Supplementary material: Discovering Synchronized Subset of Sequences: A Large Scale Solution

1. Introduction

In this supplementary document we show that the correlation between two sequences being at least ρ_θ is equivalent to the ℓ_2 norm of their difference (after z-normalization) being at most $\sqrt{2T(1-\rho_\theta)}$ (Appendix A), provide proofs for Theorem 2.1 (Appendix B) and Lemma 2.2 of the main text (Appendix C), and show that the proposition that two clusters C^i and C^j are ϵ-neighbors if Eq. (5) of the main text holds (Appendix D).

Appendix A

Here we show that if the correlation coefficient between two T-long sequences x and y is higher than ρ_θ, *i.e.* $r(x,y) > \rho_\theta$, then $||\tilde{x}-\tilde{y}|| \leq \epsilon_\theta = \sqrt{2T(1-\rho_\theta)}$, where $\tilde{x}$ and $\tilde{y}$ are the z-normalized versions of x and y, respectively. The correlation coefficient between sequences x and y of length T can be computed using the corresponding z-normalized sequences $\tilde{x}$ and $\tilde{y}$ as

$$r\,(x,y) = \frac{1}{T}\sum_{t=1}^{T} \tilde{x}_t \tilde{y}_t. \tag{S.1}$$

The square of ℓ_2 norm of $\tilde{x}-\tilde{y}$ is

$$||\tilde{x}-\tilde{y}||_2^2 = \sum_{t=1}^{T} \tilde{x}_t^2 - 2\tilde{x}_t\tilde{y}_t + \tilde{y}_t^2 = 2T - 2Tr(x,y), \tag{S.2}$$

where the last equality follows from the fact that the $\sum_t \tilde{x}_t^2 = \sum_t \tilde{y}_t^2 = T$ as z-normalized sequences have standard deviation of 1. Eq. (S.2) implies that

$$||\tilde{x}-\tilde{y}||_2 = \sqrt{2T(1-r(x,y))} \leq \sqrt{2T(1-\rho_\theta)} = \epsilon_\theta.$$

Appendix B

Here we give the proof of Theorem 1 of the main text, which we copy below for convenience.

Theorem 1.1. *Let $\mathcal{S}=\{x_i\}_{i\in\mathcal{I}}$, where $\mathcal{I} \subseteq \{i\}_{i=1}^N$, be a set of T-long sequences that satisfy condition (3) and $\mathcal{U}$ the set that contains the K-dimensional compressed ($K < \min\{T,N\}$) PCA representations of those sequences, $\mathcal{U} = \{\mathbf{u}^i\}_{i\in\mathcal{I}}$. Let $\epsilon := \epsilon_\theta\sqrt{K/(2(K+1))}$, and $\{C^j\}_j$ be a set of clusters (Definition 2) such that $\bigcup_j C^j = \mathbb{R}^K$. Then, there exists an ϵ-expanded cluster C^j_ϵ such that $\mathcal{U} \subseteq C^j_\epsilon$. Moreover, there is no $C^j_{\epsilon_0}$ with $\epsilon_0 < \epsilon$ that can in general guarantee the existence of C^j such that $\mathcal{U} \subseteq C^j_{\epsilon_0}$.*

Proof. Since $\mathcal{U}$ is a set with a finite number of points in $\mathbb{R}^K$, there exist an infinite number of Euclidean balls that contain all points in $\mathcal{U}$. Let $\mathbf{B}_r[\mathbf{u}_c]$ be the smallest Euclidean ball that contains all points in $\mathcal{U}$, where $\mathbf{u}_c$ is the center of the ball and r its radius. According to Jung's Theorem [1], the radius r of the smallest Epsilon ball $\mathbf{B}_r[\mathbf{u}_c]$ is

$$r = d\sqrt{\frac{K}{2(K+1)}}, \tag{S.3}$$

where d is the diameter of the set $\mathcal{U}$, which is defined as $d := \max_{\mathbf{u}^p,\mathbf{u}^q\in\mathcal{U}} ||\mathbf{u}^p - \mathbf{u}^q||$.

Since condition (3) holds for the all sequences in $\mathcal{S}$, the diameter of the set $\mathcal{S}$ cannot be greater than ϵ_θ. This also defines an upper bound for the diameter d of the set $\mathcal{U}$ as follows. For any given two sequences $x_p, x_q \in \mathcal{S}$,

$$\begin{aligned}
\epsilon_\theta^2 \geq ||\tilde{x}_p - \tilde{x}_q||^2 &= \sum_{k=1}^{T} |u_k^p - u_k^q|^2 \\
&= \sum_{k=1}^{K} |u_k^p - u_k^q|^2 + \sum_{k=K+1}^{T} |u_k^p - u_k^q|^2 \\
&= ||\mathbf{u}^p - \mathbf{u}^q||^2 + \sum_{k=K+1}^{T} |u_k^p - u_k^q|^2 \\
&\geq ||\mathbf{u}^p - \mathbf{u}^q||^2,
\end{aligned} \tag{S.4}$$

where the first equality holds because PCA is an orthogonal and therefore distance preserving transformation when all the coefficients are used. Since (S.4) holds for any $\mathbf{u}^p, \mathbf{u}^q \in \mathcal{U}$, the diameter d of the set $\mathcal{U}$ cannot be larger than ϵ_θ, *i.e.*,

$$d \leq \epsilon_\theta. \tag{S.5}$$

This inequality is tight, because theoretically there may be x_p and x_q that are perfectly reconstructed with the first K PCA coefficients and for which $\epsilon_\theta^2 = ||x_p - x_q||^2$, in which case $\sum_{k=K+1}^{T} |u_k^p - u_k^q|^2 = 0$ and $\epsilon_\theta^2 = ||x_p - x_q||^2 = ||\mathbf{u}^p - \mathbf{u}^q||^2$.

According to equations (S.3) and (S.5)

$$r = d\sqrt{\frac{K}{2(K+1)}} \leq \epsilon_\theta \sqrt{\frac{K}{2(K+1)}}, \tag{S.6}$$

thus, the Euclidean ball $\mathbf{B}_\epsilon[\mathbf{u}_c]$, where

$$\epsilon := \epsilon_\theta \sqrt{\frac{K}{2(K+1)}} \tag{S.7}$$

is guaranteed to contain all points in $\mathcal{U}$.

To complete the proof, we show that the ball $\mathbf{B}_\epsilon[\mathbf{u}_c]$ is contained entirely in at least one of the ϵ-expanded clusters $\{C_\epsilon^j\}_j$ obtained from clusters $\{C^j\}_j$. Since the union of the clusters $\{C^j\}_j$ covers the entire $\mathbb{R}^K$, the center of the Euclidean ball, $\mathbf{u}_c$, must be in one of the clusters, say C^j. Then, by definition of ϵ-expended clusters (see Definition 2), it holds that $\mathbf{B}_\epsilon[\mathbf{u}_c] \subseteq C_\epsilon^j$. In summary, we have shown that $\mathcal{U} \subset \mathbf{B}_\epsilon[\mathbf{u}_c] \subseteq C_\epsilon^j$.

The discussion after (S.5) together with Jung's Theorem suggests that in the absence of further information about the points in $\mathcal{U}$, one cannot find a smaller ball $\mathbf{B}_{\epsilon_0}[\mathbf{u}_c]$ that guarantees that the set $\mathcal{U}$ is contained entirely in any ϵ_0-expanded cluster. □

Appendix C

Here we give the proof of Lemma 2.2, which is copied below for convenience.

Lemma 1.2. *A point* $\mathbf{u}$ *belongs to* C_ϵ^j *if and only if* $\sum_{k=1}^{K} f\left(u_k; \theta_k^j, \theta_k^{j+1}\right) \leq \epsilon^2$, *where*

$$f\left(u_k; \theta_k^{j_k}, \theta_k^{j_k+1}\right) := \begin{cases} 0 & \text{if } u_k \in (\theta_{j_k}^j, \theta_k^{j_k+1}] \\ \min\limits_{t \in \{j_k, j_k+1\}} \left\{(\theta_k^t - u_k)^2\right\} & \text{else.} \end{cases} \tag{S.8}$$

Proof. By definition of expended clusters (Definition 2), a point $\mathbf{u}$ belongs to an expanded cluster C_ϵ^j if and only if the ℓ_2 norm of the difference between $\mathbf{u}$ and the point of C^j that is closest to $\mathbf{u}$ is at most ϵ. Let $\mathbf{v}^* = (v_1^*, \dots, v_k^*)$ be the point of C^j that is closest to $\mathbf{u}$; then, $\mathbf{u}$ belongs to C_j^ϵ if and only if $||\mathbf{u} - \mathbf{v}^*|| \leq \epsilon$, or, equivalently, $||\mathbf{u} - \mathbf{v}^*||^2 \leq \epsilon^2$. The point $\mathbf{v}^*$ by definition satisfies

$$\mathbf{v}^* = \operatorname*{argmin}_{\mathbf{v} \in C^j} ||\mathbf{u} - \mathbf{v}||^2. \tag{S.9}$$

The distance $||\mathbf{u} - \mathbf{v}||^2 = \sum_{k=1}^{K} (u_k - v_k)^2$ can be minimized by minimizing each term $(u_k - v_k)^2$. If $u_k \in (\theta_k^{j_k}, \theta_k^{j_k+1}]$, then $(u_k - v_k)^2$ can be made zero by picking $v_k^* = u_k$. Otherwise, $(u_k - v_k)^2$ is minimized by setting v_k^* to the threshold that is closest to u_k, in which case $\min\left\{(u_k - v_k)^2 : v_k \notin (\theta_k^{j_k}, \theta_k^{j_k+1}]\right\} = \min\left\{(\theta_k^{j_k} - u_k)^2, (\theta_k^{j_k+1} - u_k)^2\right\}$. Thus, the minimal (squared) distance is computed through the $f(\cdot)$ defined in (S.8), *i.e.*,

$$||\mathbf{u} - \mathbf{v}^*||^2 = \sum_{k=1}^{K} f\left(u_k; \theta_k^{j_k}, \theta_k^{j_k+1}\right). \tag{S.10}$$

Based on the argument in the beginning of this proof, $\mathbf{u} \in C_\epsilon^j$ if and only if $||\mathbf{u} - \mathbf{v}^*||^2 = \sum_{k=1}^{K} f\left(u_k; \theta_k^{j_k}, \theta_k^{j_k+1}\right) \leq \epsilon^2$. □

Appendix D

Here we show two clusters C^i and C^j are ϵ-neighbors if and only if the following inequality holds

$$\sum_{k=1}^{K} \min_{\substack{p \in \{i, i+1\} \\ t \in \{j, j+1\}}} \left\{\left(\theta_k^{p_k} - \theta_k^{t_k}\right)^2\right\} < \epsilon^2, \tag{S.11}$$

where $\{(\theta_k^{i_k}, \theta_k^{i_k+1}]\}_{k=1}^K$ and $\{(\theta_k^{j_k}, \theta_k^{j_k+1}]\}_{k=1}^K$ are the threshold intervals that define clusters C^i and C^j, respectively.

Let us recall that while defining the clusters we divide each of the K dimensions of $\mathbb{R}^K$ into M nonoverlapping intervals via a series of $M+1$ strictly increasing threshold values $\theta_k^0, \theta_k^1, \dots, \theta_k^M$ (where $\theta_k^0 = -\infty$ and $\theta_k^M = \infty$, see Section 2.2 of the main text). Also recall that each cluster C^j is defined by K intervals, where each interval is determined by two consecutive thresholds values (Definition 1), *i.e.*,

$$C^j := \{(u_1, \dots, u_K) \in \mathbb{R}^K : \theta_k^{j_k} < u_k \leq \theta_k^{j_k+1}\}. \tag{S.12}$$

Since there are K dimensions and M intervals per dimension, there are M^K unique clusters defined as above. The union of those clusters covers the entire $\mathbb{R}^K$ and since we pick the K intervals for each cluster uniquely, two different clusters do not overlap (*i.e.* $C^i \cap C^j = \emptyset$ if $i \neq j$).

By definition of ϵ-neighbourhood (Section 2.3 of main text), in order to establish that two clusters C^i and C^j are ϵ-neighbours, we need to find $\inf\limits_{\mathbf{u} \in C^i, \mathbf{v} \in C^j} ||\mathbf{u} - \mathbf{v}||$ and check if it is smaller than ϵ. Equivalently, one can find $\inf\limits_{\mathbf{u} \in C^i, \mathbf{v} \in C^j} ||\mathbf{u} - \mathbf{v}||^2$ and check if it is smaller than ϵ^2. For any $\mathbf{u} \in C^i$ and $\mathbf{v} \in C^j$, $||\mathbf{u} - \mathbf{v}||^2$ is

$$||\mathbf{u} - \mathbf{v}||^2 = \sum_{k=1}^{K} (u_k - v_k)^2. \tag{S.13}$$

One can minimize $||\mathbf{u}-\mathbf{v}||^2$ by minimizing the contribution of each dimension $(u_k - v_k)^2$ separately. Since $\theta_k^{i_k} < u_k \leq \theta_k^{i_k+1}$ and $\theta_k^{j_k} < v_k \leq \theta_k^{j_k+1}$, we can write:

$$\theta_k^{i_k} - \theta_k^{j_k+1} < u_k - v_k < \theta_k^{i_k+1} - \theta_k^{j_k}. \tag{S.14}$$

Since the threshold values of the kth dimension are defined to be strictly increasing (*i.e.* $\theta_k^0 < \theta_k^1 < \cdots < \theta_k^M$, see above), there are only three possibilities:

I. $\mathcal{I}_i = (\theta_k^{i_k}, \theta_k^{i_k+1}]$ and $\mathcal{I}_j = (\theta_k^{j_k}, \theta_k^{j_k+1}]$ are completely overlapping (when $i_k = j_k$), in which case

$$\theta_k^{i_k} - \theta_k^{j_k+1} < 0 < \theta_k^{i_k+1} - \theta_k^{j_k}. \tag{S.15}$$

That is, we can pick $u_k = v_k$ and make $u_k - v_k = 0$.

II. $\mathcal{I}_i = (\theta_k^{i_k}, \theta_k^{i_k+1}]$ is to the left of $\mathcal{I}_j = (\theta_k^{j_k}, \theta_k^{j_k+1}]$ and $\mathcal{I}_i \cap \mathcal{I}_j = \emptyset$ (when $i_k < j_k$), in which case

$$\theta_k^{i_k} - \theta_k^{j_k+1} < u_k - v_k < \theta_k^{i_k+1} - \theta_k^{j_k} < 0 \tag{S.16}$$

and we have that $\inf(u_k - v_k)^2 = (\theta_k^{i_k+1} - \theta_k^{j_k})^2$

III. $\mathcal{I}_i = (\theta_k^{i_k}, \theta_k^{i_k+1}]$ is to the right of $\mathcal{I}_j = (\theta_k^{j_k}, \theta_k^{j_k+1}]$ and $\mathcal{I}_i \cap \mathcal{I}_j = \emptyset$ (when $i_k > j_k$), in which case

$$0 < \theta_k^{i_k} - \theta_k^{j_k+1} < u_k - v_k < \theta_k^{i_k+1} - \theta_k^{j_k}, \tag{S.17}$$

and we have that $\inf(u_k - v_k)^2 = (\theta_k^{i_k} - \theta_k^{j_k+1})^2$.

The three conditions above imply that:

$$\inf_{\substack{\theta_k^{i_k} < u_k \leq \theta_k^{i_k+1} \\ \theta_k^{j_k} < v_k \leq \theta_k^{j_k+1}}} (u_k - v_k)^2 = \min_{\substack{p \in \{i, i+1\} \\ t \in \{j, j+1\}}} \left\{ \left(\theta_k^{p_k} - \theta_k^{t_k}\right)^2 \right\}. \tag{S.18}$$

Since (S.18) holds for $k = 1, \ldots, K$, we can write

$$\begin{aligned} \inf_{\mathbf{u} \in C^i, \mathbf{v} \in C^j} ||\mathbf{u} - \mathbf{v}||^2 &= \sum_{k=1}^{K} \inf_{\substack{\theta_k^{i_k} < u_k \leq \theta_k^{i_k+1} \\ \theta_k^{j_k} < v_k \leq \theta_k^{j_k+1}}} (u_k - v_k)^2 \\ &= \sum_{k=1}^{K} \min_{\substack{p \in \{i, i+1\} \\ t \in \{j, j+1\}}} \left\{ \left(\theta_k^{p_k} - \theta_k^{t_k}\right)^2 \right\}. \end{aligned} \tag{S.19}$$

By definition, two clusters C^i and C^j are ϵ-neighbors if and only if $\inf_{\mathbf{u} \in C^i, \mathbf{v} \in C^j} ||\mathbf{u} - \mathbf{v}||^2 < \epsilon^2$. According to (S.19), this is equivalent to saying that C^i and C^j are ϵ-neighbors if and only if (S.11) holds.

References

[1] Boris V Dekster. The Jung theorem for spherical and hyperbolic spaces. *Acta Mathematica Hungarica*, 67(4):315–331, 1995. 1