Robust Nonparametric Nearest Neighbor Random Process Clustering

Michael Tschannen, *Student Member, IEEE*, and Helmut Bölcskei, *Fellow, IEEE*

***Abstract*—We consider the problem of clustering noisy finite-length observations of stationary ergodic random processes according to their generative models without prior knowledge of the model statistics and the number of generative models. Two algorithms, both using the L^1-distance between estimated power spectral densities (PSDs) as a measure of dissimilarity, are analyzed. The first one, termed nearest neighbor process clustering (NNPC), relies on partitioning the nearest neighbor graph of the observations via spectral clustering. The second algorithm, simply referred to as k-means (KM), consists of a single k-means iteration with farthest point initialization and was considered before in the literature, albeit with a different dissimilarity measure. We prove that both algorithms succeed with high probability in the presence of noise and missing entries, and even when the generative process PSDs overlap significantly, all provided that the observation length is sufficiently large. Our results quantify the tradeoff between the overlap of the generative process PSDs, the observation length, the fraction of missing entries, and the noise variance. Finally, we provide extensive numerical results for synthetic and real data and find that NNPC outperforms state-of-the-art algorithms in human motion sequence clustering.**

***Index Terms*—Clustering, stationary random processes, time series, nonparametric, k-means, nearest neighbors.**

I. Introduction

Consider a set of N noisy length-M observations of stationary ergodic discrete-time random processes stemming from $L < N$ (typically $L \ll N$) different generative processes, referred to as generative models henceforth. We want to cluster these observations according to their generative models without prior knowledge of the model statistics and the number of generative models, L. This problem arises in many domains of science and engineering where (large amounts of) data have to be divided into meaningful categories in an unsupervised fashion. Concrete examples include audio and video sequences [2], electrocardiography (ECG) recordings [3], industrial production indices [4], and financial time series [5], [6].

Common measures for quantifying the (dis)similarity of generative models typically rely on process statistics estimated from observations using either parametric or nonparametric methods. Parametric methods yield good performance when the (parametric) model the estimation is based on matches the true (unknown) model well. Nonparametric methods typically outperform parametric ones in case of model mismatch [7], a likely scenario in many practical applications. Existing random process clustering methods quantify the dissimilarity of observations using the Euclidean distance between estimated process model parameters [4], [5], cepstral coefficients [3], [8], or normalized periodograms [9]. Other methods rely on divergences (e.g., Kullback-Leibler divergence) between normalized periodograms [10], [11], use the distributional distance [12] between processes [13]–[15], or the earth mover's distance between copulas of the processes [5], [16]. In all cases the resulting distances are fed into a standard clustering algorithm such as k-means or hierarchical clustering. Another line of work employs a Bayesian framework to infer the cluster assignments, e.g., according to a maximum a posteriori criterion [17]. While many of these approaches have proven effective in practice, corresponding analytical performance results are scarce. Moreover, existing analytical results are mostly concerned with the asymptotic regime where the observation length goes to infinity while the number of observations is fixed (see, e.g., [4], [10], [11], [13], [14], [18]); the finite observation-length regime has attracted significantly less attention [5], [13], [15], [19].

Contributions: We consider two process clustering algorithms that apply to nonparametric generative models and employ the L^1-distance between estimated power spectral densities (PSDs) as dissimilarity measure. The first one, termed nearest neighbor process clustering (NNPC), relies on partitioning the q-nearest neighbor graph (q is a parameter of the algorithm) of the observations via normalized spectral clustering and, to the best of our knowledge, has not been considered in the literature before. NNPC is inspired by the thresholding-based subspace clustering (TSC) algorithm [20], which clusters a set of (high-dimensional) data points into a union of low-dimensional subspaces without prior knowledge of the subspaces, their dimensions, and their orientations. The second algorithm, which will be referred to as KM, consists of a single k-means iteration with farthest point initialization [21] and was first proposed in [13], albeit with a different dissimilarity measure.

Assuming real-valued stationary ergodic Gaussian processes with arbitrary (continuous) PSDs as generative models, we characterize the performance of NNPC and KM analytically for finite-length observations—potentially with missing entries—contaminated by independent additive real-valued white Gaussian noise. We find that both algorithms succeed with high probability even when the PSDs of the generative models exhibit significant overlap, all provided that the observation length is sufficiently large and the noise variance is sufficiently small. Our analytical results quantify the tradeoff between observation length, fraction of missing entries, noise variance, and distance between the (true) PSDs of the generative models.

The authors are with the Department of Information Technology and Electrical Engineering, ETH Zurich, Switzerland (e-mail: michaelt@nari.ee.ethz.ch; boelcskei@nari.ee.ethz.ch).

Part of this paper was presented at the 2015 IEEE International Symposium on Information Theory (ISIT) [1].

Furthermore, we prove that treating the finite-length observations as vectors in Euclidean space and clustering them using the TSC algorithm [20] (which inspired NNPC) results in performance strictly inferior to that obtained for NNPC. We argue that the underlying cause is to be found in TSC employing spherical distance as dissimilarity measure, thereby ignoring the stationary process structure of the observations. In a broader context this suggests that clustering observations of random processes using dissimilarity measures conceived with Euclidean geometry in mind, a popular ad-hoc approach in practice [22], can lead to highly suboptimal performance.

We evaluate the performance of NNPC and KM on synthetic and on real data, and find that NNPC outperforms state-of-the-art algorithms in human motion sequence clustering. Furthermore, NNPC and KM are shown to yield better clustering performance than single linkage and average linkage hierarchical clustering based on the L^1-distance between estimated PSDs. We also compare (L^1-based) NNPC and KM to their respective L^2 and L^∞-cousins and find that the original variants consistently yield better or the same results.

Relation to prior work: Numerical studies of time series clustering based on spectral clustering of the q-nearest neighbor graph using different dissimilarity measures (albeit not the L^1-distance, or, for that matter, other L^p-distances, between estimated PSDs) were reported in [23]. In [24] time series clustering is formulated as a community detection problem in graphs, but no analytical performance results are provided. KM with distributional distance as dissimilarity measure was proven in [13]—for more general (i.e., not necessarily Gaussian) generative models—to deliver correct clustering with probability approaching 1 as the observation length goes to infinity. We note, however, that estimating the distributional distance is computationally more demanding than estimating the L^1-distance between PSDs.

Notation: We use lowercase boldface letters to denote vectors, uppercase boldface letters to designate matrices, and the superscript $^\top$ stands for transposition. v_i is the ith entry of the vector $\mathbf{v}$. For the matrix $\mathbf{A}$, $\mathbf{A}_{i,j}$ denotes the entry in the i-th row and j-th column, $\mathbf{A}_i$ its i-th row, $\|\mathbf{A}\|_{2\to 2}$ its spectral norm, $\|\mathbf{A}\|_F \coloneqq (\sum_{i,j}|\mathbf{A}_{i,j}|^2)^{1/2}$ its Frobenius norm, and (for $\mathbf{A}$ square) $\mathrm{tr}(\mathbf{A}) = \sum_i \mathbf{A}_{i,i}$ its trace. $\mathbf{I}$ and $\mathbf{1}$ stand for the identity matrix and the all ones matrix (the latter not necessarily square), respectively. For matrices $\mathbf{A}$ and $\mathbf{B}$ of identical dimensions, $\mathbf{A}\circ\mathbf{B}$ is the Hadamard product, i.e., $(\mathbf{A}\circ\mathbf{B})_{i,j} = \mathbf{A}_{i,j}\mathbf{B}_{i,j}$. For the vector $\mathbf{b}\in\{0,1\}^n$, we let $\mathbf{P}_\mathbf{b} \coloneqq \mathrm{diag}(b_1,\ldots,b_n)$. The i-th element of a sequence x is denoted by $x[i]$. For a positive integer N, $[N]$ stands for the set $\{1,2,\ldots,N\}$. The (circular) convolution of $f,g\in L^2([0,1))$ is defined as $(f*g)(y) \coloneqq \int_0^1 f(x)\tilde{g}(y-x)\mathrm{d}x$, $y\in[0,1)$, where $\tilde{g}$ is the 1-periodic extension of g. log refers to the natural logarithm. $\mathbb{E}[X]$ denotes the expectation of the random variable X and the notation $Y\sim X$ indicates that the random variable Y has the same distribution as X. We say that a subgraph H of a graph G is connected if every pair of nodes in H can be joined by a path with nodes exclusively in H. A connected subgraph H of G is called a connected component of G if there are no edges between H and the remaining nodes in G.

II. Formal problem statement and algorithms

We consider the following clustering problem. Given the unlabeled data set $\mathcal{X} = \mathcal{X}_1\cup\ldots\cup\mathcal{X}_L$ of cardinality N, where $\mathcal{X}_\ell = \{x_i^{(\ell)}\}_{i=1}^{n_\ell}$ contains noisy length-M observations $x_i^{(\ell)}$—possibly with missing entries—of the real-valued discrete-time stationary ergodic random process $X^{(\ell)}[m]$, $m\in\mathbb{Z}$, corresponding to the ℓ-th generative model, find the partition $\mathcal{X}_1,\ldots,\mathcal{X}_L$. The statistics of the generative models and of the noise processes, and the number of generative models, are all assumed unknown.

Both clustering algorithms considered in this paper are based on the following measure for the distance between pairs of processes. With the PSD of $X^{(\ell)}$ denoted by $s^{(\ell)}(f)$, $f\in[0,1)$, we define the distance (dissimilarity) between the processes $X^{(k)}$ and $X^{(\ell)}$ as $d(X^{(k)},X^{(\ell)}) \coloneqq \frac{1}{2}\int_0^1 |s^{(k)}(f)-s^{(\ell)}(f)|\mathrm{d}f$. As argued below, for the algorithms to be meaningful, the different processes have to be of the same or at least of comparable power, which motivates the normalization $\int_0^1 s^{(\ell)}(f)\mathrm{d}f = 1$, $\ell\in[L]$. Now, this implies that $d(X^{(k)},X^{(\ell)}) \le \frac{1}{2}\int_0^1 |s^{(k)}(f)|\mathrm{d}f + \frac{1}{2}\int_0^1 |s^{(\ell)}(f)|\mathrm{d}f = \frac{1}{2}\int_0^1 s^{(k)}(f)\mathrm{d}f + \frac{1}{2}\int_0^1 s^{(\ell)}(f)\mathrm{d}f = 1$, and hence $d(X^{(k)},X^{(\ell)})\in[0,1]$. The distance measure $d(X^{(k)},X^{(\ell)})$ is close to 1 when $s^{(k)}$ and $s^{(\ell)}$ are concentrated on disjoint frequency bands and close to 0 when they exhibit similar support sets and shapes. In contrast, for general L^p-distances $d_{L^p}(X^{(k)},X^{(\ell)}) \coloneqq (\int_0^1 |s^{(k)}(f)-s^{(\ell)}(f)|^p\mathrm{d}f)^{\frac{1}{p}}$, with $p>1$, it is easy to see that $\int_0^1 s^{(\ell)}(f)\mathrm{d}f = 1$, $\ell\in[L]$, does not imply a uniform upper bound for $d_{L^p}(X^{(k)},X^{(\ell)})$. For example, $d_{L^\infty}(X^{(k)},X^{(\ell)})$ can become arbitrarily large if we set $s^{(k)}(f) = 1$, $f\in[0,1)$, and let $s^{(\ell)}$ have a sharp peak at some frequency $f_0\in[0,1)$, while maintaining $\int_0^1 s^{(\ell)}(f)\mathrm{d}f = 1$.

We now present the NNPC and the KM algorithms. Recall that NNPC is inspired by the TSC algorithm introduced in [20], and KM is obtained by replacing the distance measure in Algorithm 1 in [13] by the distance measure d defined above. In principle, NNPC and KM are applicable to general (real-valued) time series, in particular also to non-stationary random processes, but the definition of d above is obviously motivated by stationarity.

The NNPC algorithm. *Given a set $\mathcal{X}$ of N length-M observations, the number of generative models L (the estimation of L from $\mathcal{X}$ is discussed below), and the parameter q, carry out the following steps.*

Step 1: *For every $x_i\in\mathcal{X}$, estimate the PSD $\hat{s}_i(f)$ via the Blackman-Tukey (BT) estimator according to*

$$\hat{s}_i(f) \coloneqq \sum_{m=-M+1}^{M-1} g[m]\hat{r}_i[m]e^{-\mathrm{i}2\pi fm}, \quad \textit{where} \tag{1}$$

$$\hat{r}_i[m] \coloneqq \frac{1}{M}\sum_{n=0}^{M-|m|-1} x_i[n+m]x_i[n], \quad |m|\le M-1,$$

and $g[m]$, $m\in\mathbb{Z}$, is an even window function (i.e., $g[m] = g[-m]$) with $g[m]=0$ for $|m|\ge M$, and with bounded non-negative discrete-time Fourier transform (DTFT).

Step 2: *For every $x_i \in \mathcal{X}$, identify the set $\mathcal{T}_i \subset [N]\backslash\{i\}$ of cardinality q defined through*

$$d(x_i, x_j) \leq d(x_i, x_v), \quad \text{for all } j \in \mathcal{T}_i \text{ and all } v \notin \mathcal{T}_i,$$

where

$$d(x_i, x_j) \coloneqq \frac{1}{2} \int_0^1 |\hat{s}_i(f) - \hat{s}_j(f)| \, \mathrm{d}f. \tag{2}$$

Step 3: *Let $\mathbf{z}_j \in \mathbb{R}^N$ be the vector with ith entry $\exp(-2\, d(x_i, x_j))$, if $i \in \mathcal{T}_j$, and 0, if $i \notin \mathcal{T}_j$.*
Step 4: *Construct the adjacency matrix $\mathbf{A}$ according to $\mathbf{A} = \mathbf{Z} + \mathbf{Z}^\top$, where $\mathbf{Z} = [\mathbf{z}_1 \, \ldots \, \mathbf{z}_N]$.*
Step 5: *Apply normalized spectral clustering [25] to $(\mathbf{A}, L)$.*

Step 2 of NNPC determines the q-nearest neighbors of every observation w.r.t. to the distance measure d. We henceforth denote the corresponding nearest neighbor graph with adjacency matrix $\mathbf{A}$ constructed in Step 4 by G. The parameter q determines the minimum degree of G. Choosing q too small results in the observations stemming from a given generative model forming multiple connected components in G and hence not being assigned to the same cluster in Step 5. This problem can be countered by taking q larger, which, however, increases the chances of observations originating from different generative models being connected in G, thereby increasing the likelihood of incorrect cluster assignments. These tradeoffs are identical to those associated with the choice of the parameter q in TSC [20]. Note that spectral clustering is robust in the sense that it may deliver correct clustering even when G contains edges connecting observations that originate from different generative models, as long as the corresponding edge weights are sufficiently small.

The number of generative models, L, may be estimated in Step 4 based on the adjacency matrix $\mathbf{A}$ using the *eigengap heuristic* [25] (note that L is needed only in Step 5), which relies on the fact that the number of zero eigenvalues of the normalized Laplacian of G equals the number of connected components in G.

The KM algorithm [13]. *Given a set $\mathcal{X}$ of N length-M observations and the number of generative models L, carry out the following steps.*
Step 1: *Initialize $c_1 \coloneqq 1$ and $\hat{\mathcal{X}}_\ell \coloneqq \{\}$, for all $\ell \in [L]$.*
Step 2: *For every $x_i \in \mathcal{X}$, estimate the PSD $\hat{s}_i(f)$ via the BT estimator* (1)*.*
Step 3: for $p = 2$ **to** L **do:**

$$c_p \coloneqq \arg\max_{i \in [N]} \left(\min_{\ell \in [p-1]} d(x_i, x_{c_\ell}) \right),$$

with d as defined in (2)*.*
Step 4: for $i = 1$ **to** N **do:**

$$\ell^\star \leftarrow \arg\min_{\ell \in [L]} d(x_i, x_{c_\ell})$$
$$\hat{\mathcal{X}}_{\ell^\star} \leftarrow \hat{\mathcal{X}}_{\ell^\star} \cup \{x_i\}$$

KM selects the cluster centers in Step 3 and determines the assignments of the observations to these cluster centers in Step 4. Specifically, the algorithm selects x_1 as the first cluster center and then recursively determines the remaining cluster centers by maximizing the minimum distance to the cluster centers already chosen. In Step 4, it then assigns each observation to the closest cluster center (see Fig. 1). Intuitively, KM recovers the correct cluster assignments if the clusters are separated well enough. In practice, performing additional k-means iterations by alternating between cluster center refinement (simply by taking the refined center to be the average of the observations assigned to it) and reassignment of the data points to the refined cluster centers, can often improve performance. Numerical results on the effect of additional k-means iterations are provided in Sec. V. Our analytical results, however, all pertain to the case of a single k-means iteration per the definition of the KM algorithm above. Note that besides the number of clusters, L, KM does not have other parameters such as q in NNPC.

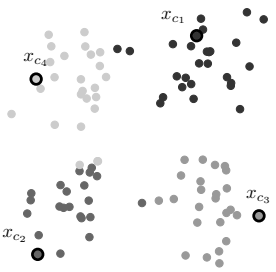

Fig. 1. Clustering of an example data set in $\mathbb{R}^2$ determined by KM with farthest point intialization and based on Euclidean distance.

Both NNPC and KM are based on comparisons of distances between observations, and are, therefore, meaningful only if the underlying processes $X^{(\ell)}$ are of comparable power $\int_0^1 s^{(\ell)}(f)\mathrm{d}f$. Indeed, when this is not the case, the distance between the observations is determined predominantly by the difference in power rather than the difference in PSD support sets and shapes. Note that the assumption of comparable power is not critical as, in practice, we can normalize the observations.

The choice of the window function g in (1) determines the bias-variance tradeoff of the BT estimator and through the distance estimates $d(x_i, x_j)$ ultimately the bias-variance tradeoff of NNPC and KM. For a discussion of window choice considerations for the BT estimator in a general context, we refer the reader to [7, Sec. 2.6]. We only remark here that the variance of the BT estimator goes to 0 as $M \to \infty$ under rather mild conditions on the process PSD and for $g \in \ell_1$ [26, Appendix B4]; the statistical data model employed in this paper (and described in the next section) satisfies these conditions on the PSDs.

We finally briefly discuss computational aspects of NNPC and KM for $L \ll N$, the situation typically encountered in practice. The BT PSD estimates (1) can be computed efficiently using the FFT. NNPC is a spectral clustering algorithm and as such requires the $N(N-1)/2$ distances between all pairs of observations to construct G. NNPC furthermore needs to determine the L eigenvectors corresponding to the L smallest eigenvalues of the $N \times N$ normalized graph Laplacian, which requires $O(N^3)$ operations (without exploiting potentially present structural properties of the Laplacian such as, e.g., sparsity). Spectral clustering then performs standard k-means clustering on the rows of the resulting $N \times L$ matrix of eigenvectors. The computational complexity of NNPC therefore becomes challenging for large N. Several spectral clustering methods suitable for data sets of up to millions of

observations are available in the literature, see, e.g., [27]–[29]. KM, on the other hand, is computationally considerably less expensive, requiring only $O(NL^2)$ distance computations.

We finally note that both NNPC and KM along with the corresponding analytical performance guarantees presented in the next section can easily be generalized to stationary ergodic vector-processes $\mathbf{x}^{(\ell)}[m] \in \mathbb{R}^n$, $m \in \mathbb{Z}$. Specifically, with the spectral density matrices $\mathbf{S}^{(\ell)}(f) \coloneqq \sum_{m=-\infty}^{\infty} \mathbb{E}\big[\mathbf{x}^{(\ell)}[m](\mathbf{x}^{(\ell)}[0])^\top\big] e^{-\mathrm{i}2\pi f m} \in \mathbb{R}^{n\times n}$, $\ell \in [L]$, one defines the distance measure

$$d(\mathbf{x}^{(k)}, \mathbf{x}^{(\ell)}) = \sum_{u,v\in[n]} \int_0^1 |\mathbf{S}_{u,v}^{(k)}(f) - \mathbf{S}_{u,v}^{(\ell)}(f)| \,\mathrm{d}f$$

and employs the BT estimator in (1) component-wise to estimate $\mathbf{S}^{(\ell)}(f)$. As this requires the computation of distances between all scalar random process components, evaluating $d(\mathbf{x}^{(k)}, \mathbf{x}^{(\ell)})$ in the vector case incurs $n(n+1)/2$ (exploiting the symmetry of $\mathbf{S}^{(\ell)}(f)$) times the cost in the scalar case. All other steps of NNPC and KM remain unchanged and hence have the same computational complexity as in the scalar case. For simplicity of exposition, we focus on the scalar case throughout the paper.

III. Analytical performance results

We start by describing the statistical data model underlying our analytical performance results. Recall that both NNPC and KM are, in principle, applicable to general real-valued time series including non-stationary processes. The performance analysis conducted here applies, however, to stationary processes. In addition, we take into account additive noise and potentially missing entries. Specifically, we assume that the $x_i^{(\ell)}$ are obtained as contiguous length-M observations of $\check{X}^{(\ell)}[m] \coloneqq U^{(\ell)}[m](X^{(\ell)}[m] + W^{(\ell)}[m]), m \in \mathbb{Z}$, where $U^{(\ell)}$ is a Bernoulli process with i.i.d. entries according to $\mathrm{P}\big[U^{(\ell)}[m] = 1\big] = 1 - \mathrm{P}\big[U^{(\ell)}[m] = 0\big] = p > 0$ (we henceforth refer to p as sampling probability), $X^{(\ell)}$ is zero-mean stationary Gaussian with PSD $s^{(\ell)}(f)$, and $W^{(\ell)}$ is a zero-mean white Gaussian noise process with variance σ^2. The autocorrelation functions (ACFs) $r^{(\ell)}[m] \coloneqq \int_0^1 s^{(\ell)}(f) e^{\mathrm{i}2\pi f m} \mathrm{d}f$ of the $X^{(\ell)}$ are assumed absolutely summable, i.e., $\sum_{m=-\infty}^{\infty} |r^{(\ell)}[m]| < \infty$, $\ell \in [L]$, which implies continuity of the $s^{(\ell)}(f)$ and thereby ergodicity of the corresponding processes $X^{(\ell)}$ [30]. Moreover, we take the PSDs to be normalized according to $\int_0^1 s^{(\ell)}(f)\mathrm{d}f = 1$, $\ell \in [L]$, and we let $B \coloneqq \max_{\ell\in[L]} \sup_{f\in[0,1)} s^{(\ell)}(f)$. We further assume that $U^{(\ell)}$, $X^{(\ell)}$, and $W^{(\ell)}$ are mutually independent. As a consequence, the noisy process $\tilde{X}^{(\ell)}[m] \coloneqq X^{(\ell)}[m] + W^{(\ell)}[m]$ and the Bernoulli process $U^{(\ell)}$ are jointly stationary ergodic so that $\check{X}^{(\ell)}[m] = U^{(\ell)}[m]\tilde{X}^{(\ell)}[m]$ is stationary ergodic by [31, Prop. 3.36]. Furthermore, we denote the ACF of the noisy process $\tilde{X}^{(\ell)}[m]$ by $\tilde{r}^{(\ell)}[m]$ and note that $\tilde{r}^{(\ell)}[m] = r^{(\ell)}[m] + \sigma^2\delta[m]$. It follows from $\check{X}^{(\ell)}[m] = U^{(\ell)}[m]\tilde{X}^{(\ell)}[m]$ that $\check{r}^{(\ell)}[m] = u[m]\tilde{r}^{(\ell)}[m]$, where $u[m] \coloneqq p$ for $m = 0$, and $u[m] \coloneqq p^2$, else. For each ℓ, the $x_i^{(\ell)}$ may either stem from independent realizations of $\check{X}^{(\ell)}$ or correspond to different (possibly overlapping) length-M segments of a given realization of $\check{X}^{(\ell)}$. In the latter case the $x_i^{(\ell)}$ will not be statistically independent in general. This is, however, not an issue as statistical independence is not required in our analysis, neither across observations stemming from a given generative model nor across observations originating from different generative models.

Multiplication of $\tilde{X}^{(\ell)}$ by the Bernoulli process $U^{(\ell)}$ models, e.g., a sampling device which acquires only every $(1/p)$-th sample on average. Moreover, in practice we could deliberately subsample in order to speed up the computation of the distances d when the observation length M is large. Specifically, with observation length M and sampling probability p, we get $\approx (1-p)M$ entries of $x_i^{(\ell)}$ that are set to 0, which can be exploited when computing the BT estimates using the FFT [32].

Naïvely applying the BT estimator to the $x_i^{(\ell)}$ delivers PSD estimates that, owing to $\check{r}^{(\ell)}[m] = u[m]\tilde{r}^{(\ell)}[m]$, can be severely biased compared to estimates that would be obtained from observations with no missing entries. Indeed, as $\check{r}^{(\ell)}[0] = p\,\tilde{r}^{(\ell)}[0]$ and $\check{r}^{(\ell)}[m] = p^2\tilde{r}^{(\ell)}[m]$, for $m \neq 0$, for small p, $u[m]$ assigns a much larger weight to lag $m = 0$ than to the lags $m \neq 0$. To correct this bias, we assume in the remainder of the paper (in particular also in the analytical results below) that the BT estimates in (1) are computed for the window function $\hat{g}[m] \coloneqq g[m]/u[m]$, $m \in \mathbb{Z}$, i.e., g in Algorithms 1 and 2 is replaced by $\hat{g}$. While $\hat{g}$ remains even and supported on $\{-M+1, \ldots, M-1\}$, BT PSD estimates based on $\hat{g}$ are not guaranteed to be non-negative (in contrast to estimates based on g directly [7, Sec. 2.5.2]) as the DTFT of $\hat{g}$ may not be non-negative. This is, however, not an issue as we consider distances between PSDs only and do not explicitly make use of the positivity property of PSDs. We note that bias correction requires knowledge of p, which can be obtained in practice simply by estimating the average number of non-zero entries in the $x_i^{(\ell)}$. In addition, we will assume that $g[0] = 1$ and g has a bounded DTFT $g(f)$, i.e., $0 \leq g(f) \leq A < \infty$, $f \in [0, 1)$. An example of such a window function is the Bartlett window (see (10)) used in the experiments in Sec. V. Our performance results will be seen to depend on the maximum ACF moment $\mu_{\max} \coloneqq \max_{\ell\in[L]} \mu^{(\ell)}$, where $\mu^{(\ell)} \coloneqq \sum_{m=-\infty}^{\infty} |h[m]||r^{(\ell)}[m]|$ with

$$h[m] \coloneqq \begin{cases} 1 - g[m](1 - |m|/M), & \text{for } |m| < M \\ 1, & \text{otherwise.} \end{cases} \tag{3}$$

We are now ready to state our main results. For the NNPC algorithm, we provide a sufficient condition for the following *no false connections (NFC) property* to hold. Recall that G is the nearest neighbor graph with adjacency matrix $\mathbf{A}$, as constructed in Step 4 of NNPC.

Definition 1 (No False Connections Property)**.** *G satisfies the no false connections property if, for all $\ell \in [L]$, all nodes corresponding to $\mathcal{X}_\ell$ are connected exclusively to nodes corresponding to $\mathcal{X}_\ell$.*

We henceforth say that "NNPC succeeds" if the NFC property is satisfied. Although the NFC property alone does not guarantee correct clustering, it was found to be a sensible performance measure for subspace clustering algorithms (see, e.g., [20], [33], [34]). To ensure correct clustering one would

additionally need the subgraph of G corresponding to $\mathcal{X}_\ell$ to be connected, for each $\ell \in [L]$ [25, Prop. 4; Sec. 7]. Establishing conditions for this to hold appears to be difficult, at least for the statistical data model considered here.

Theorem 1. *Let $\mathcal{X}$ be generated according to the statistical data model described above and assume that $q \leq \min_{\ell\in[L]}(n_\ell - 1)$. Then, the clustering condition*

$$\min_{\substack{k,\ell\in[L]:\\ k\neq\ell}} d(X^{(k)}, X^{(\ell)}) > \frac{8\sqrt{2}A(B+\sigma^2+\sqrt{2}(1+p)(1+\sigma^2))}{p^2}\sqrt{\frac{\log M}{M}}+2\mu_{\max} \tag{4}$$

guarantees that G satisfies the NFC property with probability at least $1 - 6N/M^2$.

The condition $q \leq \min_{\ell\in[L]}(n_\ell - 1)$ is necessary for the NFC property to hold as choosing $q > \min_{\ell\in[L]}(n_\ell-1)$ would force NNPC to select observations from $\mathcal{X}\backslash\mathcal{X}_\ell$ for at least one of the data points $x_i^{(\ell)}$. As the n_ℓ are unknown in practice, one has to guess q while taking into account the tradeoffs related to the choice of q as discussed in Sec. II.

Our main result for KM comes with a performance guarantee that is stronger than the NFC property, namely it ensures correct clustering; accordingly, "KM succeeds" henceforth refers to KM delivering correct clustering. This stronger result is possible as KM does not entail a spectral clustering step and is therefore much easier to analyze. On the other hand, NNPC typically outperforms KM in practice, as seen in the numerical results in Sec. V.

Theorem 2. *Let $\mathcal{X}$ be generated according to the statistical data model described above. Then, under the clustering condition* (4)*, the partition $\hat{\mathcal{X}}_1, \ldots, \hat{\mathcal{X}}_L$ of $\mathcal{X}$ inferred by KM corresponds to the true partition $\mathcal{X}_1, \ldots, \mathcal{X}_L$ with probability at least $1 - 6N/M^2$.*

The proofs of Theorems 1 and 2 are provided in Appendix A. We first note that the clustering condition (4) depends on a few model parameters only and all constants involved are explicit. Furthermore, the condition is identical for NNPC and KM, although the performance guarantee we obtain for KM (namely correct clustering) is stronger than that for NNPC (namely the NFC property). This is a consequence of both proofs relying on the same "separation condition" (namely (11) in Appendix A) and the clustering condition (4) being sufficient for this separation condition to hold (see Appendix A for further details).

Theorems 1 and 2 essentially state that NNPC and KM succeed even when the PSDs $s^{(\ell)}$ of the $X^{(\ell)}$ overlap significantly and the observations have missing entries and are contaminated by strong noise, all this provided that the observation length M is sufficiently large and the window function g is chosen to guarantee small $\mu_{\max}$. The clustering condition (4) suggests (recall that it is sufficient only) a tradeoff between the amount of overlap of pairs of PSDs $\{s^{(k)}, s^{(\ell)}\}$ (through $\min_{k,\ell\in[L]:\, k\neq\ell} d(X^{(k)}, X^{(\ell)})$), the observation length M, the sampling probability p, and the noise variance σ^2. It indicates, for example, that both algorithms tolerate shorter observation length M, more missing entries (i.e., smaller p), and stronger noise (i.e., larger σ^2) as the pairs $\{s^{(k)}, s^{(\ell)}\}$, $k \neq \ell$, overlap less and hence $\min_{k,\ell\in[L]:\, k\neq\ell} d(X^{(k)}, X^{(\ell)})$ is larger. Keeping σ^2 and p fixed, the first term on the RHS of (4), which accounts for the PSD estimation error owing to finite observation length M, vanishes as M becomes large. Since $d(X^{(k)}, X^{(\ell)}) \in [0, 1]$, we need $\mu_{\max} \ll 1$ to ensure that the clustering condition can be satisfied for finite M. To see how this can be accomplished, we consider $r^{(\ell)}$ of small effective support relative to M, i.e., $r^{(\ell)}[m] \approx 0$ for $m \geq M_0$ with $M_0 \ll M$, which is essentially equivalent to requiring that the $s^{(\ell)}$ be sufficiently smooth. We then choose g such that $g[m] \approx 1$ for $m \leq M_0$ and note that this ensures $h[m] \approx 1 - g[m](1 - |m|/M) \approx 1 - g[m] \approx 0$, for $m \leq M_0 \ll M$. Thanks to $\mu^{(\ell)} = \sum_{m=-\infty}^{\infty} |h[m]||r^{(\ell)}[m]| \approx \sum_{m=-M_0}^{M_0} |h[m]||r^{(\ell)}[m]| \ll 1$, we then get $\mu_{\max} \ll 1$. The clustering condition (4) can hence, indeed, be satisfied for finite M if the $r^{(\ell)}$ have small effective support. Note that the choice of g will affect the constant A (recall that $0 \leq g(f) \leq A < \infty$, $f \in [0, 1)$). Specifically, windows g of larger effective support have larger corresponding A in general.

To ensure high probability of success, we need to take $M \gg \sqrt{N}$, i.e., the observation length has to be large relative to the square root of the number of observations. We note that the results in Theorems 1 and 2 can easily be extended to colored noise processes, as long as the noise PSDs are identical for all $\ell \in [L]$.

We emphasize that the vast majority of analytical performance results for random process clustering available in the literature pertain to the asymptotic regime $M \to \infty$, with N fixed. The findings in [10], [11] are closest in spirit to ours and show that pairs of observations stemming from different generative models can be discriminated consistently (in the statistical sense), for $M \to \infty$, via a PSD-based distance measure, provided that the PSDs of all pairs of generative models differ on a set of positive Lebesgue measure.

Finally, we note that generalization of our analytical results to processes other than Gaussian such as, e.g., subgaussian processes, seems difficult as a version of the concentration inequality [35, Lem. 1], upon which the proofs of Theorems 1 and 2 rely, does not appear to be available for non-Gaussian random vectors with dependent entries (see [36] for details). For i.i.d. subgaussian processes such an inequality was reported in [37]; this is, however, not of interest here as i.i.d. processes have flat PSDs.

IV. Comparison with thresholding-based subspace clustering

For finite observation length M, the random process clustering problem considered here can also be cast as a classical subspace clustering problem simply by interpreting the observations $x_i^{(\ell)}$ as vectors $\mathbf{x}_i^{(\ell)} \in \mathbb{R}^M$. Numerical results, not reported here[1], demonstrate, however, that this approach leads to NNPC significantly outperforming its subspace clustering cousin, the TSC algorithm [20]. Our next result, Proposition 1

[1]but available at http://www.nari.ee.ethz.ch/commth/research/

below, provides analytical underpinning for this observation. Before stating the formal result, we develop some intuition. To this end, we consider statistically independent observations and set $p = 1$ (i.e., no missing entries). We then note that the clustering condition (4) for NNPC ensures that (using (15) and (16) in (11) together with (22) and (32), cf. Appendices A and B)

$$\mathrm{P}\Big[d(x_j^{(k)}, x_i^{(\ell)}) \leq d(x_v^{(\ell)}, x_i^{(\ell)})\Big] < \frac{6}{M^2}, \tag{5}$$

for $i \neq v$, all j, and $k \neq \ell$. This guarantees that the probability of the NFC property being violated becomes small for M large, in particular, even when the PSD pairs $\{s^{(k)}, s^{(\ell)}\}$, $k \neq \ell$, overlap substantially and $\mathrm{SNR} \coloneqq r^{(\ell)}[0]/\sigma^2 = 1/\sigma^2 < 1$, $\ell \in [L]$. For TSC (which constructs the sets $\mathcal{T}_i$ such that $|\langle \mathbf{x}_j, \mathbf{x}_i\rangle| \geq |\langle \mathbf{x}_v, \mathbf{x}_i\rangle|$ for all $j \in \mathcal{T}_i$ and all $v \notin \mathcal{T}_i$) applied to $\{\mathbf{x}_i^{(\ell)}\}_{i\in n_\ell, \ell\in[L]}$ the probability corresponding to the LHS of (5) is $\mathrm{P}[|\langle \mathbf{x}_j^{(k)}, \mathbf{x}_i^{(\ell)}\rangle| \geq |\langle \mathbf{x}_v^{(\ell)}, \mathbf{x}_i^{(\ell)}\rangle|]$. The next proposition establishes that this probability remains bounded away from 0 even when M grows large, unless the observations are noiseless and all the PSD pairs $\{s^{(k)}, s^{(\ell)}\}$, $k \neq \ell$, are supported on essentially disjoint frequency bands. These conditions are, however, hardly encountered in practice, and the corresponding clustering problem can be considered easy. The superior performance of NNPC as compared to TSC stems from the TSC similarity measure not exploiting the stationarity of the generative models. We proceed to the formal statement.

Proposition 1. *Let $x_i^{(\ell)}$ be a contiguous length-M observation of $\tilde{X}^{(\ell)}$ (note that we consider the case $p = 1$). Assume that the $x_i^{(\ell)}$ are independent across $\ell \in [L]$ and $i \in [n_\ell]$. Denote the vectors containing the elements of the $x_i^{(\ell)}$ by $\mathbf{x}_i^{(\ell)} \in \mathbb{R}^M$ and the corresponding covariance matrices by $\tilde{\mathbf{R}}^{(\ell)} \coloneqq \mathbf{R}^{(\ell)} + \sigma^2\mathbf{I}$, with $\mathbf{R}^{(\ell)}_{v,w} = r^{(\ell)}[w-v] = r^{(\ell)}[v-w]$, $\ell \in [L]$. Then, for $k \neq \ell$ and $v \neq i$, we have*

$$\mathrm{P}\left[\left|\left\langle \mathbf{x}_j^{(k)}, \mathbf{x}_i^{(\ell)}\right\rangle\right| \geq \left|\left\langle \mathbf{x}_v^{(\ell)}, \mathbf{x}_i^{(\ell)}\right\rangle\right|\right] \geq \frac{1}{5\pi}\arctan\left(\frac{\sqrt{\mathrm{tr}\left(\tilde{\mathbf{R}}^{(k)}\tilde{\mathbf{R}}^{(\ell)}\right)}}{5\sqrt{3}\sqrt{\mathrm{tr}\left(\tilde{\mathbf{R}}^{(\ell)}\tilde{\mathbf{R}}^{(\ell)}\right)}}\right). \tag{6}$$

Proof: See Appendix C.

Remark 1. *Note that, in contrast to Theorems 1 and 2, Proposition 1 assumes the observations to be statistically independent. This assumption turns out to be critical in the proof of Proposition 1.*

We next show, as announced, that the RHS of (6) remains strictly positive even when M grows large, unless the observations are noiseless and all pairs of PSDs have essentially disjoint support. To this end, we examine the behavior of $(1/M)\mathrm{tr}(\tilde{\mathbf{R}}^{(k)}\tilde{\mathbf{R}}^{(\ell)})$, $k \neq \ell$, and $(1/M)\mathrm{tr}(\tilde{\mathbf{R}}^{(\ell)}\tilde{\mathbf{R}}^{(\ell)})$ (the motivation for the normalization by M will become clear later). First, note that $(1/M)\mathrm{tr}(\tilde{\mathbf{R}}^{(\ell)}\tilde{\mathbf{R}}^{(\ell)}) < \infty$ as $\sum_{m=-\infty}^{\infty}\left(\tilde{r}^{(\ell)}[m]\right)^2 < \infty$ by virtue of $\tilde{r}^{(\ell)} = r^{(\ell)}[m] + \sigma^2\delta[m] \in \ell_1$, which, in turn, follows from the assumption $r^{(\ell)} \in \ell_1$. The probability in (6) is hence bounded away from 0 unless $(1/M)\mathrm{tr}(\tilde{\mathbf{R}}^{(k)}\tilde{\mathbf{R}}^{(\ell)}) \approx 0$. It therefore remains to identify conditions for $(1/M)\mathrm{tr}(\tilde{\mathbf{R}}^{(k)}\tilde{\mathbf{R}}^{(\ell)}) \approx 0$ to hold. To this end, we note that

$$\begin{aligned}\frac{1}{M}\mathrm{tr}\left(\tilde{\mathbf{R}}^{(k)}\tilde{\mathbf{R}}^{(\ell)}\right) &= \frac{1}{M}\sum_{m=0}^{M-1}\sum_{n=0}^{M-1}\tilde{r}^{(k)}[n-m]\tilde{r}^{(\ell)}[n-m] \\ &= \sum_{m\in\mathcal{M}}\left(1-\frac{|m|}{M}\right)\tilde{r}^{(k)}[m]\tilde{r}^{(\ell)}[m] \qquad (7)\\ &= \int_0^1 (w * \tilde{s}^{(k)})(f)\tilde{s}^{(\ell)}(f)\mathrm{d}f, \qquad (8)\end{aligned}$$

where (7) is due to the Toeplitz structure and the symmetry of $\tilde{\mathbf{R}}^{(k)}$ and $\tilde{\mathbf{R}}^{(\ell)}$, (8) is by Parseval's Theorem, $\mathcal{M} \coloneqq \{-M+1, -M+2, \ldots, M-1\}$, and $w(f) \coloneqq \sum_{m\in\mathcal{M}}(1 - |m|/M)e^{-\mathrm{i}2\pi fm} = \sin^2(\pi fM)/(M\sin^2(\pi f))$. As $w(f)$ is strictly positive on the interval $[0, 1)$ (apart from its zeros which are supported on a set of measure 0) and the $\tilde{s}^{(\ell)}(f)$, $\ell \in [L]$, are non-negative, (8) is bounded away from 0 for finite M. As M grows large, $w(f)$ approaches the Dirac delta distribution, i.e., the "leakage" induced by w becomes small and we have (8) $\approx \int_0^1 \tilde{s}^{(k)}(f)\tilde{s}^{(\ell)}(f)\mathrm{d}f$. This integral vanishes for all $k \neq \ell$ if and only if $\sigma^2 = 0$ (recall that $\tilde{s}^{(\ell)}(f) = s^{(\ell)}(f) + \sigma^2$, $\ell \in [L]$) and all pairs $\{s^{(k)}, s^{(\ell)}\}$, $k \neq \ell$, are supported on essentially disjoint frequency bands. This establishes the claim made above and concludes the argument.

V. Numerical results[2]

We present numerical results for NNPC and KM on synthetic and on real data. In addition, we report results for KM followed by 100 k-means iterations (see the discussion in Sec. II); this variant of KM will be referred to as iterated k-means (KMit). Furthermore, we compare NNPC, KM, and KMit with single linkage (SL), average linkage (AL), and complete linkage (CL) hierarchical clustering [38, Sec. 14.3.12], all based on the L^1-distance measure (2). We also investigate variants of NNPC, KM, and KMit with the L^1-distance measure replaced by $d_{L^2}(x_i, x_j) \coloneqq (\int_0^1 |\hat{s}_i(f) - \hat{s}_j(f)|^2\,\mathrm{d}f)^{\frac{1}{2}}$, and variants of NNPC and KM with the L^1-distance measure replaced by $d_{L^\infty}(x_i, x_j) \coloneqq \sup_{f\in[0,1)}|\hat{s}_i(f) - \hat{s}_j(f)|$ (we do not consider KMit here as d_{L^∞}-based k-means iterations do not seem sensible). NNPC and KM were implemented strictly according to the corresponding algorithm descriptions in Sec. II. For SL, AL, and CL, we use the functions built into Matlab. Throughout, performance is measured in terms of the clustering error (CE), i.e., the fraction of misclustered data points, defined as

$$\mathrm{CE}(\hat{\mathbf{c}}, \mathbf{c}) = \min_{\pi}\left(1 - \frac{1}{N}\sum_{i=1}^{N} 1_{\{\pi(\hat{c}_i)=\pi(c_i)\}}\right),$$

where $\mathbf{c} \in [L]^N$ and $\hat{\mathbf{c}} \in [L]^N$ are the true and the estimated assignments, respectively, and the minimum is taken over all permutations $\pi\colon [L] \to [L]$. We report running times (excluding time for loading the data) obtained on a MacBook Pro with a 2.5 GHz Intel Core i7 CPU with 16 GB RAM.

[2]Matlab code available at http://www.nari.ee.ethz.ch/commth/research/

A. Synthetic data

We investigate the tradeoff between the minimum distance $\min_{k,\ell\in[L]:\, k\neq\ell} d(X^{(k)}, X^{(\ell)})$, the observation length M, the sampling probability p, and the noise variance σ^2 as indicated by the clustering condition (4). Recall that the clustering condition is only sufficient (and for NNPC guarantees the NFC property only). It is therefore unclear a priori to what extent the CE, indeed, follows the behavior indicated by the clustering condition.

We consider $L = 2$ second-order AR generative processes with PSDs of the form

$$s_{a,\nu}(f) = \frac{b^2(a,\nu)}{\left|1 - 2a\cos(\nu)e^{\mathrm{i}2\pi f} + a^2 e^{\mathrm{i}4\pi f}\right|^2}, \tag{9}$$

where $\nu \in [0,\pi]$, $a \in (0,1)$, and $b^2(a,\nu) = 1/(\int_0^1 1/\left|1 - 2a\cos(\nu)e^{\mathrm{i}2\pi f} + a^2 e^{\mathrm{i}4\pi f}\right|^2 \mathrm{d}f)$ ensures that $\int_0^1 s_{a,\nu}(f)\mathrm{d}f = 1$. Fig. 2 shows examples of $s_{a,\nu}(f)$ for different choices of a and ν. In the ensuing experiments, we set $s^{(1)}(f) = s_{0.6,0.7\pi}(f)$ and $s^{(2)}(f) = s_{0.6,\nu_2}(f)$, where ν_2 is variable and controls the locations of the peaks of $s^{(2)}$ and thereby the distance $d(X^{(1)}, X^{(2)})$. Indeed, varying ν_2 shifts the locations of the peaks of $s^{(2)}$ while essentially maintaining its shape. For the BT PSD estimator, we use a Bartlett window of length W defined as

$$g_W^B[m] := \begin{cases} 1 - |m|/\lfloor W/2\rfloor, & \text{for } |m| \leq \lfloor W/2\rfloor \\ 0, & \text{otherwise,} \end{cases} \tag{10}$$

and we set $W = 101$. Note that g_W^B satisfies the assumptions made about g in Sec. III. The number of generative models $L = 2$ is assumed known throughout. The performance of NNPC is found (corresponding results are not shown here) to be rather insensitive to the choice of the parameter q as long as $10 \leq q \leq 25$; we set $q = 10$. For a given quadruple (ν_2, M, σ, p), a realization of the data set $\mathcal{X}$ is obtained by sampling $n = 25$ independent observations from $\check{X}^{(1)}$ and $\check{X}^{(2)}$ each, and the CE is estimated by averaging over 10 such independent realizations of $\mathcal{X}$. We do not normalize the BT PSD estimates to unit power.

Fig. 3 shows that NNPC, KM, and KMit all exhibit roughly the same qualitative behavior as a function of $d(X^{(1)}, X^{(2)})$, M, $1/p$, and σ. In particular, for large enough $d(X^{(1)}, X^{(2)})$ all three algorithms yield a CE close to 0 even when σ^2 exceeds the signal power (i.e., when $\mathrm{SNR} < 1$), when the observations have missing entries ($p < 1$), and when M is small. All three algorithms tolerate more noise and more missing entries as the observation length increases. These numerical results are in line with the *qualitative* tradeoff indicated by the (sufficient) clustering condition (4). The numerical constants in (4) are, however, too big for the clustering condition (4) to be sharp. NNPC consistently achieves the lowest CE, followed by KMit, and KM. The performance advantage of NNPC over KM and KMit can be attributed to the spectral clustering step, which leads to increased robustness to noise and missing entries. Finally, we note that KMit often yields a significantly lower CE than KM.

The results in Fig. 5 indicate that the *qualitative* dependence of the CE on $d(X^{(1)}, X^{(2)})$, M, $1/p$, and σ for SL, AL, and CL is essentially identical to that for NNPC, KM, and KMit. For large σ and small $d(X^{(1)}, X^{(2)})$, M, or p, SL and AL lead, however, to a significantly larger CE than NNPC, KM, and KMit. The CE for CL is comparable to, but slightly larger than, that of KMit and significantly larger than that of NNPC.

Comparing the CE for NNPC, KM, and KMit in Fig. 3 with that obtained for their d_{L^2} and d_{L^∞}-cousins in Figs. 4 and 6, respectively, we note that, for all values of $d(X^{(1)}, X^{(2)})$, M, σ, and p the d_{L^2}-based variants of NNPC, KM, and KMit and the d_{L^∞}-based variants of NNPC and KM yield the same or larger CE than the respective original variants. This justifies usage of the L^1-based distance measure (2) also from a practical point of view. Finally, we note that normalizing the model PSDs (9) according to $(\int_0^1 s_{a,\nu}^2(f)\mathrm{d}f)^{\frac{1}{2}} = 1$ for d_{L^2} and $\sup_{f\in[0,1)} s_{a,\nu}(f) = 1$ for d_{L^∞} does not have a noticeable impact on clustering performance.

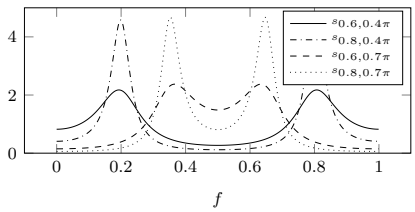

Fig. 2. Example PSDs of the form (9).

B. Real data

We perform experiments on two data sets, namely on human motion data and on EEG data.

Human motion data: We consider the problem of clustering sequences of human motion data according to the underlying activities performed. Specifically, we consider the experiment conducted in [14], [39], which uses the Carnegie Mellon Motion Capture database[3] containing motion sequences of 149 subjects performing various activities. The clustering algorithm in [39] first fits a linear dynamical system model to each motion sequence and then performs standard k-means clustering with the estimated model parameters (organized into vectors) as data points. In [14] an online clustering algorithm based on KM in combination with distributional distance is proposed. The motion vector-sequences in the Carnegie Mellon Motion Capture database describe the temporal evolution of marker positions on different body parts, recorded through optical tracking. The experiment in [14], [39] is based on subjects #16 and #35 for which the database contains 49 and 33 sequences, respectively, labeled either as "walking" or "running". We cluster the (scalar-valued) sequences describing the motion of the marker placed on the right foot of the subjects. It is argued in [14] that these sequences can be considered stationary ergodic. We assume the number of generative models $L = 2$ to be known and set $q = 5$ (good performance was observed for $4 \leq q \leq 10$). For the BT estimator, we use the Bartlett window g_W^B, defined in (10), with W given by the sequence length, and we normalize the BT PSD estimates to unit power. Table I lists the CE,

[3]available at http://mocap.cs.cmu.edu

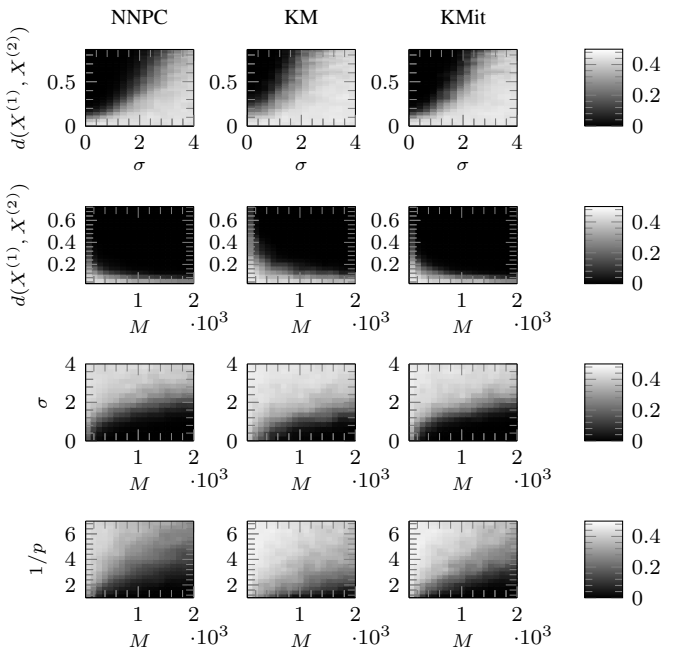

Fig. 3. Results of the synthetic data experiment. First row: CE as a function of σ and $d(X^{(1)}, X^{(2)})$ for $M = 400$ and $p = 1$. Second row: CE as a function of M and $d(X^{(1)}, X^{(2)})$ for $\sigma = 0.5$ and $p = 1$. Third row: CE as a function of M and σ for $\nu_2 = 0.62\pi$ ($d(X^{(1)}, X^{(2)}) \approx 0.2$) and $p = 1$. Bottom row: CE as a function of M and $1/p$ for $\nu_2 = 0.62\pi$ and $\sigma = 0.5$.

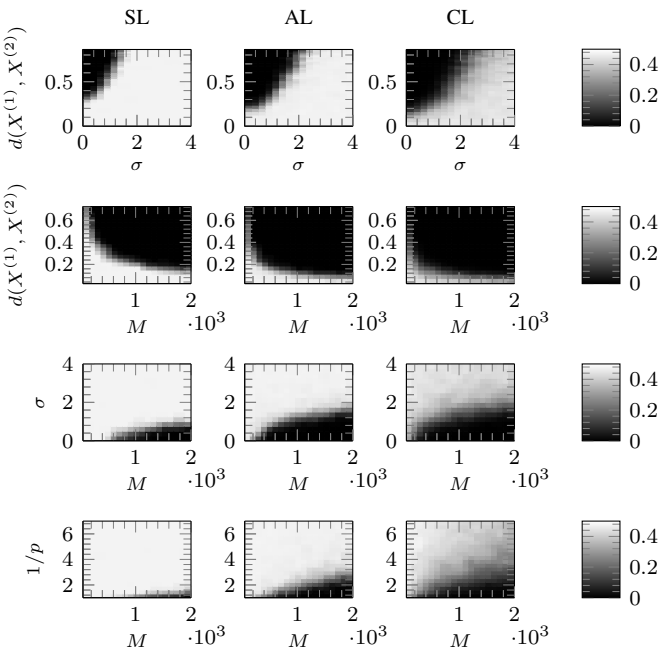

Fig. 5. CE for single linkage, average linkage, and complete linkage hierarchical clustering as a function of $d(X^{(1)}, X^{(2)})$, M, σ, and p, using the same values as in the setup in Fig. 3 for the model parameters that are not varied.

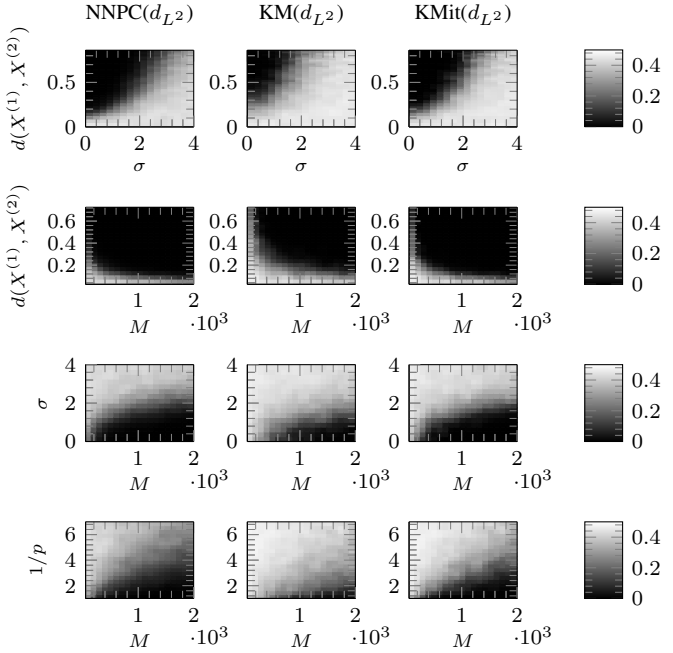

Fig. 4. CE as a function of $d(X^{(1)}, X^{(2)})$, M, σ, and p for variants of NNPC, KM, and KMit based on d_{L^2}, using the same values as in the setup in Fig. 3 for the model parameters that are not varied.

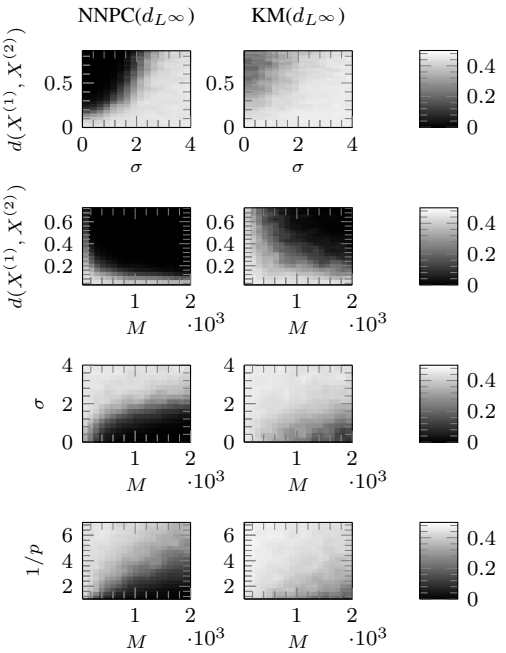

Fig. 6. CE as a function of $d(X^{(1)}, X^{(2)})$, M, σ, and p for variants of NNPC and KM based on d_{L^∞}, using the same values as in the setup in Fig. 3 for the model parameters that are not varied.

the running times in seconds, and for comparison with the results in [14], [39] also the entropy S of the clustering confusion matrix (see [39, Sec. 6] for the definition of S). This comparison reveals that for subject #35 NNPC, KM, and KMit all outperform the algorithm in [39] and match the performance of that in [14], while for subject #16 NNPC significantly outperforms both the algorithms in [14], [39] as well as KM and KMit.

EEG data: We perform an experiment similar to that in [40, Sec. 5], which considers clustering of segments of EEG recordings of healthy subjects and of subjects experi-

TABLE I
CE, S, AND RUNNING TIME t (IN SECONDS) FOR CLUSTERING OF HUMAN MOTION SEQUENCES

	NNPC			KM			KMit			[14]	[39]
subject	CE	S	t	CE	S	t	CE	S	t	S	S
#16	0.02	0.09	0.206	0.24	0.55	0.029	0.20	0.49	0.038	0.21	0.37
#35	0	0	0.185	0	0	0.017	0	0	0.024	0	0.10

encing epileptic seizure according to whether seizure activity is present or not. It is argued in [41] that EEG recordings can be modeled as stationary ergodic random processes. We use subsets A and E of the publicly available[4] EEG data set described in [42]. Each of these two subsets contains 100 EEG segments of 23.6s duration, acquired at a sampling rate of 173.61Hz. We refer to [42] for a more detailed description of acquisition and preprocessing aspects.

We compare the performance of NNPC, KM, and KMit as a function of W and q (for NNPC). We center each EEG segment by subtracting its (estimated) mean and use a Bartlett window g_W^B, as defined in (10), of variable length W for the BT PSD estimator. Furthermore, we normalize the PSD estimates to unit power and assume the number of clusters $L = 2$ to be known. Fig. 7 shows the CE obtained for NNPC as a function of the window length W and of q, as well as the CE obtained for KM and KMit as a function of W. It can be seen that NNPC is robust to small variations of q and W around the pair (q, W) corresponding to the minimum CE (marked by a white dot in Fig. 7). Similarly, KM and KMit yield a CE close to their respective minima for a large range of values for W. In Table II, we report the minimum CE achieved by each algorithm, along with the corresponding running times and CE-minimizing values for W and q (in the case of NNPC), all chosen based on results depicted in Fig. 7. The minimum CE obtained for NNPC is significantly lower than that corresponding to KM and KMit.

TABLE II
CLUSTERING EEG SEGMENTS: MINIMUM CE, RUNNING TIME t (IN SECONDS), AND CORRESPONDING PARAMETER CHOICES.

	min CE	t	W	q
NNPC	0.005	0.694	840	3
KM	0.360	0.482	640	-
KMit	0.095	0.954	520	-

APPENDIX A
PROOFS OF THEOREMS 1 AND 2

The central element in the proofs of Theorems 1 and 2 is the following result, proven in Appendix B.

[4]http://ntsa.upf.edu/downloads/andrzejak-rg-et-al-2001-indications-nonlinear-deterministic-and-finite-dimensional

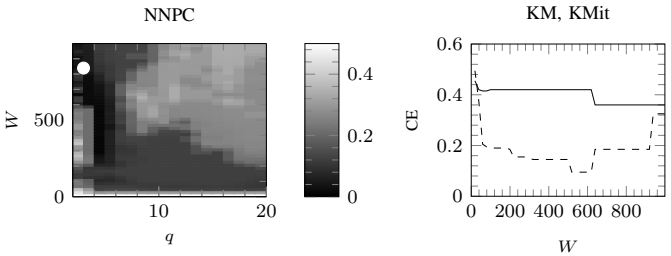

Fig. 7. Left: CE of NNPC for EEG recordings as a function of q and W. The white dot in the left figure shows the location of minimum CE. Right: CE of KM (solid line) and KMit (dashed line) as a function of W.

Theorem 3. *Consider a data set $\mathcal{X}$ generated according to the statistical data model described in Sec. III. Then, the clustering condition* (4) *implies that*

$$\min_{\substack{k,\ell\in[L]:\\ k\neq\ell}} \min_{\substack{i\in[n_\ell],\\ j\in[n_k]}} d(x_j^{(k)}, x_i^{(\ell)}) > \max_{\ell\in[L]} \max_{\substack{i,j\in[n_\ell]:\\ i\neq j}} d(x_i^{(\ell)}, x_j^{(\ell)}) \tag{11}$$

holds with probability at least $1 - 6N/M^2$.

Theorem 3 says that under the clustering condition (4) observations stemming from the same generative model are closer (in terms of the distance measure d) than observations originating from different generative models. This property is known in the clustering literature as the *strict separation property* [43]. We now show how Theorems 1 and 2 follow directly from the strict separation property.

Proof of Theorem 1: Under the condition $q \leq \min_{\ell\in[L]}(n_\ell - 1)$ the NFC property is a direct consequence of (11), which by Theorem 3, is implied by the clustering condition (4). The condition $q \leq \min_{\ell\in[L]}(n_\ell - 1)$ is necessary for the NFC property to hold as choosing $q > \min_{\ell\in[L]}(n_\ell - 1)$ would force NNPC to select observations from $\mathcal{X}\backslash\mathcal{X}_\ell$ for at least one of the data points $x_i^{(\ell)}$, thereby resulting in a violation of the NFC property.

Proof of Theorem 2: The proof is effected by first showing that in Step 3 KM selects an observation with a different underlying generative model in every iteration, i.e., the set of cluster centers $\{x_{c_\ell}\}_{\ell=1}^L$ contains exactly one observation from each generative model, provided that the clustering condition (4) and hence, by Theorem 3, (11) holds. The argument is then concluded by noting that (11) implies directly that the partition $\hat{\mathcal{X}}_1, \ldots, \hat{\mathcal{X}}_L$ obtained in Step 4 corresponds to the true partition $\mathcal{X}_1, \ldots, \mathcal{X}_L$.

It remains to establish that the cluster centers x_{c_ℓ} selected in Step 3 of KM, indeed, all originate from different generative models. This is accomplished by induction. For $v = 1$ the claim holds trivially, as we have selected a single cluster center only, namely x_{c_1}. The base case is hence established. For the inductive step, suppose that after the v-th iteration in Step 3 of KM the observations $\{x_{c_1}, \ldots, x_{c_v}\}$ all come from different generative models, and assume w.l.o.g. that the generative model underlying x_{c_ℓ} has index ℓ, $\ell \in [v]$. In iteration $v + 1$ (i.e., for the selection of $x_{c_{v+1}}$), we have

$$\max_{i\in[N]} \min_{\ell\in[v]} d(x_i, x_{c_\ell}^{(\ell)})$$

$$= \max\left\{ \max_{\substack{k\in[v],\\ i\in[n_k]}} \min_{\ell\in[v]} d(x_i^{(k)}, x_{c_\ell}^{(\ell)}),\ \max_{\substack{k\in[L]\backslash[v],\\ i\in[n_k]}} \min_{\ell\in[v]} d(x_i^{(k)}, x_{c_\ell}^{(\ell)}) \right\}$$

$$= \max\Bigg\{ \underbrace{\max_{\substack{k\in[v],\\ i\in[n_k]}} d(x_i^{(k)}, x_{c_k}^{(k)})}_{\leq \max_{\ell\in[L]} \max_{i,j\in[n_\ell]:\ i\neq j} d(x_i^{(\ell)}, x_j^{(\ell)})},\ \underbrace{\max_{\substack{k\in[L]\backslash[v],\\ i\in[n_k]}} \min_{\ell\in[v]} d(x_i^{(k)}, x_{c_\ell}^{(\ell)})}_{\geq \min_{\substack{k\in[L]\backslash[v],\\ i\in[n_k]}} \min_{\ell\in[v]} d(x_i^{(k)}, x_{c_\ell}^{(\ell)}) \geq \min_{k,\ell\in[L]:\ k\neq\ell} \min_{\substack{i\in[n_k],\\ j\in[n_\ell]}} d(x_i^{(k)}, x_j^{(\ell)})} \Bigg\} \tag{12}$$

$$= \max_{\substack{k\in[L]\backslash[v],\\ i\in[n_k]}} \min_{\ell\in[v]} d(x_i^{(k)}, x_{c_\ell}^{(\ell)}), \tag{13}$$

where we applied (11) to get (13) from (12). Note that in the maximization in (13) k runs over $[L]\backslash[v]$ (i.e., the maximization in (13) is over the observations in $\mathcal{X}\backslash(\mathcal{X}_1 \cup \cdots \cup \mathcal{X}_v)$), which implies that $x_{c_{v+1}}$ is guaranteed to correspond to a generative model that is different from those underlying $x_{c_1}, \ldots, x_{c_v}$. This completes the induction argument.

Appendix B
Proof of Theorem 3

We start by quantifying the deviation of the estimated distances $d(x_j^{(k)}, x_i^{(\ell)})$ from the true distances $d(X^{(k)}, X^{(\ell)})$ due to the PSD estimation error caused by finite observation length, noise, and missing entries.

Let $\tilde{s}^{(\ell)}(f) \coloneqq s^{(\ell)}(f) + \sigma^2$, $f \in [0,1)$, $\ell \in [L]$, be the PSD of the noisy observation $\tilde{X}^{(\ell)}$ and denote the corresponding ACF by $\tilde{r}^{(\ell)}$. With $\hat{s}_i^{(\ell)}(f)$ as defined in (1) (recall that we use the modified window $\hat{g}[m] = g[m]/u[m]$ in the BT estimator (1)), set $e_i^{(\ell)}(f) \coloneqq \hat{s}_i^{(\ell)}(f) - \tilde{s}^{(\ell)}(f)$, and let $\varepsilon \coloneqq \max_{\ell\in[L], i\in[n_\ell]} \sup_{f\in[0,1)} |e_i^{(\ell)}(f)|$. We have for all $k, \ell \in [L]$, $j \in [n_k]$, $i \in [n_\ell]$,

$$\begin{aligned}
d(x_j^{(k)}, x_i^{(\ell)}) &= \frac{1}{2}\int_0^1 \left|\hat{s}_j^{(k)}(f) - \hat{s}_i^{(\ell)}(f)\right| \mathrm{d}f \\
&= \frac{1}{2}\int_0^1 \left|s^{(k)}(f) + \sigma^2 + e_j^{(k)}(f) - (s^{(\ell)}(f) + \sigma^2 + e_i^{(\ell)}(f))\right| \mathrm{d}f \\
&\leq \frac{1}{2}\int_0^1 \left|s^{(k)}(f) - s^{(\ell)}(f)\right| \mathrm{d}f + \frac{1}{2}\int_0^1 \left|e_j^{(k)}(f) - e_i^{(\ell)}(f))\right| \mathrm{d}f \\
&\leq d(X^{(k)}, X^{(\ell)}) + \frac{1}{2}\int_0^1 \left|e_j^{(k)}(f)\right| \mathrm{d}f + \frac{1}{2}\int_0^1 \left|e_i^{(\ell)}(f)\right| \mathrm{d}f
\end{aligned} \tag{14}$$

$$\leq d(X^{(k)}, X^{(\ell)}) + \varepsilon. \tag{15}$$

Applying the reverse triangle inequality, it follows similarly that

$$d(x_j^{(k)}, x_i^{(\ell)}) \geq d(X^{(k)}, X^{(\ell)}) - \varepsilon, \tag{16}$$

for all $k, \ell \in [L]$, $j \in [n_k]$, $i \in [n_\ell]$. Replacing the RHS of (11) by the upper bound in (15) and the LHS by the lower bound in (16), we find that (11) is implied by

$$\min_{k,\ell\in[L]:\ k\neq\ell} d(X^{(k)}, X^{(\ell)}) > 2\varepsilon. \tag{17}$$

We continue by upper-bounding ε. To this end, define $\mathbf{Q}_m \in \{0,1\}^{M\times M}$ according to $(\mathbf{Q}_m)_{u,v} = 1$, if $v - u = m$, and $(\mathbf{Q}_m)_{u,v} = 0$, else, and let $\hat{\mathbf{G}}(f) \coloneqq \sum_{m\in\mathcal{M}} \hat{g}[m]\cos(2\pi f m)\mathbf{Q}_m$. Now, with $\mathbf{x} \in \mathbb{R}^M$ the random vector whose elements are given by $x_i^{(\ell)}$, it holds for $m \in \mathcal{M} = \{-M+1, -M+2, \ldots, M-1\}$ that

$$\hat{r}_i^{(\ell)}[m] = \frac{\mathbf{x}^\top \mathbf{Q}_m \mathbf{x}}{M} = \frac{\mathbf{y}^\top \mathbf{C}^\top \mathbf{P}_{\boldsymbol{\xi}}^\top \mathbf{Q}_m \mathbf{P}_{\boldsymbol{\xi}} \mathbf{C}\mathbf{y}}{M},$$

where we used $\mathbf{x} = \mathbf{P}_{\boldsymbol{\xi}}\mathbf{C}\mathbf{y}$, with the entries of $\mathbf{y}$ i.i.d. standard normal, $\mathbf{C} = (\mathbf{R} + \sigma^2\mathbf{I})^{1/2} \in \mathbb{R}^{M\times M}$ with $\mathbf{R}_{v,w} = r^{(\ell)}[w-v]$ the (Toeplitz) covariance matrix corresponding to M consecutive elements of $\tilde{X}^{(\ell)}$, and $\boldsymbol{\xi} \in \{0,1\}^M$ indicates the locations of the observed entries of $\mathbf{x}$. Note that $\mathbf{R}$ is identical for all contiguous length-M segments of $\tilde{X}^{(\ell)}$ thanks to stationarity, and $\mathbf{C}$ is symmetric because $\mathbf{R} + \sigma^2\mathbf{I}$ is symmetric. We next develop an upper bound on ε according to

$$\begin{aligned}
&\sup_{f\in[0,1)} \left|e_i^{(\ell)}(f)\right| = \sup_{f\in[0,1)} \left|\hat{s}_i^{(\ell)}(f) - \tilde{s}^{(\ell)}(f)\right| \\
&= \sup_{f\in[0,1)} \left| \sum_{m\in\mathcal{M}} \hat{g}[m]\hat{r}_i^{(\ell)}[m] e^{-\mathrm{i}2\pi f m} - \sum_{m\in\mathbb{Z}} \tilde{r}^{(\ell)}[m] e^{-\mathrm{i}2\pi f m} \right| \\
&= \sup_{f\in[0,1)} \Bigg| \underbrace{\sum_{m\in\mathcal{M}} \frac{\hat{g}[m]}{M} \left(\mathbf{x}^\top \mathbf{Q}_m \mathbf{x} - \mathbb{E}_{\mathbf{y}}\left[\mathbf{x}^\top \mathbf{Q}_m \mathbf{x}\right]\right) e^{-\mathrm{i}2\pi f m}}_{\frac{1}{M}\left(\mathbf{x}^\top\left(\sum_{m\in\mathcal{M}} \hat{g}[m]\cos(2\pi f m)\mathbf{Q}_m\right)\mathbf{x} - \mathbb{E}_{\mathbf{y}}\left[\mathbf{x}^\top\left(\sum_{m\in\mathcal{M}} \hat{g}[m]\cos(2\pi f m)\mathbf{Q}_m\right)\mathbf{x}\right]\right)} \\
&\quad + \sum_{m\in\mathcal{M}} \frac{\hat{g}[m]}{M} \mathbb{E}_{\mathbf{y}}\left[\mathbf{x}^\top \mathbf{Q}_m \mathbf{x}\right] e^{-\mathrm{i}2\pi f m} - \sum_{m\in\mathbb{Z}} \tilde{r}^{(\ell)}[m] e^{-\mathrm{i}2\pi f m} \Bigg| \\
&\leq \sup_{f\in[0,1)} \Bigg| \underbrace{\frac{1}{M}\left(\mathbf{y}^\top \mathbf{C}^\top \mathbf{P}_{\boldsymbol{\xi}}^\top \hat{\mathbf{G}}(f) \mathbf{P}_{\boldsymbol{\xi}} \mathbf{C}\mathbf{y} - \mathbb{E}_{\mathbf{y}}\left[\mathbf{y}^\top \mathbf{C}^\top \mathbf{P}_{\boldsymbol{\xi}}^\top \hat{\mathbf{G}}(f) \mathbf{P}_{\boldsymbol{\xi}} \mathbf{C}\mathbf{y}\right]\right)}_{=:\alpha_i^{(\ell)}(f)} \Bigg| \\
&\quad + \Bigg| \underbrace{\sum_{m\in\mathcal{M}} \frac{\hat{g}[m]}{M} \mathbb{E}_{\mathbf{y}}\left[\mathbf{x}^\top \mathbf{Q}_m \mathbf{x}\right] - \sum_{m\in\mathbb{Z}} \tilde{r}^{(\ell)}[m]}_{=:\beta_i^{(\ell)}} \Bigg|,
\end{aligned} \tag{18}$$

where we used the fact that $\hat{g}[m](\mathbf{x}^\top \mathbf{Q}_m \mathbf{x} - \mathbb{E}_{\mathbf{y}}\left[\mathbf{x}^\top \mathbf{Q}_m \mathbf{x}\right])$ is a real-valued even sequence ($\hat{g}[m]$ and $\mathbf{x}^\top \mathbf{Q}_m \mathbf{x} = M\hat{r}_i^{(\ell)}[m]$ are real-valued even by definition, the latter property implies that $\mathbb{E}_{\mathbf{y}}\left[\mathbf{x}^\top \mathbf{Q}_m \mathbf{x}\right]$ is also real-valued and even). It now follows from (18) that

$$\varepsilon \leq \max_{\ell\in[L], i\in[n_\ell]} \left(\sup_{f\in[0,1)} \left|\alpha_i^{(\ell)}(f)\right| + \left|\beta_i^{(\ell)}\right| \right)$$

and hence (4) implies (11) via (17) on the event $\mathcal{F}^\star \coloneqq \bigcap_{\ell\in[L], i\in[n_\ell]} (\mathcal{F}_{1,i}^{(\ell)} \cap \mathcal{F}_{2,i}^{(\ell)})$ with

$$\mathcal{F}_{1,i}^{(\ell)} \coloneqq \left\{ \sup_{f\in[0,1)} \left|\alpha_i^{(\ell)}(f)\right| < \frac{4A(B+\sigma^2)}{p^2}\sqrt{\frac{2\log M}{M}} \right\} \text{ and}$$

$$\mathcal{F}_{2,i}^{(\ell)} \coloneqq \left\{ \left|\beta_i^{(\ell)}\right| < 8(1+p)\frac{A(1+\sigma^2)}{p^2}\sqrt{\frac{\log M}{M}} + \mu_{\max} \right\}.$$

With the upper bound on $\mathrm{P}[\bar{\mathcal{F}}_{1,i}^{(\ell)}]$ resulting from (22) and that on $\mathrm{P}[\bar{\mathcal{F}}_{2,i}^{(\ell)}]$ in (32), application of the union bound according to

$$\mathrm{P}[\mathcal{F}^\star] \geq 1 - \sum_{\ell\in[L],i\in[n_\ell]} \left(\mathrm{P}\left[\bar{\mathcal{F}}_{1,i}^{(\ell)}\right] + \mathrm{P}\left[\bar{\mathcal{F}}_{2,i}^{(\ell)}\right]\right) \geq 1 - \frac{6N}{M^2} \tag{19}$$

completes the proof.

We proceed to the upper bound on $\mathrm{P}[\bar{\mathcal{F}}_{1,i}^{(\ell)}]$.

Upper bound on $\mathrm{P}[\bar{\mathcal{F}}_{1,i}^{(\ell)}]$: Conditioning on $\boldsymbol{\xi}$ and setting $\mathbf{B} \coloneqq \mathbf{C}^\top \mathbf{P}_{\boldsymbol{\xi}}^\top \hat{\mathbf{G}}(f) \mathbf{P}_{\boldsymbol{\xi}} \mathbf{C}$, we establish an upper bound on the tail probability of $\sup_{f\in[0,1)} \left|\alpha_i^{(\ell)}(f)\right|$ by invoking a well-known concentration of measure result for quadratic forms in Gaussian random vectors [35, Lem. 1], namely

$$\mathrm{P}\Big[\left|\mathbf{y}^\top \mathbf{B}\mathbf{y} - \mathbb{E}\big[\mathbf{y}^\top \mathbf{B}\mathbf{y}\big]\right| \geq \left\|\mathbf{B}+\mathbf{B}^\top\right\|_F \sqrt{\delta} + 2\|\mathbf{B}\|_{2\to 2}\delta \Big| \boldsymbol{\xi}\Big] \leq 2e^{-\delta}. \tag{20}$$

Next, we note that $\|\mathbf{B}+\mathbf{B}^\top\|_F \leq 2\|\mathbf{B}\|_F \leq 2\sqrt{M}\|\mathbf{B}\|_{2\to 2}$ and

$$\begin{aligned}\|\mathbf{B}\|_{2\to 2} &\leq \| \underbrace{\mathbf{C}}_{=\mathbf{R}+\sigma^2\mathbf{I}} \|_{2\to 2}^2 \underbrace{\|\mathbf{P}_{\boldsymbol{\xi}}\|_{2\to 2}^2}_{\leq 1} \|\hat{\mathbf{G}}(f)\|_{2\to 2} \\ &\leq \frac{A(B+\sigma^2)}{p^2},\end{aligned}$$

where the second inequality follows as both $\mathbf{R}$ and $\hat{\mathbf{G}}(f)$ are symmetric Toeplitz matrices and hence, by [44, Lem. 4.1], $\|\mathbf{R}\|_{2\to 2} \leq \sup_{f\in[0,1)} s^{(\ell)}(f) \leq B$ and

$$\begin{aligned}\|\hat{\mathbf{G}}(f)\|_{2\to 2} &\leq \sup_{f'\in[0,1)} \hat{g}(f') \\ &= \sup_{f'\in[0,1)} \frac{1}{p^2} g(f') + \underbrace{\left(\frac{1}{p} - \frac{1}{p^2}\right)}_{\leq 0} \underbrace{g[0]}_{=1} \\ &\leq \frac{1}{p^2} \sup_{f'\in[0,1)} g(f') = \frac{A}{p^2},\end{aligned} \tag{21}$$

where we used $\hat{g}[m] = (1/p^2)g[m] + (1/p - 1/p^2)g[0]\delta[m]$. Now, setting $\delta = 2\log(M)$ in (20) and using $\delta/M \leq \sqrt{\delta/M} < 1$, for $M \geq 1$, yields

$$\mathrm{P}\Big[\bar{\mathcal{F}}_{1,i}^{(\ell)} \Big| \boldsymbol{\xi}\Big] = \mathrm{P}\left[\sup_{f\in[0,1)} \left|\alpha_i^{(\ell)}(f)\right| \geq \frac{4A(B+\sigma^2)}{p^2}\sqrt{\frac{2\log M}{M}} \Big| \boldsymbol{\xi}\right] \leq \frac{2}{M^2}. \tag{22}$$

The proof is concluded by noting that this bound holds uniformly over $\boldsymbol{\xi} \in \{0,1\}^M$ so that $\mathrm{P}[\bar{\mathcal{F}}_{1,i}^{(\ell)}] \leq 2/M^2$.

Upper bound on $\mathrm{P}[\bar{\mathcal{F}}_{2,i}^{(\ell)}]$: Setting $\tilde{\mathbf{G}} \coloneqq \sum_{m\in\mathcal{M}} \hat{g}[m]\mathbf{Q}_m$, we start by rewriting the first sum in the definition of $\beta_i^{(\ell)}$ in (18) as

$$\begin{aligned}\sum_{m\in\mathcal{M}} \frac{\hat{g}[m]}{M} \mathbb{E}_{\mathbf{y}}\big[\mathbf{x}^\top \mathbf{Q}_m \mathbf{x}\big] &= \frac{1}{M}\mathbb{E}_{\mathbf{y}}\left[\mathbf{x}^\top \Big(\sum_{m\in\mathcal{M}} \hat{g}[m]\mathbf{Q}_m\Big)\mathbf{x}\right] \\ &= \frac{1}{M}\mathbb{E}_{\mathbf{y}}\Big[\mathbf{y}^\top \mathbf{C}^\top \mathbf{P}_{\boldsymbol{\xi}}^\top \tilde{\mathbf{G}} \mathbf{P}_{\boldsymbol{\xi}} \mathbf{C}\mathbf{y}\Big] \\ &= \frac{1}{M}\operatorname{tr}(\mathbf{C}^\top \mathbf{P}_{\boldsymbol{\xi}}^\top \tilde{\mathbf{G}} \mathbf{P}_{\boldsymbol{\xi}} \mathbf{C}) \\ &= \frac{1}{M}\operatorname{tr}(\mathbf{P}_{\boldsymbol{\xi}}^\top \tilde{\mathbf{G}} \mathbf{P}_{\boldsymbol{\xi}} \underbrace{\mathbf{C}\mathbf{C}^\top}_{=\mathbf{R}+\sigma^2\mathbf{I}}) \qquad (23) \\ &= \frac{1}{M}\sum_{u,v\in[M]} \xi_u \xi_v \tilde{\mathbf{G}}_{u,v} (\underbrace{\mathbf{R}_{v,u}}_{=\mathbf{R}_{u,v}} + \sigma^2\delta[u-v]) \qquad (24) \\ &= \frac{1}{M}\boldsymbol{\xi}^\top (\tilde{\mathbf{G}} \circ (\mathbf{R}+\sigma^2\mathbf{I}))\boldsymbol{\xi}. \qquad (25)\end{aligned}$$

Now, setting $\mathbf{D} \coloneqq \tilde{\mathbf{G}} \circ (\mathbf{R}+\sigma^2\mathbf{I})$ and using (25), we have

$$\begin{aligned}&\left|\beta_i^{(\ell)}\right| \\ &= \left|\frac{1}{M}\boldsymbol{\xi}^\top \mathbf{D}\boldsymbol{\xi} - \frac{1}{M}\mathbb{E}\Big[\boldsymbol{\xi}^\top \mathbf{D}\boldsymbol{\xi}\Big] + \frac{1}{M}\mathbb{E}\Big[\boldsymbol{\xi}^\top \mathbf{D}\boldsymbol{\xi}\Big] - \sum_{m\in\mathbb{Z}} \tilde{r}^{(\ell)}[m]\right| \\ &\leq \left|\frac{1}{M}\Big(\boldsymbol{\xi}^\top \mathbf{D}\boldsymbol{\xi} - \mathbb{E}\Big[\boldsymbol{\xi}^\top \mathbf{D}\boldsymbol{\xi}\Big]\Big)\right| + \left|\frac{1}{M}\mathbb{E}\Big[\boldsymbol{\xi}^\top \mathbf{D}\boldsymbol{\xi}\Big] - \sum_{m\in\mathbb{Z}} \tilde{r}^{(\ell)}[m]\right| \qquad (26) \\ &\leq \left|\underbrace{\frac{1}{M}\Big(\boldsymbol{\xi}^\top \mathbf{D}\boldsymbol{\xi} - \mathbb{E}\Big[\boldsymbol{\xi}^\top \mathbf{D}\boldsymbol{\xi}\Big]\Big)}_{=:\gamma_i^{(\ell)}}\right| + \mu_{\max}. \qquad (27)\end{aligned}$$

Here, the last inequality is a consequence of the following upper bound on the second term in (26)

$$\begin{aligned}&\left|\frac{1}{M}\mathbb{E}\Big[\boldsymbol{\xi}^\top \mathbf{D}\boldsymbol{\xi}\Big] - \sum_{m\in\mathbb{Z}} \tilde{r}^{(\ell)}[m]\right| \\ &= \left|\frac{1}{M}\sum_{v,w\in[M]} \mathbb{E}[\xi_v\xi_w]\,\tilde{\mathbf{G}}_{v,w}(\mathbf{R}_{v,w} + \sigma^2\delta[v-w]) - \sum_{m\in\mathbb{Z}} \tilde{r}^{(\ell)}[m]\right| \qquad (28) \\ &= \left|\frac{1}{M}\sum_{v,w\in[M]} \underbrace{\mathbb{E}[\xi_v\xi_w]\,\hat{g}[v-w]}_{u[v-w]\hat{g}[v-w]=g[v-w]}\; \tilde{r}^{(\ell)}[v-w] - \sum_{m\in\mathbb{Z}} \tilde{r}^{(\ell)}[m]\right| \\ &= \left|\underbrace{g[0]}_{=1}(r^{(\ell)}[0]+\sigma^2) - (r^{(\ell)}[0]+\sigma^2) + \sum_{m\in\mathcal{M}\setminus\{0\}} \left(\frac{M-|m|}{M} g[m] r^{(\ell)}[m] - r^{(\ell)}[m]\right) - \sum_{m\in\mathbb{Z}\setminus\mathcal{M}} r^{(\ell)}[m]\right| \qquad (29) \\ &\leq \sum_{m\in\mathbb{Z}} |h[m]||r^{(\ell)}[m]| \\ &\leq \mu_{\max}, \qquad (30)\end{aligned}$$

where (28) follows from the equality (25)=(24) and from $\tilde{r}^{(\ell)}[m] = r^{(\ell)}[m] + \sigma^2\delta[m]$. We continue by establishing a

bound on the tail probability of $|\gamma_i^{(\ell)}|$. To this end, we note that

$$\begin{aligned}\|\mathbf{D}\|_{2\to 2} &= \|\tilde{\mathbf{G}} \circ (\mathbf{R}+\sigma^2\mathbf{I})\|_{2\to 2} \\ &= (1+\sigma^2)\left\|\tilde{\mathbf{G}} \circ \left(\frac{\mathbf{R}+\sigma^2\mathbf{I}}{1+\sigma^2}\right)\right\|_{2\to 2} \\ &\leq (1+\sigma^2)\|\tilde{\mathbf{G}}\|_{2\to 2} \\ &\leq \frac{A(1+\sigma^2)}{p^2}, \end{aligned} \tag{31}$$

where we used the fact that $(\mathbf{R}+\sigma^2\mathbf{I})/(1+\sigma^2)$ is a symmetric positive semi-definite matrix with ones on its main diagonal, and we employed [45, Thm. 5.5.11] in the first inequality, and steps analogous to those in (21) to obtain the second inequality.

Now, using (27) we get

$$\begin{aligned}\mathrm{P}\left[\bar{\mathcal{F}}_{2,i}^{(\ell)}\right] &\leq \mathrm{P}\left[\left|\gamma_i^{(\ell)}\right| \geq 8(1+p)\frac{A(1+\sigma^2)}{p^2}\sqrt{\frac{\log M}{M}}\right] \\ &< \mathrm{P}\left[\left|\gamma_i^{(\ell)}\right| > 8(1+p)\|\mathbf{D}\|_{2\to 2}\sqrt{\frac{\log M}{M}}\right] < \frac{4}{M^2}, \end{aligned} \tag{32}$$

where the second inequality follows from the upper bound on $\|\mathbf{D}\|_{2\to 2}$ in (31) and the third inequality is an application of Lemma 1 with $\mathbf{H} \coloneqq \mathbf{D}$ and $t \coloneqq 8(1+p)\|\mathbf{D}\|_{2\to 2}\sqrt{M\log M}$.

A final remark concerns the concentration inequality for quadratic forms in Boolean random vectors reported in the following Lemma 1. Such concentration inequalities, or more generally, concentration inequalities for multivariate polynomials of Boolean random variables have been studied extensively in the context of random graph theory [46]. Unfortunately, the bounds available in the literature typically come in terms of functions of the entries of $\mathbf{H}$ that do not lead to crisp statements in the context of the process clustering problem considered here. We therefore develop a new concentration result in Lemma 1, which depends on $\|\mathbf{H}\|_{2\to 2}$ only. The proof of this result is based on techniques developed in [37].

Lemma 1. *Let $\mathbf{H} \in \mathbb{R}^{M\times M}$ be a (deterministic) symmetric matrix and let $\boldsymbol{\xi} \in \{0,1\}^M$ be a random vector with i.i.d. Bernoulli entries drawn according to $\mathrm{P}[\xi_i = 1] = 1 - \mathrm{P}[\xi_i = 0] = p$, $i \in [M]$. Then, we have*

$$\begin{aligned}&\mathrm{P}\left[\left|\boldsymbol{\xi}^\top\mathbf{H}\boldsymbol{\xi} - \mathbb{E}\left[\boldsymbol{\xi}^\top\mathbf{H}\boldsymbol{\xi}\right]\right| > t\right] \\ &\qquad < 4\exp\left(-\frac{t^2}{32(1+p)^2M\|\mathbf{H}\|_{2\to 2}^2}\right). \end{aligned} \tag{33}$$

Proof. The proof is effected by adapting the proof of [37, Thm. 1.1], which provides a concentration inequality for quadratic forms in zero-mean subgaussian random vectors. We start by decomposing $\boldsymbol{\xi}^\top\mathbf{H}\boldsymbol{\xi} - \mathbb{E}\left[\boldsymbol{\xi}^\top\mathbf{H}\boldsymbol{\xi}\right]$ according to

$$\begin{aligned}&\boldsymbol{\xi}^\top\mathbf{H}\boldsymbol{\xi} - \mathbb{E}\left[\boldsymbol{\xi}^\top\mathbf{H}\boldsymbol{\xi}\right] \\ &\quad= \sum_{i\in[M]} \mathbf{H}_{i,i}(\xi_i^2 - \mathbb{E}[\xi_i^2]) + \sum_{\substack{i,j\in[M]:\\ i\neq j}} \mathbf{H}_{i,j}(\xi_i\xi_j - \mathbb{E}[\xi_i\xi_j]) \\ &\quad= \underbrace{\sum_{i\in[M]} \mathbf{H}_{i,i}(\xi_i - p)}_{=:S_{\mathrm{diag}}} + \underbrace{\sum_{\substack{i,j\in[M]:\\ i\neq j}} \mathbf{H}_{i,j}(\xi_i\xi_j - p^2)}_{=:S_{\mathrm{off}}}, \end{aligned}$$

where we used the fact that the ξ_i, $i \in [M]$, are $\{0,1\}$-valued and statistically independent. Now, we have

$$\begin{aligned}\mathrm{P}\left[\left|\boldsymbol{\xi}^\top\mathbf{H}\boldsymbol{\xi} - \mathbb{E}\left[\boldsymbol{\xi}^\top\mathbf{H}\boldsymbol{\xi}\right]\right| > t\right] &\leq \mathrm{P}[|S_{\mathrm{diag}}| + |S_{\mathrm{off}}| > t] \\ &\leq \mathrm{P}[|S_{\mathrm{diag}}| > t/2] + \mathrm{P}[|S_{\mathrm{off}}| > t/2] \qquad (34) \\ &\leq 2\exp\left(-\frac{2t^2}{\sum_{i\in[M]}(\mathbf{H}_{i,i})^2}\right) \\ &\qquad + 2\exp\left(-\frac{t^2}{32(1+p)^2M\|\mathbf{H}\|_{2\to 2}^2}\right) \qquad (35) \\ &< 4\exp\left(-\frac{t^2}{32(1+p)^2M\|\mathbf{H}\|_{2\to 2}^2}\right), \end{aligned}$$

where (35) follows from the upper bounds on $\mathrm{P}[|S_{\mathrm{diag}}| > t/2]$ and $\mathrm{P}[|S_{\mathrm{off}}| > t/2]$ established below, and the last inequality is thanks to $\sum_{i\in[M]}(\mathbf{H}_{i,i})^2 \leq M\max_{i\in[M]}(\mathbf{H}_{i,i})^2 \leq M\|\mathbf{H}\|_{2\to 2}^2$ obtained from $\|\mathbf{H}\|_{2\to 2}^2 = \max_{\|\mathbf{x}\|_2=1}\|\mathbf{H}\mathbf{x}\|_2^2 \geq \max_{\|\mathbf{x}\|_2=1,\mathbf{x}\in\{0,1\}^M}\|\mathbf{H}\mathbf{x}\|_2^2 = \max_{i\in[M]}\sum_{j\in[M]}(\mathbf{H}_{j,i})^2 \geq \max_{i\in[M]}(\mathbf{H}_{i,i})^2$.

Upper bound on $\mathrm{P}[|S_{\mathrm{diag}}| > t/2]$: Note that the $\mathbf{H}_{i,i}(\xi_i - p)$, $i \in [M]$, are independent, bounded, zero-mean random variables with $a_i \leq \mathbf{H}_{i,i}(\xi_i - p) \leq b_i$, $a_i, b_i \in \mathbb{R}$, $i \in [M]$. We can therefore apply Hoeffding's inequality [47, Thm. 2.8], which upon noting that $(b_i - a_i)^2 = \mathbf{H}_{i,i}^2$ yields

$$\mathrm{P}[|S_{\mathrm{diag}}| > t/2] < 2\exp\left(-\frac{2t^2}{\sum_{i\in[M]}(\mathbf{H}_{i,i})^2}\right).$$

Upper bound on $\mathrm{P}[|S_{\mathrm{off}}| > t/2]$: We start by decoupling [48, Sec. 8.4] the sum S_{off} over the off-diagonal entries of $\mathbf{H}$, then upper-bound the moment generating function of S_{off}, and use the resulting upper bound to get an upper bound on $\mathrm{P}[S_{\mathrm{off}} > t/2]$ via the exponential Chebyshev inequality. The final result follows by noting that $\mathrm{P}[S_{\mathrm{off}} > t/2] = \mathrm{P}[S_{\mathrm{off}} < -t/2]$ and applying the union bound.

To decouple S_{off}, consider i.i.d. Bernoulli random variables $\nu_i \in \{0,1\}$, $i \in [M]$, with $\mathrm{P}[\nu_i = 0] = \mathrm{P}[\nu_i = 1] = 1/2$, and set $\boldsymbol{\nu} = [\nu_1 \ \ldots \ \nu_M]^\top$. With

$$S_\nu \coloneqq \sum_{i,j\in[M]} \nu_i(1-\nu_j)\mathbf{H}_{i,j}(\xi_i - p)(\xi_j + p),$$

we have $S_{\mathrm{off}} = 4\mathbb{E}_{\boldsymbol{\nu}}[S_\nu]$ thanks to the symmetry of $\mathbf{H}$ (i.e., $\mathbf{H}_{i,j} = \mathbf{H}_{j,i}$), and $\mathbb{E}[\nu_i(1-\nu_j)] = 1/4$, for $i \neq j$, and $\mathbb{E}[\nu_i(1-\nu_j)] = 0$, for $i = j$. Setting $\mathcal{I}_\nu \coloneqq \{i \in [M] \colon \nu_i = 1\}$, we can express S_ν as

$$\begin{aligned}S_\nu &= \sum_{i\in\mathcal{I}_\nu, j\in\overline{\mathcal{I}_\nu}} \mathbf{H}_{i,j}(\xi_i - p)(\xi_j + p) \\ &= \sum_{i\in\mathcal{I}_\nu}(\xi_i - p)\Bigg(\sum_{j\in\overline{\mathcal{I}_\nu}} \mathbf{H}_{i,j}(\xi_j + p)\Bigg). \end{aligned} \tag{36}$$

We continue by upper-bounding the moment generating function of S_{off} via Jensen's inequality according to

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\xi}}[\exp(\lambda S_{\text{off}})] &= \mathbb{E}_{\boldsymbol{\xi}}[\exp(\lambda 4\mathbb{E}_{\boldsymbol{\nu}}[S_\nu])] \\ &\leq \mathbb{E}_{\boldsymbol{\xi},\boldsymbol{\nu}}[\exp(4\lambda S_\nu)], \end{aligned} \tag{37}$$

where $\lambda > 0$ is a deterministic parameter. It follows from (36) that S_ν, conditioned on $\boldsymbol{\nu}$ and on the ξ_j, with $j \in \overline{\mathcal{I}_\nu}$, is a linear combination of independent bounded zero-mean random variables. We therefore have

$$\begin{aligned}&\mathbb{E}_{\xi_i, i\in\mathcal{I}_\nu}[\exp(4\lambda S_\nu)] \\ &= \mathbb{E}_{\xi_i, i\in\mathcal{I}_\nu}\left[\exp\left(4\lambda\sum_{i\in\mathcal{I}_\nu}(\xi_i - p)\Big(\sum_{j\in\overline{\mathcal{I}_\nu}}\mathbf{H}_{i,j}(\xi_j+p)\Big)\right)\right] \\ &= \prod_{i\in\mathcal{I}_\nu}\mathbb{E}_{\xi_i}\left[\exp\left(4\lambda(\xi_i - p)\Big(\sum_{j\in\overline{\mathcal{I}_\nu}}\mathbf{H}_{i,j}(\xi_j+p)\Big)\right)\right] \qquad (38) \\ &\leq \prod_{i\in\mathcal{I}_\nu}\exp\left(2\lambda^2\Big(\sum_{j\in\overline{\mathcal{I}_\nu}}\mathbf{H}_{i,j}(\xi_j+p)\Big)^2\right) \qquad (39) \\ &= \exp\left(2\lambda^2\sum_{i\in\mathcal{I}_\nu}\Big(\underbrace{\sum_{j\in\overline{\mathcal{I}_\nu}}\mathbf{H}_{i,j}(\xi_j+p)}_{=\mathbf{H}_i(\mathbf{I}-\mathbf{P}_{\boldsymbol{\nu}})(\boldsymbol{\xi}+p\mathbf{1})}\Big)^2\right) \\ &= \exp\left(2\lambda^2\|\mathbf{P}_{\boldsymbol{\nu}}\mathbf{H}(\mathbf{I}-\mathbf{P}_{\boldsymbol{\nu}})(\boldsymbol{\xi}+p\mathbf{1})\|_2^2\right) \\ &\leq \exp\left(2\lambda^2\underbrace{\|\mathbf{P}_{\boldsymbol{\nu}}\|_{2\to2}^2}_{\leq 1}\|\mathbf{H}\|_{2\to2}^2\underbrace{\|\mathbf{I}-\mathbf{P}_{\boldsymbol{\nu}}\|_{2\to2}^2}_{\leq 1}\underbrace{\|\boldsymbol{\xi}+p\mathbf{1}\|_2^2}_{\leq M(1+p)^2}\right) \\ &\leq \exp\left(2\lambda^2(1+p)^2M\|\mathbf{H}\|_{2\to2}^2\right), \qquad (40)\end{aligned}$$

where we used the independence of the ξ_i, $i \in \mathcal{I}_\nu$, to get (38), and Hoeffding's Lemma in the step leading from (38) to (39). Note that instead of Hoeffding's Lemma we could also apply [49, Thm. 2.1] to get a sharper bound on (38), but this would not lead to a different scaling behavior of (33) in terms of p or M.

Combining (40) with (37) and noting that the bound (40) does not depend on $\boldsymbol{\nu}$ and ξ_j, $j \in \overline{\mathcal{I}_\nu}$, it follows that

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\xi},\boldsymbol{\nu}}[\exp(\lambda S_{\text{off}})] &\leq \mathbb{E}_{\boldsymbol{\xi},\boldsymbol{\nu}}[\exp(4\lambda S_\nu)] \\ &= \mathbb{E}_{\boldsymbol{\nu}}\Big[\mathbb{E}_{\xi_j, j\in\overline{\mathcal{I}_\nu}}[\mathbb{E}_{\xi_i, i\in\mathcal{I}_\nu}[\exp(4\lambda S_\nu)]]\Big] \\ &\leq \mathbb{E}_{\boldsymbol{\nu}}\Big[\mathbb{E}_{\xi_j, j\in\overline{\mathcal{I}_\nu}}\Big[\exp\left(2\lambda^2(1+p)^2M\|\mathbf{H}\|_{2\to2}^2\right)\Big]\Big] \\ &= \exp\left(2\lambda^2(1+p)^2M\|\mathbf{H}\|_{2\to2}^2\right). \end{aligned} \tag{41}$$

We finally use (41) and the exponential Chebyshev inequality to get the upper bound

$$\mathrm{P}[S_{\text{off}} > t/2] \leq \exp\left(-\lambda t/2 + 2\lambda^2(1+p)^2M\|\mathbf{H}\|_{2\to2}^2\right), \tag{42}$$

which holds for all $\lambda > 0$. Minimizing (42) over $\lambda > 0$ yields

$$\mathrm{P}[S_{\text{off}} > t/2] \leq \exp\left(-\frac{t^2}{32(1+p)^2M\|\mathbf{H}\|_{2\to2}^2}\right). \tag{43}$$

□

Appendix C
Proof of Proposition 1

Recall that $\mathbf{x}_i^{(\ell)} = \mathbf{C}^{(\ell)}\mathbf{y}_i^{(\ell)}$, $\ell \in [L]$, $i \in [n_\ell]$, where $\mathbf{y}_i^{(\ell)}$ is an i.i.d. standard normal random vector and $\mathbf{C}^{(\ell)} \coloneqq (\tilde{\mathbf{R}}^{(\ell)})^{1/2}$. Setting $\sigma^{(k,\ell)} \coloneqq \|\mathbf{C}^{(k)^\top}\mathbf{C}^{(\ell)}\mathbf{y}_i^{(\ell)}\|_2$, conditional on $\mathbf{y}_i^{(\ell)}$, $\langle \mathbf{x}_j^{(k)}, \mathbf{x}_i^{(\ell)}\rangle$ and $\langle \mathbf{x}_v^{(\ell)}, \mathbf{x}_i^{(\ell)}\rangle$, for $k \neq \ell$ and $v \neq i$, are independent (as a consequence of the mutual independence of the $\mathbf{x}_i^{(\ell)}$, $\ell \in [L]$, $i \in [n_\ell]$, which is by assumption) and distributed according to $\mathcal{N}(0, {\sigma^{(k,\ell)}}^2)$ and $\mathcal{N}(0, {\sigma^{(\ell,\ell)}}^2)$, respectively. Conditional on $\mathbf{y}_i^{(\ell)}$, or equivalently, conditional on $\sigma^{(k,\ell)}$ and $\sigma^{(\ell,\ell)}$, $|\langle \mathbf{x}_j^{(k)}, \mathbf{x}_i^{(\ell)}\rangle|$ and $|\langle \mathbf{x}_v^{(\ell)}, \mathbf{x}_i^{(\ell)}\rangle|$ hence have half-normal distributions and we get

$$\begin{aligned}&\mathrm{P}\left[\left|\left\langle \mathbf{x}_j^{(k)}, \mathbf{x}_i^{(\ell)}\right\rangle\right| < \left|\left\langle \mathbf{x}_v^{(\ell)}, \mathbf{x}_i^{(\ell)}\right\rangle\right| \middle| \frac{\sigma^{(k,\ell)}}{\sigma^{(\ell,\ell)}}\right] \\ &\quad= \int_0^\infty \frac{\sqrt{2}}{\sigma^{(\ell,\ell)}\sqrt{\pi}} e^{-\frac{x^2}{2\sigma^{(\ell,\ell)^2}}} \int_0^x \frac{\sqrt{2}}{\sigma^{(k,\ell)}\sqrt{\pi}} e^{-\frac{y^2}{2\sigma^{(k,\ell)^2}}}\,\mathrm{d}y\,\mathrm{d}x \\ &\quad= \int_0^\infty \frac{\sqrt{2}}{\sigma^{(\ell,\ell)}\sqrt{\pi}} e^{-\frac{x^2}{2\sigma^{(\ell,\ell)^2}}} \operatorname{erf}\left(\frac{x}{\sigma^{(k,\ell)}\sqrt{2}}\right)\mathrm{d}x \\ &\quad= 1 - \frac{2}{\pi}\arctan\left(\frac{\sigma^{(k,\ell)}}{\sigma^{(\ell,\ell)}}\right), \end{aligned} \tag{44}$$

where we used the integral formula [50, Eqn. 2, p. 7] $\int_0^\infty \operatorname{erf}(ax)e^{-b^2x^2}\mathrm{d}x = (\pi/2 - \arctan(b/a))/(b\sqrt{\pi})$, with $a = 1/(\sigma^{(k,\ell)}\sqrt{2})$ and $b = 1/(\sigma^{(\ell,\ell)}\sqrt{2})$ to arrive at (44).

Denoting the probability density function of $\sigma^{(k,\ell)}/\sigma^{(\ell,\ell)}$ by p_σ, we get for fixed $\beta > 0$,

$$\begin{aligned}&\mathrm{P}\left[\left|\left\langle \mathbf{x}_j^{(k)}, \mathbf{x}_i^{(\ell)}\right\rangle\right| \geq \left|\left\langle \mathbf{x}_v^{(\ell)}, \mathbf{x}_i^{(\ell)}\right\rangle\right|\right] \\ &= \int_0^\infty \left(1 - \mathrm{P}\left[\left|\left\langle \mathbf{x}_j^{(k)}, \mathbf{x}_i^{(\ell)}\right\rangle\right| < \left|\left\langle \mathbf{x}_v^{(\ell)}, \mathbf{x}_i^{(\ell)}\right\rangle\right| \middle| x\right]\right) p_\sigma(x)\mathrm{d}x \\ &= \int_0^\infty \frac{2}{\pi}\arctan(x)\, p_\sigma(x)\mathrm{d}x \\ &\geq \int_\beta^\infty \frac{2}{\pi}\arctan(x)\, p_\sigma(x)\mathrm{d}x \\ &\geq \frac{2}{\pi}\arctan(\beta)\int_\beta^\infty p_\sigma(x)\mathrm{d}x \\ &= \frac{2}{\pi}\arctan(\beta)\,\mathrm{P}\left[\frac{\sigma^{(k,\ell)}}{\sigma^{(\ell,\ell)}} \geq \beta\right]. \end{aligned} \tag{45}$$

We continue by setting

$$\beta \coloneqq \frac{\sqrt{\operatorname{tr}\left(\tilde{\mathbf{R}}^{(k)}\tilde{\mathbf{R}}^{(\ell)}\right)}}{5\sqrt{3}\sqrt{\operatorname{tr}\left(\tilde{\mathbf{R}}^{(\ell)}\tilde{\mathbf{R}}^{(\ell)}\right)}}$$

and obtain

$$\begin{aligned}\mathrm{P}\left[\frac{\sigma^{(k,\ell)}}{\sigma^{(\ell,\ell)}} \geq \beta\right] &\geq \mathrm{P}\left[\left\{\sigma^{(k,\ell)} \geq \frac{1}{\sqrt{3}}\sqrt{\operatorname{tr}\left(\tilde{\mathbf{R}}^{(k)}\tilde{\mathbf{R}}^{(\ell)}\right)}\right\}\right. \\ &\qquad\left.\cap\left\{\sigma^{(\ell,\ell)} \leq 5\sqrt{\operatorname{tr}\left(\tilde{\mathbf{R}}^{(\ell)}\tilde{\mathbf{R}}^{(\ell)}\right)}\right\}\right] \\ &\geq 1 - \mathrm{P}\left[\sigma^{(k,\ell)} < \frac{1}{\sqrt{3}}\sqrt{\operatorname{tr}\left(\tilde{\mathbf{R}}^{(k)}\tilde{\mathbf{R}}^{(\ell)}\right)}\right] \end{aligned}$$

$$- \mathrm{P}\left[\sigma^{(\ell,\ell)} > 5\sqrt{\mathrm{tr}\Big(\tilde{\mathbf{R}}^{(\ell)}\tilde{\mathbf{R}}^{(\ell)}\Big)}\right]$$
$$> 1 - e^{-\frac{1}{9}} - e^{-8} > \frac{1}{10}, \tag{46}$$

where the second inequality follows from a union bound argument, and the third from

$$\mathrm{P}\left[\sigma^{(k,\ell)} < \frac{1}{\sqrt{3}}\sqrt{\mathrm{tr}\Big(\tilde{\mathbf{R}}^{(k)}\tilde{\mathbf{R}}^{(\ell)}\Big)}\right] \leq e^{-\frac{1}{9}} \tag{47}$$

and

$$\mathrm{P}\left[\sigma^{(\ell,\ell)} > 5\sqrt{\mathrm{tr}\Big(\tilde{\mathbf{R}}^{(\ell)}\tilde{\mathbf{R}}^{(\ell)}\Big)}\right] \leq e^{-8}, \tag{48}$$

both proven below. Inserting (46) into (45) yields the desired result.

Proof of (47): We start by noting that ${\sigma^{(k,\ell)}}^2 = \|\mathbf{C}^{(k)^\top}\mathbf{C}^{(\ell)}\mathbf{y}_i^{(\ell)}\|_2^2 = {\mathbf{y}_i^{(\ell)}}^\top \mathbf{C}^{(\ell)^\top}\tilde{\mathbf{R}}^{(k)}\mathbf{C}^{(\ell)}\mathbf{y}_i^{(\ell)}$ can be written as $\sigma^{(k,\ell)^2} \sim \sum_{m=1}^{M}\lambda_m z_m^2$, where λ_m, $m \in [M]$, denotes the non-negative eigenvalues of ${\mathbf{C}^{(\ell)}}^\top\tilde{\mathbf{R}}^{(k)}\mathbf{C}^{(\ell)}$ and z_m, $m \in [M]$, are independent standard normal random variables. Setting $\boldsymbol{\lambda} = [\lambda_1 \ \ldots \ \lambda_M]^\top$ and applying the lower tail bound [51, Lem. 1] for linear combinations of independent χ^2 random variables yields, for $t > 0$,

$$\mathrm{P}\Big[{\sigma^{(k,\ell)}}^2 \leq \|\boldsymbol{\lambda}\|_1 - 2\|\boldsymbol{\lambda}\|_2\sqrt{t}\Big] \leq e^{-t}. \tag{49}$$

The inequality (47) is obtained from (49) by noting that $\|\boldsymbol{\lambda}\|_2 \leq \|\boldsymbol{\lambda}\|_1$ and $\|\boldsymbol{\lambda}\|_1 = \mathrm{tr}\big({\mathbf{C}^{(\ell)}}^\top\tilde{\mathbf{R}}^{(k)}\mathbf{C}^{(\ell)}\big) = \mathrm{tr}\big(\tilde{\mathbf{R}}^{(k)}\mathbf{C}^{(\ell)}{\mathbf{C}^{(\ell)}}^\top\big) = \mathrm{tr}\big(\tilde{\mathbf{R}}^{(k)}\tilde{\mathbf{R}}^{(\ell)}\big)$, and by setting $t = 1/9$ in (49).

Proof of (48): Noting that $\sigma^{(\ell,\ell)} = f(\mathbf{y}_i^{(\ell)}) = \|{\mathbf{C}^{(\ell)}}^\top\mathbf{C}^{(\ell)}\mathbf{y}_i^{(\ell)}\|_2 = \|\tilde{\mathbf{R}}^{(\ell)}\mathbf{y}_i^{(\ell)}\|_2$ is Lipschitz with Lipschitz constant $\|\tilde{\mathbf{R}}^{(\ell)}\|_{2\to 2}$, we can invoke a well-known concentration inequality for Lipschitz functions of Gaussian random vectors with independent standard normal entries (see, e.g., [48, Thm. 8.40]) to get, for $t > 0$,

$$\mathrm{P}\Big[\Big\|\tilde{\mathbf{R}}^{(\ell)}\mathbf{y}_i^{(\ell)}\Big\|_2 - \mathbb{E}\Big[\Big\|\tilde{\mathbf{R}}^{(\ell)}\mathbf{y}_i^{(\ell)}\Big\|_2\Big] \geq t\Big] \leq \exp\left(-\frac{t^2}{2\|\tilde{\mathbf{R}}^{(\ell)}\|_{2\to 2}^2}\right). \tag{50}$$

The inequality (48) is now implied by $\mathbb{E}[\|\tilde{\mathbf{R}}^{(\ell)}\mathbf{y}_i^{(\ell)}\|_2] \leq \sqrt{\mathbb{E}\Big[\|\tilde{\mathbf{R}}^{(\ell)}\mathbf{y}_i^{(\ell)}\|_2^2\Big]} = \sqrt{\mathrm{tr}(\tilde{\mathbf{R}}^{(\ell)}\tilde{\mathbf{R}}^{(\ell)})} = \|\tilde{\mathbf{R}}^{(\ell)}\|_F$ (where we used Jensen's inequality), $\|\tilde{\mathbf{R}}^{(\ell)}\|_{2\to 2} \leq \|\tilde{\mathbf{R}}^{(\ell)}\|_F$, and (50) with $t = 4\|\tilde{\mathbf{R}}^{(\ell)}\|_F$.

REFERENCES

[1] M. Tschannen and H. Bölcskei, "Nonparametric nearest neighbor random process clustering," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, pp. 1207–1211, June 2015.

[2] Y. Wang, Z. Liu, and J.-C. Huang, "Multimedia content analysis – Using both audio and visual clues," *IEEE Signal Process. Mag.*, vol. 17, no. 6, pp. 12–36, 2000.

[3] K. Kalpakis, D. Gada, and V. Puttagunta, "Distance measures for effective clustering of ARIMA time-series," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, pp. 273–280, 2001.

[4] M. Corduas and D. Piccolo, "Time series clustering and classification by the autoregressive metric," *Comput. Stat. & Data Analysis*, vol. 52, no. 4, pp. 1860–1872, 2008.

[5] G. Marti, S. Andler, F. Nielsen, and P. Donnat, "Clustering financial time series: How long is enough?," in *Proc. Int. Joint Conf. Art. Intell. (IJCAI)*, pp. 2583–2589, 2016.

[6] G. Marti, F. Nielsen, P. Donnat, and S. Andler, "On clustering financial time series: A need for distances between dependent random variables," in *Computational Information Geometry*, pp. 149–174, Springer, 2017.

[7] P. Stoica and R. L. Moses, *Spectral Analysis of Signals*. Pearson/Prentice Hall, Upper Saddle River, NJ, 2005.

[8] J. Boets, K. De Cock, B. De Moor, and M. Espinoza, "Clustering time series, subspace identification and cepstral distances," *Commun. Inf. & Syst.*, vol. 5, no. 1, pp. 69–96, 2005.

[9] J. Caiado, N. Crato, and D. Peña, "A periodogram-based metric for time series classification," *Comput. Stat. & Data Analysis*, vol. 50, no. 10, pp. 2668–2684, 2006.

[10] Y. Kakizawa, R. H. Shumway, and M. Taniguchi, "Discrimination and clustering for multivariate time series," *J. Am. Stat. Assoc.*, vol. 93, no. 441, pp. 328–340, 1998.

[11] J. A. Vilar and S. Pértega, "Discriminant and cluster analysis for Gaussian stationary processes: Local linear fitting approach," *J. Nonparametric Stat.*, vol. 16, no. 3-4, pp. 443–462, 2004.

[12] R. M. Gray, *Probability, Random Processes, and Ergodic Properties*. Springer Science & Business Media, 2009.

[13] D. Ryabko, "Clustering processes," in *Proc. Int. Conf. Mach. Learn. (ICML)*, pp. 919–926, June 2010.

[14] A. Khaleghi, D. Ryabko, J. Mary, and P. Preux, "Online clustering of processes," in *Proc. Int. Conf. Art. Intell. Stat. (AISTATS)*, pp. 601–609, 2012.

[15] A. Khaleghi, D. Ryabko, J. Mary, and P. Preux, "Consistent algorithms for clustering time series," *J. Mach. Learn. Res.*, vol. 17, no. 3, pp. 1–32, 2016.

[16] G. Marti, F. Nielsen, and P. Donnat, "Optimal copula transport for clustering multivariate time series," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pp. 2379–2383, 2016.

[17] Y. Xiong and D.-Y. Yeung, "Time series clustering with ARMA mixtures," *Pattern Recogn.*, vol. 37, no. 8, pp. 1675–1689, 2004.

[18] P. Borysov, J. Hannig, and J. Marron, "Asymptotics of hierarchical clustering for growing dimension," *J. Multivar. Analysis*, vol. 124, pp. 465–479, 2014.

[19] D. Ryabko and J. Mary, "A binary-classification-based metric between time-series distributions and its use in statistical and learning problems," *J. Mach. Learn. Res.*, vol. 14, no. 1, pp. 2837–2856, 2013.

[20] R. Heckel and H. Bölcskei, "Robust subspace clustering via thresholding," *IEEE Trans. Inf. Theory*, vol. 61, no. 11, pp. 6320–6342, 2015.

[21] I. Katsavounidis, C.-C. Jay Kuo, and Z. Zhang, "A new initialization technique for generalized Lloyd iteration," *IEEE Signal Process. Lett.*, vol. 1, no. 10, pp. 144–146, 1994.

[22] P. Esling and C. Agon, "Time-series data mining," *ACM Comput. Surv. (CSUR)*, vol. 45, no. 1, p. 12, 2012.

[23] M. Tucci and M. Raugi, "Analysis of spectral clustering algorithms for linear and nonlinear time series," in *Proc. IEEE Int. Conf. Intell. Syst. Design Appl. (ISDA)*, pp. 925–930, 2011.

[24] L. N. Ferreira and L. Zhao, "Time series clustering via community detection in networks," *Inf. Sci.*, vol. 326, pp. 227–242, 2016.

[25] U. von Luxburg, "A tutorial on spectral clustering," *Stat. Comput.*, vol. 17, no. 4, pp. 395–416, 2007.

[26] S. M. Kay, *Modern Spectral Estimation*. Prentice Hall, 1988.

[27] D. Yan, L. Huang, and M. I. Jordan, "Fast approximate spectral clustering," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining (KDD)*, pp. 907–916, ACM, 2009.

[28] M. Li, X.-C. Lian, J. T. Kwok, and B.-L. Lu, "Time and space efficient spectral clustering via column sampling," in *Proc. IEEE Conf. Comput. Vision Pattern Recogn. (CVPR)*, pp. 2297–2304, 2011.

[29] W.-Y. Chen, Y. Song, H. Bai, C.-J. Lin, and E. Y. Chang, "Parallel spectral clustering in distributed systems," *IEEE Trans. Pattern Analysis Mach. Intell.*, vol. 33, no. 3, pp. 568–586, 2011.

[30] G. Maruyama, "The harmonic analysis of stationary stochastic processes," *Memoirs Fac. Sci. Kyushu Univ. Ser. A, Math.*, vol. 4, no. 1, pp. 45–106, 1949.

[31] H. White, *Asymptotic Theory for Econometricians*. Academic Press, 2014.

[32] D. Skinner, "Pruning the decimation in-time FFT algorithm," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 24, no. 2, pp. 193–194, 1976.

[33] E. Elhamifar and R. Vidal, "Sparse subspace clustering: Algorithm, theory, and applications," *IEEE Trans. Pattern Analysis Mach. Intell.*, vol. 35, no. 11, pp. 2765–2781, 2013.

[34] M. Soltanolkotabi, E. Elhamifar, and E. J. Candès, "Robust subspace clustering," *Ann. Stat.*, vol. 42, no. 2, pp. 669–699, 2014.

[35] L. Demanet, P.-D. Létourneau, N. Boumal, H. Calandra, J. Chiu, and S. Snelson, "Matrix probing: A randomized preconditioner for the wave-equation Hessian," *Appl. Comput. Harmon. Analysis*, vol. 32, no. 2, pp. 155–168, 2012.

[36] R. Adamczak, "A note on the Hanson-Wright inequality for random vectors with dependencies," *Electron. Commun. Probab.*, vol. 20, no. 72, pp. 1–13, 2015.

[37] M. Rudelson and R. Vershynin, "Hanson-Wright inequality and sub-gaussian concentration," *Electron. Commun. Probab.*, vol. 18, pp. no. 82, 1–9, 2013.

[38] J. Friedman, T. Hastie, and R. Tibshirani, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer New York, 2009.

[39] L. Li and B. A. Prakash, "Time series clustering: Complex is simpler!," in *Proc. Int. Conf. Mach. Learn. (ICML)*, pp. 185–192, 2011.

[40] E. A. Maharaj and P. D'Urso, "Fuzzy clustering of time series in the frequency domain," *Inf. Sci.*, vol. 181, no. 7, pp. 1187–1211, 2011.

[41] S. Sanei and J. A. Chambers, *EEG Signal Processing*. John Wiley & Sons, 2008.

[42] R. G. Andrzejak, K. Lehnertz, F. Mormann, C. Rieke, P. David, and C. E. Elger, "Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state," *Phys. Rev. E*, vol. 64, no. 6, p. 061907, 2001.

[43] M.-F. Balcan, A. Blum, and S. Vempala, "A discriminative framework for clustering via similarity functions," in *Proc. Annual ACM Symp. Theory Comp.*, pp. 671–680, 2008.

[44] R. M. Gray, "Toeplitz and circulant matrices: A review," *Foundations & Trends Commun. Inf. Theory*, vol. 2, no. 3, pp. 155–239, 2006.

[45] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*. Cambridge University Press, 1991.

[46] W. Schudy and M. Sviridenko, "Concentration and moment inequalities for polynomials of independent random variables," in *Proc. ACM-SIAM Symp. Discrete Algo. (SODA)*, pp. 437–446, 2012.

[47] S. Boucheron, G. Lugosi, and P. Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

[48] S. Foucart and H. Rauhut, *A Mathematical Introduction to Compressive Sensing*. Springer, Berlin, Heidelberg, 2013.

[49] V. Buldygin and K. Moskvichova, "The sub-gaussian norm of a binary random variable," *Theory Probab. Math. Stat.*, vol. 86, pp. 33–49, 2013.

[50] E. W. Ng and M. Geller, "A table of integrals of the error functions," *J. Res. Natl. Bureau Standards B*, vol. 73, pp. 1–20, 1969.

[51] B. Laurent and P. Massart, "Adaptive estimation of a quadratic functional by model selection," *Ann. Stat.*, vol. 28, no. 5, pp. 1302–1338, 2000.