1. Introduction

In this supplementary document we show that the correlation between two sequences being at least $\rho_0$ is equivalent to the $\ell_2$ norm of their difference (after $z$-normalization) being at most $\sqrt{2T(1-\rho_0)}$ (Equation (S.2)), 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 $\epsilon$-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 $\rho_0$, i.e., $r(x,y) > \rho_0$, then $||\tilde{x} - \tilde{y}||_2 \leq \epsilon_0 = \sqrt{2T(1-\rho_0)}$, 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

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

The square of $\ell_2$ norm of $\tilde{x} - \tilde{y}$ is

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

$$||\tilde{x} - \tilde{y}||_2 = \sqrt{2T(1-r(x, y))} \leq \sqrt{2T(1-\rho_0)} = \epsilon_0.$$ \hspace{1cm} (S.3)

Appendix B

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

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

**Proof.** Since $U$ 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 $U$. Let $B_r[u_c]$ be the smallest Euclidean ball that contains all points in $U$, where $u_c$ is the center of the ball and $r$ its radius. According to Jung’s Theorem [1], the radius $r$ of the smallest Epsilon ball $B_r[u_c]$ is

$$r = d \sqrt{\frac{K}{2(K+1)}},$$ \hspace{1cm} (S.4)

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

Since condition (3) holds for the all sequences in $S$, the diameter of the set $S$ cannot be greater than $\epsilon_0$. This also defines an upper bound for the diameter $d$ of the set $U$ as follows. For any given two sequences $x_p, x_q \in S$,

$$\epsilon_0^2 \geq ||\tilde{x}_p - \tilde{x}_q||_2 = \sum_{k=1}^{K} |u^p_k - u^q_k|^2 \geq \sum_{k=1}^{K} |u^p_k - u^q_k|^2 + \sum_{k=K+1}^{T} |u^p_k - u^q_k|^2$$

$$\geq ||u^p - u^q||_2^2,$$ \hspace{1cm} (S.5)

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 $u^p, u^q \in U$, the diameter $d$ of the set $U$ cannot be larger than $\epsilon_0$, i.e.,

$$d \leq \epsilon_0.$$ \hspace{1cm} (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 $\epsilon_0^2 = ||x_p - x_q||_2^2$, in which
case $\sum_{k=K+1}^{T} |u_k^p - u_k^q|^2 = 0$ and $\epsilon_\theta^2 = ||x_p - x_q||^2 = ||u^p - u^q||^2$.

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

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

thus, the Euclidean ball $B_\epsilon[u_c]$, where

$$\epsilon := \epsilon_\theta \sqrt{\frac{K}{2(K+1)}} \quad (S.7)$$
is guaranteed to contain all points in $U$.

To complete the proof, we show that the ball $B_\epsilon[u_c]$ is contained entirely in at least one of the $\epsilon$-expanded clusters $\{C_j^i\}_j$, 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, $u_c$, must be in one of the clusters, say $C_j^i$. Then, by definition of $\epsilon$-expanded clusters (see Definition 2), it holds that $B_\epsilon[u_c] \subseteq C_j^i$. In summary, we have shown that $U \subseteq B_\epsilon[u_c] \subseteq C_j^i$.

The discussion after (S.5) together with Jung’s Theorem suggests that in the absence of further information about the points in $U$, one cannot find a smaller ball $B_\epsilon[u_c]$ that guarantees that the set $U$ is contained entirely in any $\epsilon_\theta$-expanded cluster.

**Appendix C**

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

**Lemma 1.2.** A point $u$ belongs to $C_j^i$ if and only if $\sum_{k=1}^{K} f(u_k; \theta_j^k, \theta_{j+1}^k) \leq \epsilon^2$, where

$$f(u_k; \theta_j^k, \theta_{j+1}^k) := \begin{cases} 0 & \text{if } u_k \in (\theta_j^k, \theta_{j+1}^k] \\ \min_{t \in (j,j+1]} \{ (\theta_t^k - u_k)^2 \} & \text{else.} \end{cases} \quad (S.8)$$

**Proof.** By definition of expanded clusters (Definition 2), a point $u$ belongs to an expanded cluster $C_j^i$ if and only if the $\ell_2$ norm of the difference between $u$ and the point of $C_j^i$ that is closest to $u$ is at most $\epsilon$. Let $v^* = (v_1^*, \ldots, v_K^*)$ be the point of $C_j^i$ that is closest to $u$; then, $u$ belongs to $C_j^i$ if and only if $||u - v^*|| \leq \epsilon$, or, equivalently, $||u - v^*||^2 \leq \epsilon^2$.

The point $v^*$ by definition satisfies

$$v^* = \arg\min_{v \in C_j^i} ||u - v||^2. \quad (S.9)$$

The distance $||u - v||^2 = \sum_{k=1}^{K} (u_k - v_k)^2$ can be minimized by minimizing each term $(u_k - v_k)^2$. If $u_k \in (\theta_j^k, \theta_{j+1}^k]$, 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

$$min \left\{ (u_k - v_k)^2 : v_k \notin (\theta_j^k, \theta_{j+1}^k] \right\} = \min \left\{ (\theta_j^k - u_k)^2, (\theta_{j+1}^k - u_k)^2 \right\}. \quad (S.10)$$

Thus, the minimal (squared) distance is computed through the $f(\cdot)$ defined in (S.8), i.e.,

$$||u - v^*||^2 = \sum_{k=1}^{K} f(u_k; \theta_j^k, \theta_{j+1}^k). \quad (S.11)$$

Based on the argument in the beginning of this proof, $u \in C_j^i$ if and only if $||u - v^*||^2 = \sum_{k=1}^{K} f(u_k; \theta_j^k, \theta_{j+1}^k) \leq \epsilon^2$.

**Appendix D**

Here we show two clusters $C^i$ and $C^j$ are $\epsilon$-neighbours if and only if the following inequality holds

$$\sum_{k=1}^{K} \min_{\rho \in \{i,i+1\}} \min_{t \in \{j,j+1\}} \left\{ (\theta_t^k - \theta_t^\rho)^2 \right\} < \epsilon^2, \quad (S.12)$$

where $\left\{ (\theta_t^j, \theta_{t+1}^j) \right\}_{k=1}^{K}$ and $\left\{ (\theta_t^i, \theta_{t+1}^i) \right\}_{k=1}^{K}$ 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 $\theta_0^j, \theta_1^j, \ldots, \theta_M^j$ (where $\theta_0^j = -\infty$ and $\theta_M^j = \infty$, 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.,

$$C^j := \{ (u_1, \ldots, u_K) \in \mathbb{R}^K : \theta_j^j < u_k \leq \theta_{j+1}^j \}. \quad (S.13)$$

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

$$||u - v||^2 = \sum_{k=1}^{K} (u_k - v_k)^2. \quad (S.14)$$
One can minimize $||u-v||^2$ 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:

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

Since the threshold values of the $k$th dimension are defined to be strictly increasing (i.e. $\theta_0^k < \theta_1^k < \cdots < \theta_k^M$, see above), there are only three possibilities:

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

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

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

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

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

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

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

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

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

The three conditions above imply that:

$$\inf_{\theta_k^{i_k} < u_k \leq \theta_k^{i_k+1}, \theta_k^{j_k} < v_k \leq \theta_k^{j_k+1}} (u_k - v_k)^2 = \min\left\{ (\theta_k^{p_k} - \theta_k^{q_k})^2 \right\}. \quad (S.18)$$

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

$$\inf_{u \in C^i, v \in C^j} ||u-v||^2 = \sum_{k=1}^{K} \inf_{\theta_k^{i_k} < u_k \leq \theta_k^{i_k+1}, \theta_k^{j_k} < v_k \leq \theta_k^{j_k+1}} (u_k - v_k)^2$$

$$= \sum_{k=1}^{K} \min_{p \in \{1, \ldots, i_k\}, q \in \{1, \ldots, j_k\}} \left\{ (\theta_k^{p_k} - \theta_k^{q_k})^2 \right\}. \quad (S.19)$$

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

References