Sanna Passino and Heard's contribution to the Discussion of "Statistical exploration of the Manifold Hypothesis" by Whiteley, Gray and Rubin-Delanchy

Francesco Sanna Passino, Nicholas A. Heard

Department of Mathematics, Imperial College London
f.sannapassino@imperial.ac.uk; n.heard@imperial.ac.uk

We congratulate the authors on their excellent contribution and the thoughtful perspectives they brought to this topic. In this contribution, we focus on the assumption of independence between the latent variables $Z_1, \dots, Z_n$ in the Latent Metric Space (LMS) model in Section 2. Data often arrive sequentially, leading to intrinsic row-wise dependencies between data matrices observed at $T > 1$ time points, $\mathbf{Y}_t = (Y_{i,j,t})_{1 \leq i \leq n,\, 1 \leq j \leq p} \in \mathbb{R}^{n \times p}, t = 1, \dots, T$. This scenario is commonly encountered with longitudinal data, or dynamic networks.

A temporal LMS model. We could define latent trajectories $\mathbf{Z}_i = (Z_{i,1}, \dots, Z_{i,T}) \in \mathcal{Z}^T$, sampled independently for each individual $i = 1, \dots, n$. Observed data matrices $\mathbf{Y}_t \in \mathbb{R}^{n \times p}$ would then be generated as

$$Y_{i,j,t} = X_j(Z_{i,t}) + \sigma E_{i,j,t}, \qquad i = 1, \dots, n, \quad j = 1, \dots, p, \quad t = 1, \dots, T, \tag{1}$$

imposing the same assumptions as Section 2. Temporal dependence within each $\mathbf{Z}_i$ induces dependence between corresponding rows of the matrices $\mathbf{Y}_t$, while the latent geometry is preserved through time by the shared metric space $\mathcal{Z}$.

Joint dimension reduction. Dimension reduction of the matrices $\mathbf{Y}_1, \dots, \mathbf{Y}_T$ can proceed using the unfolded data matrix $\tilde{\mathbf{Y}} = [\mathbf{Y}_1 \mid \dots \mid \mathbf{Y}_T]^\top \in \mathbb{R}^{nT \times p}$. Let $s \leq \min\{p, nT\}$, and let the columns of $\tilde{\mathbf{V}}_{\tilde{\mathbf{Y}}} \in \mathbb{R}^{p \times s}$ be the orthonormal eigenvectors associated with the s largest eigenvalues of $\tilde{\mathbf{Y}}^\top \tilde{\mathbf{Y}} \in \mathbb{R}^{p \times p}$. In the terminology of the paper, the *dimension-s dynamic PCA embedding* is:

$$\left[\tilde{\boldsymbol{\zeta}}_{1,1} \,\middle|\, \dots \,\middle|\, \tilde{\boldsymbol{\zeta}}_{n,T}\right]^\top = \tilde{\mathbf{Y}}\, \tilde{\mathbf{V}}_{\tilde{\mathbf{Y}}} \in \mathbb{R}^{nT \times s}. \tag{2}$$

These quantities can be interpreted as *stable time-indexed principal component scores* summarising the evolution of each latent trajectory in a shared low-dimensional space.

Illustrative example: latent trajectories on a torus. Consider the setting of Section 3.5, in which the latent space $\mathcal{Z}$ corresponds to a torus embedded in $\mathbb{R}^3$, parameterised by two radii $\rho_1 > \rho_2 > 0$. A point $z \in \mathcal{Z}$ can be expressed via two angles (θ_1, θ_2) and a map $h : \mathbb{R}^2 \to \mathbb{R}^3$:

$$z = h(\theta_1, \theta_2) = \begin{bmatrix} (\rho_1 + \rho_2 \cos\theta_2)\cos\theta_1 & (\rho_1 + \rho_2 \cos\theta_2)\sin\theta_1 & \rho_2 \sin\theta_2 \end{bmatrix}^\top.$$

Each individual follows a latent trajectory on the torus, described by $Z_{i,t} = h(\theta_{1,i,t}, \theta_{2,i,t})$. Here we consider paths obtained via *fractional Brownian motion (fBm)* processes with Hurst parameter $H_i \in (0, 1)$. For each angle index $k = 1, 2$ and individual $i = 1, \ldots, n$, the fBm model with normally distributed initial angles assumes

$$(\theta_{k,i,1}, \ldots, \theta_{k,i,T}) \sim \mathbb{N}(\mathbf{0},\, \sigma_0^2\, \mathbf{1}\mathbf{1}^\top + \mathbf{K}_i),$$

for $\sigma_0 > 0$, where $\mathbf{K}_i$ is the $T \times T$ fBm covariance matrix with entries

$$K_{s,t,i} = (s^{2H_i} + t^{2H_i} + |t - s|^{2H_i})/2.$$

Following the data-generating structure from Section 3.5 for X_j, $j = 1, \ldots, p$, and $\mathbf{E}_t$, we sample observations from model (1) with $T = 30$, $p = 50$, $\sigma = 0.01$, for $n = 1000$ individuals evolving as an fBm process on the torus ($\rho_1 = 0.75, \rho_2 = 0.15$) with $H_i = 0.9$. The latent trajectories are then estimated by the principal component scores (2). Results for the six highest time-indexed principal component scores, for six randomly selected trajectories, are given in Figures 1b–1c, along with their true counterparts on the torus in Figure 1a.

(a) Six sampled $\mathbf{Z}_i$ paths. (b) Dimensions 1-3 of $\tilde{\boldsymbol{\zeta}}_{i,t}$. (c) Dimensions 4-6 of $\tilde{\boldsymbol{\zeta}}_{i,t}$.

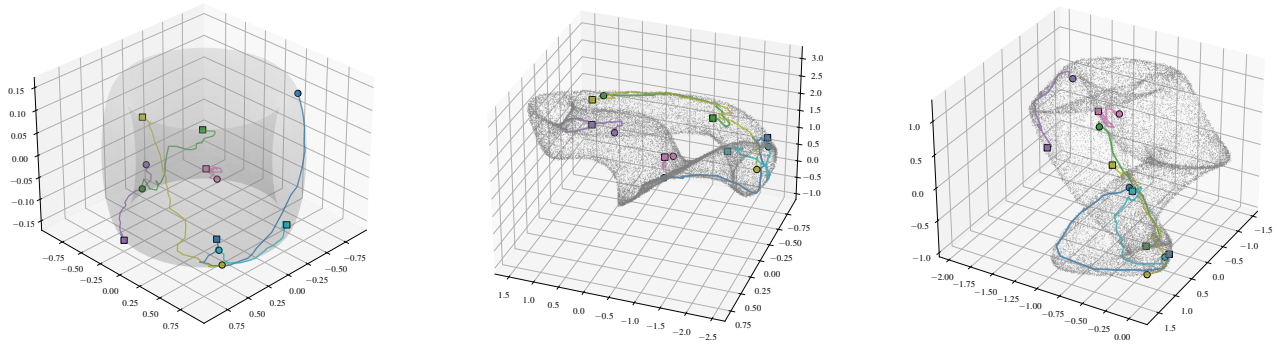

Figure 1: Latent fBm trajectories on a torus and resulting principal component scores.

As in Section 3.5, whilst the global shapes of the true latent positions and the estimates differ, the inter-point distances seem relatively well-preserved. In the temporal case, we also observe realistic within-individual distances, with relative path lengths approximately preserved.

Discussion. In conclusion, we invite the authors to further comment on where the requirement for $Z_1, \ldots, Z_n$ to be independent is exploited. Clarifying this point could shed further light on the proposed methodology, and provide additional intuition on how the results might extend to settings where the latent variables are dependent. Such an extension could, in our opinion, considerably enhance the applicability of the model in temporal contexts, as we attempted to demonstrate in this brief example.