(\theta))$'s for the same choices of t and the θ-grid. Here $ax(t) = 2(t+1), bx(t) = 2/(t+1)$ and $ay(t) = 2(t^{0.5}+1), by(t) = 2/(t^{0.5}+1)$. The black dashed lines represent true curves at four time points and the red dashed lines are the estimated curves. The fit is excellent so that they almost lie on top of each other. At $t = 1$, X and Y both have the same shape.

7. Human Mimicry Application

We apply TACIFA to data from a simple social interaction in which one participant was instructed to imitate the head movements of another. The interaction occurred over Skype, and the videos of both participants were recorded. OpenFace software (Baltrusaitis et al., 2018) was used to extract regression scores for the X and Y coordinates of facial features around the mouth, as well as the pitch, yaw, and roll of head positions, from each frame of each video. These facial features are extracted and normalized before comparing the corresponding time series. Here, we analyze a session where one individual was instructed to imitate the other participant's head movement throughout the interaction. We also apply our method to two related sessions where the role of imitator/imitate changes during the session, with results in Section 3 of Supplementary Materials. Although these social interactions were intentionally constrained to help assess the current methodology under consideration, they represent the types of dynamic social interactions that are of interest to psychologists, autism clinicians, and social robotics developers.

The duration of the experiment is rescaled into [0, 1]. The choices of hyperparameters for estimation are kept the same as in the two simulation setups above except for the number of B-splines. We again run a similar cross validation procedure, and set the number of bases at 8. We collect 5000 MCMC samples after 5000 burn-in samples. We truncate the columns of the loading matrices that have mean absolute contribution less than 0.0001. We plot the estimated warping function along with credible bands and the values of $SP(\boldsymbol{\Psi}\boldsymbol{\Gamma}_1), SP(\boldsymbol{\Psi}\boldsymbol{\Gamma}_2), SP(\boldsymbol{\Lambda}\boldsymbol{\Xi}_1)$, and $SP(\boldsymbol{\Lambda}\boldsymbol{\Xi}_2)$ as in the simulation analyses. Recall that $SP_{i,j}(A) = \big(|0.5 - P(A[i,j] > 0)|\big)/0.5$ where $P(A[i,j] > 0)$ stands for the posterior probability estimated from the MCMC samples of A after performing the post-processing steps defined in Section 4.1.

We apply TACIFA to the time courses of 20 facial features from around the mouth and chin, along with three predictors of head position. We begin by evaluating the loading matrices of the shared and individual factors. There should be a large shared space in this experiment, as we know one person was imitating the head movements of the other, and all of the features examined were related to the head. We plot $SP(\boldsymbol{\Psi}\boldsymbol{\Gamma}_1), SP(\boldsymbol{\Psi}\boldsymbol{\Gamma}_2), SP(\boldsymbol{\Lambda})$, and $SP(\boldsymbol{\Lambda}\boldsymbol{\Xi}_2)$ in Figure 10. Half of the 20 facial features examined in this experiment were roughly the mirror image of the others, due to facial symmetry. As a consequence, we might predict that the shared space should not have more than 13 factors. Consistent with this hypothesis, there are 13 important shared features in Figure 7. In addition, all of the features examined in this experiment are related to head movement, so we might predict very little individual variation in the time courses. This prediction is consistent with the low importance of all the individual-specific factors shown in Figure 10.

Next, we examine the TACIFA estimated warping function and accompanying uncertainty quantification. Figure 11 shows that the estimated warping function is below the $M(t) = t$ line throughout the experiment. This indicates that the TACIFA approach correctly estimated that one individual was following the other individual in time through the experiment. Derivative DTW was the only other method that achieved that. Furthermore, all these methods also suggest that the participants switched leadership roles multiple times, which is not true.

Next, we compare the TACIFA out of sample prediction MSEs to those of two-stage approaches, and compute the similarity. The TACIFA MSEs are 4.25 and 2.21, with 95% and 98% frequentist coverage within 95% posterior predictive credible bands, relative to the estimated variances 4.34 and 2.61 for the first and second individuals, respectively. These MSEs are lower than those of the two stage approaches, which are around 9. A detailed table is in the Supplementary Materials.

Finally, we assess the similarity of the two time series and test whether greater numbers of features influence the similarity measure. Let $\mathbf{X}_m$ and $\mathbf{Y}_m$ denote the paired time series with m set of features (maximum of 10) around the chin along with the three predictors on head position. We have a total of 10 possible features in this analysis. We get $\text{Syn}(\mathbf{X}_3, \mathbf{Y}_3)$=0.80, $\text{Syn}(\mathbf{X}_6, \mathbf{Y}_6)$=0.85 and $\text{Syn}(\mathbf{X}_{10}, \mathbf{Y}_{10})$=0.85. These high values are reasonable, since all the features examined will be influenced by head movement and head movements were intentionally coordinated. The results also indicate that similarity values increase as the number of relevant features increases.

8. Discussion

There are many possibilities of future research building on TACIFA. It is natural to generalize to D many matrices, which would require D different individual-specific loadings $\mathbf{\Gamma}_1, \ldots, \mathbf{\Gamma}_D$ along with $D-1$ different warping functions. In addition, in settings such as our motivating social mimicry application, there may be data available from n pairs of interacting individuals. In such a case, it is natural to develop a hierarchical extension of the proposed approach that can borrow information across individuals and make inferences about population parameters. Another direction is to build static Bayesian models to estimate the joint and individual structures under the orthogonality assumption by dropping the warping function from our proposed model to accounting for group differences. The current implementation for updating $\mathbf{\Lambda}$ prohibits its use for large p as the computational complexity in updating a $p \times r$ dimensional $\mathbf{\Lambda}$ at each iteration is of order rp^2. Thus, developing computationally efficient posterior computation algorithms is another direction to ensure broader applicability of our proposed method. Future work will also consider the cases where the data matrices $\mathbf{X}$ and $\mathbf{Y}$ have an unequal number of time points. Although theoretically our proposed model can accommodate this case, the computational complexity may be high.

A further important and challenging direction is to generalize the proposed methods to allow for more complex types of interactions. Two individuals who are interacting may not simply imitate each other, but have more nuanced and diverse types of coordination. For example, one individual may nod their head or laugh in response to the funny facial expressions another individual intentionally makes, or one individual may close their eyes when the other individual sticks out their tongue. Accommodating such complexity will require a more complex dynamic latent structure than that described here.

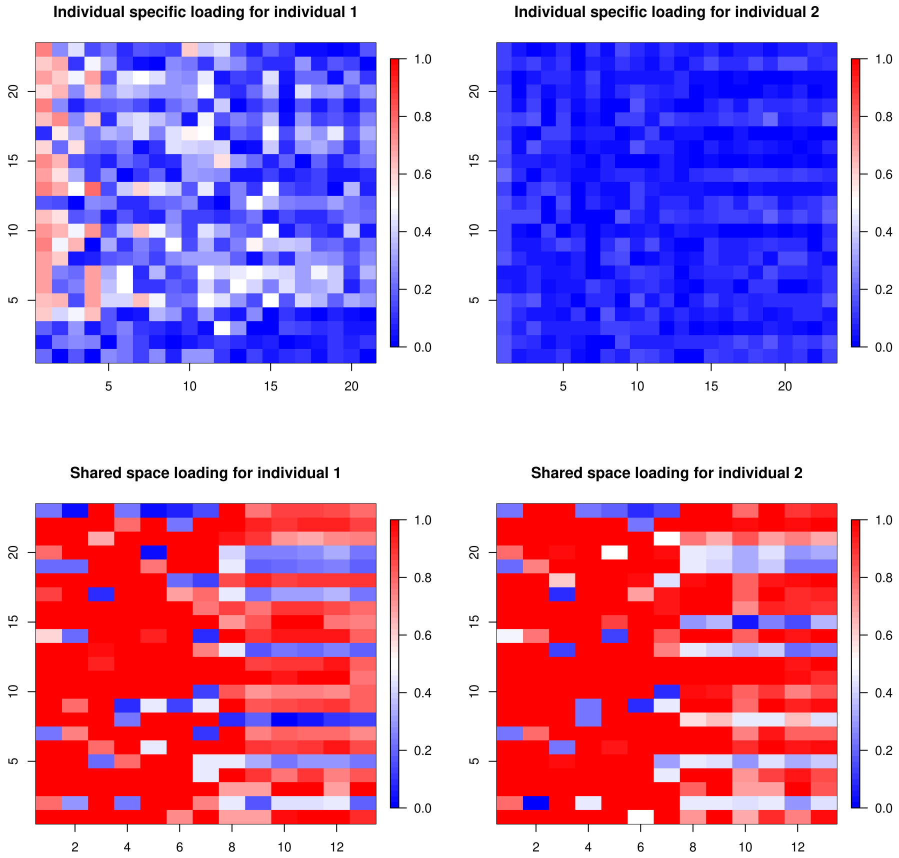


Figure 10: Plot of the summary measure as evidence of importance of the entries of loading matrices in human mimicry dataset (A). Each column represents one factor. The columns with higher proportion of red correspond to the factors with higher importance.

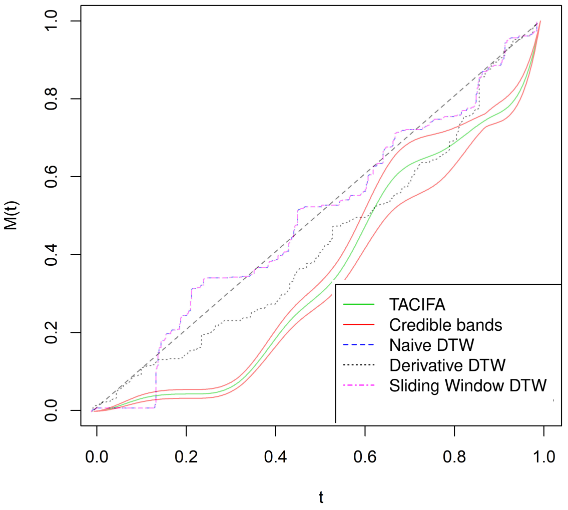


Figure 11: Estimated warping function in human mimicry dataset (A). The green curve is the estimated function along with the 95% pointwise credible bands in red. The estimated curve is always below the dashed line, indicating the second person is mimicked throughout the experiment

References

Omar Aguilar and Mike West. Bayesian dynamic factor models and portfolio allocation. *Journal of Business & Economic Statistics*, 18:338–357, 2000.

Christian Aßmann, Jens Boysen-Hogrefe, and Markus Pape. Bayesian analysis of static and dynamic factor models: An ex-post approach towards the rotation problem. *Journal of Econometrics*, 192:190–206, 2016.

Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. Openface 2.0: Facial behavior analysis toolkit. In *Automatic Face & Gesture Recognition (FG 2018), 2018 13th IEEE International Conference on*, pages 59–66. IEEE, 2018.

Karthik Bharath and Sebastian Kurtek. Partition-based sampling of warp maps for curve alignment. *arXiv preprint arXiv:1708.04891*, 2017.

Anirban Bhattacharya and David B Dunson. Sparse Bayesian infinite factor models. *Biometrika*, 98:291–306, 2011.

Zhengping Che, Xinran He, Ke Xu, and Yan Liu. DECADE: a deep metric learning model for multivariate time series. In *KDD workshop on mining and learning from time series*, 2017.

Wen Cheng, Ian L Dryden, Xianzheng Huang, et al. Bayesian registration of functions and curves. *Bayesian Analysis*, 11:447–475, 2016.

Gerda Claeskens, Bernard W Silverman, and Leen Slaets. A multiresolution approach to time warping achieved by a Bayesian prior–posterior transfer fitting strategy. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72:673–694, 2010.

Carl De Boor. *A practical guide to splines*, volume 27. Springer-Verlag New York, 1978.

Roberta De Vito, Ruggero Bellio, Lorenzo Trippa, and Giovanni Parmigiani. Bayesian multi-study factor analysis for high-throughput biological data. *Annals of Applied Statistics (Future Papers)*, 2021.

Simon Duane, Anthony D Kennedy, Brian J Pendleton, and Duncan Roweth. Hybrid Monte Carlo. *Physics Letters B*, 195:216–222, 1987.

Qing Feng, Meilei Jiang, Jan Hannig, and JS Marron. Angle-based joint and individual variation explained. *Journal of Multivariate Analysis*, 166:241–265, 2018.

Sylvia Fruehwirth-Schnatter and Hedibert Freitas Lopes. Sparse Bayesian factor analysis when the number of factors is unknown. *arXiv preprint arXiv:1804.04231*, 2018.

Daniel Gervini and Theo Gasser. Self-modelling warping functions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66:959–971, 2004.

Subhashis Ghosal and Aad Van der Vaart. *Fundamentals of nonparametric Bayesian inference*, volume 44. Cambridge University Press, 2017.

Xuming He and Peide Shi. Monotone B-spline smoothing. *Journal of the American Statistical Association*, 93:643–650, 1998.

Shashank Jere, Justin Dauwels, Muhammad Tayyab Asif, Nikola Mitro Vie, Andrzej Cichocki, and Patrick Jaillet. Extracting commuting patterns in railway networks through matrix decompositions. In *Control Automation Robotics & Vision (ICARCV), 2014 13th International Conference on*, pages 541–546. IEEE, 2014.

Lucy Johnston. Behavioral mimicry and stigmatization. *Social Cognition*, 20:18–35, 2002.

Sebastian Kurtek. A geometric approach to pairwise Bayesian alignment of functional data using importance sampling. *Electronic Journal of Statistics*, 11:502–531, 2017.

Jessica L Lakin and Tanya L Chartrand. Using nonconscious behavioral mimicry to create affiliation and rapport. *Psychological Science*, 14:334–339, 2003.

Gen Li and Irina Gaynanova. A general framework for association analysis of heterogeneous data. *The Annals of Applied Statistics*, 12:1700–1726, 2018.

Lizhen Lin and David B Dunson. Bayesian monotone regression using Gaussian process projection. *Biometrika*, 101:303–317, 2014.

Jennifer Listgarten, Radford M Neal, Sam T Roweis, and Andrew Emili. Multiple alignment of continuous time series. In *Advances in Neural Information Processing Systems*, pages 817–824, 2005.

Eric F Lock and David B Dunson. Bayesian consensus clustering. *Bioinformatics*, 29: 2610–2616, 2013.

Eric F Lock, Katherine A Hoadley, James Stephen Marron, and Andrew B Nobel. Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *The Annals of Applied Statistics*, 7:523, 2013.

Hedibert Freitas Lopes and Mike West. Bayesian model assessment in factor analysis. *Statistica Sinica*, 14:41–67, 2004.

Yi Lu, Radu Herbei, and Sebastian Kurtek. Bayesian registration of functions with a Gaussian process prior. *Journal of Computational and Graphical Statistics*, 26:894–904, 2017.

Lauren E Marsh, Geoffrey Bird, and Caroline Catmur. The imitation game: Effects of social cues on 'imitation'are domain-general in nature. *NeuroImage*, 139:368–375, 2016.

Radford M Neal et al. Mcmc using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2:2, 2011.

Brian Neelon and David B Dunson. Bayesian isotonic regression and trend analysis. *Biometrics*, 60:398–406, 2004.

Carlotta Orsenigo and Carlo Vercellis. Combining discrete SVM and fixed cardinality warping distances for multivariate time series classification. *Pattern Recognition*, 43:3787–3794, 2010.

James O Ramsay et al. Monotone regression splines in action. *Statistical Science*, 3:425–441, 1988.

Priyadip Ray, Lingling Zheng, Joseph Lucas, and Lawrence Carin. Bayesian joint analysis of heterogeneous genomics data. *Bioinformatics*, 30:1370–1376, 2014.

Veronika Ročková and Edward I George. Fast Bayesian factor analysis via automatic rotations to sparsity. *Journal of the American Statistical Association*, 111:1608–1622, 2016.

Martijn Schouteden, Katrijn Van Deun, Tom F Wilderjans, and Iven Van Mechelen. Performing DISCO-SCA to search for distinctive and common information in linked data. *Behavior Research Methods*, 46:576–587, 2014.

George AF Seber. *Multivariate observations*, volume 252. John Wiley & Sons, 2009.

Weining Shen and Subhashis Ghosal. Adaptive Bayesian procedures using random series priors. *Scandinavian Journal of Statistics*, 42:1194–1213, 2015.

Thomas S Shively, Thomas W Sager, and Stephen G Walker. A Bayesian approach to non-parametric monotone function estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71:159–175, 2009.

Donatello Telesca and Lurdes Y T Inoue. Bayesian hierarchical curve registration. *Journal of the American Statistical Association*, 103:328–339, 2008.

George Trigeorgis, Mihalis A Nicolaou, Björn W Schuller, and Stefanos Zafeiriou. Deep canonical time warping for simultaneous alignment and representation learning of sequences. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40:1128–1138, 2017.

Tsung-Heng Tsai, Mahlet G Tadesse, Yue Wang, and Habtom W Ressom. Profile-based lc-ms data alignment-a Bayesian approach. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 10:494–503, 2013.

Jérôme Vial, Hicham Noçairi, Patrick Sassiat, Sreedhar Mallipatu, Guillaume Cognon, Didier Thiébaut, Béatrice Teillet, and Douglas N Rutledge. Combination of dynamic time warping and multivariate analysis for the comparison of comprehensive two-dimensional gas chromatograms: application to plant extracts. *Journal of Chromatography A*, 1216: 2866–2872, 2009.

Guoxu Zhou, Andrzej Cichocki, Yu Zhang, and Danilo P Mandic. Group component analysis for multiblock data: Common and individual feature extraction. *IEEE Transactions on Neural Networks and Learning Systems*, 27:2426–2439, 2016.