Statistics and Its Interface Volume 16 (2023) 319–335

# Adaptive clustering and feature selection for categorical time series using interpretable frequency-domain features

Scott A. Bruce*

This article presents a novel approach to clustering and feature selection for categorical time series via interpretable frequency-domain features. A distance measure is introduced based on the spectral envelope and optimal scalings, which parsimoniously characterize prominent cyclical patterns in categorical time series. Using this distance, partitional clustering algorithms are introduced for accurately clustering categorical time series. These adaptive procedures offer simultaneous feature selection for identifying important features that distinguish clusters and fuzzy membership when time series exhibit similarities to multiple clusters. Clustering consistency of the proposed methods is investigated, and simulation studies are used to demonstrate clustering accuracy with various underlying group structures. The proposed methods are used to cluster sleep stage time series for sleep disorder patients in order to identify particular oscillatory patterns associated with sleep disruption.

AMS 2000 subject classifications: Primary 62M10, 62M15, 62H30; secondary 62H86.
Keywords and phrases: Categorical time series, Multiple time series, Optimal scaling, Partitional clustering, Spectral envelope, Unsupervised learning.

## 1. INTRODUCTION

Categorical time series are prevalent across a variety of scientific disciplines including climate science [3], sleep medicine [24], and genomics [41]. For many studies, cyclical patterns in categorical time series are of particular interest and can be useful for characterizing differences among populations [22, 25] and within a population of interest [24]. This is due to the valuable and interpretable information contained in the frequency domain that can be used to better understand the scientific nature of differences among groups [42]. Specifically, there is a dearth of methods for clustering categorical time series based on their cyclical behavior, and the goal of this article is to offer a theoretically-justified clustering framework and collection of interpretable, adaptive frequency-domain algorithms for identifying clusters and their defining features within a population of categorical time series.

*ORCID: 0000-0002-7904-4211.

For example, this article is motivated by a sleep study in which participants with different types of sleep disorders are monitored during a night of sleep via polysomnography. This study can be used to better understand differences in nocturnal physiology within the population of sleep disorder patients. During sleep, the body cycles through different stages: movement/wakefulness, rapid eye movement (REM) sleep, and non-rapid eye movement (NREM) sleep, which is further divided into light sleep (S1,S2) and deep sleep (S3, S4). These stages are characterized by particular brain activity patterns measured via electroencephalography (EEG) [37] and recorded at regular intervals throughout the night for each participant. For illustration, Figure 1 displays examples of full night sleep stage time series for four individuals with different sleep-related pathologies: insomnia (INS), nocturnal frontal lobe epilepsy (NFLE), periodic leg movements (PLM), and rapid eye movement behavior disorder (RBD). Sleep disorders tend to disrupt normal cyclical behavior in different ways, which can negatively impact overall health and well-being [33]. The analysis presented herein seeks to identify common profiles, or clusters, of sleep stage

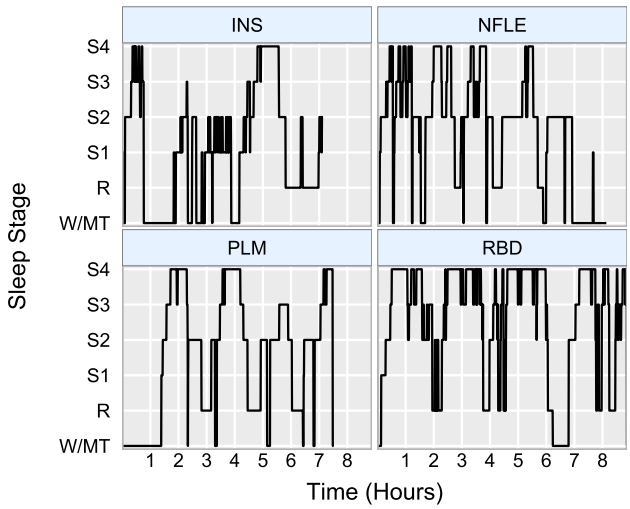


*Figure 1. Sleep stage time series for four sleep study participants with different sleep-related pathologies: insomnia (INS), nocturnal frontal lobe epilepsy (NFLE), periodic leg movements (PLM), and REM behavior disorder (RBD).*

time series among sleep disorder patients to better understand common ways in which sleep disruptions alter typical nocturnal cyclical patterns. Additionally, it is of interest to identify which features of sleep stage time series are more useful in distinguishing clusters. It is possible that cyclical patterns with particular frequencies or involving particular sleep stages are more helpful in characterizing different types of sleep disruptions, so an analysis that can adaptively identify distinctive features is desirable. Such an analysis can provide helpful tools in diagnosing sleep disorder and evaluating treatment efficacy.

Clustering of real-valued time series has been well-studied; see [27, 40, 30] for a general introduction and survey of methods and [19] for a review of frequency-domain methods. However, clustering of categorical time series has received considerably less attention. The majority of statistical methods for categorical time series analysis have been developed for analyzing a single categorical time series. Some examples include the Markov chain model of [2], the link function approach of [11], the likelihood-based method of [12], and the spectral envelope approach for analyzing a single time series introduced in [41]. A comprehensive discussion of this research direction can be found in [13]. However, the features introduced in these works for characterizing categorical time series have been used more recently for feature-based clustering. In the statistical literature, [34] introduces clustering methods for categorical time series based on time-homogeneous first-order Markov chains via group-specific transition matrices. More recently, [8] introduces a time-homogeneous first-order semi-Markov model, which allows for more flexible modeling of group-specific sojourn time distributions compared to standard Markov models. In the computer science literature, [15] introduces a distance measure and clustering algorithm that accounts for differences in the raw values of the time series and in the first-order temporal correlation structures.

For applications where cyclical patterns are of interest, Markov models offer little help in understanding key differences in cyclical patterns among categorical time series. Also, first-order methods can miss important features of categorical time series with prominent cyclical patterns attributed to higher order lag dependency or higher order dependence structures [18]. However, these first-order models cannot be easily extended due to exponential growth in the feature space for higher-order models. To this end, we propose using the spectral envelope and its corresponding set of optimal scalings [41] as low-dimensional, interpretable features for clustering categorical time series. Use of these features is motivated by noticing that most categorical time series can be represented in terms of their prominent oscillatory patterns, characterized by the spectral envelope, and by the set of mappings from categories to numerical values that accentuate specific oscillatory patterns, characterized by the optimal scalings. [24] introduces the spectral envelope surface for quantifying the association between the oscillatory patterns of a collection of categorical time series and continuous covariates. However, it is not immediately useful for clustering. To the best of my knowledge, this article presents the first statistical approach for frequency-domain clustering of categorical time series.

Clustering algorithms for real data applications should also be capable of adapting to commonly encountered situations arising in practice. For example, not all features may be equally helpful in distinguishing clusters. In these situations, clustering accuracy can be improved by focusing on important features that exhibit meaningful differences across clusters [46]. It is also possible that series may exhibit characteristics resembling more than one cluster. In many cases, this is due to time series dynamics that are drifting or switching in a vague manner not focused on a particular point in time [31]. To preserve this information, [31] introduces a class of fuzzy clustering models for time series that allows for time series to exhibit partial memberships to multiple clusters. This article introduces adaptive algorithms that can offer feature selection and fuzzy clustering by extending the frameworks of [46, 31] for use in clustering categorical time series.

The proposed partitional clustering methods have a similar structure and are briefly described as follows. Each time series to be clustered is represented as a vector-valued time series through the use of indicator variables. The smoothed spectral density matrix of this vector-valued time series is then obtained, and the spectral envelope and optimal scalings at each frequency are computed from the estimated spectral matrix. An initial partition is then randomly determined, and the spectral envelope and optimal scalings for each cluster are estimated respectively. A distance measure that considers both of these features is then used to assess the distance from each series to each cluster, and each series is reassigned to its nearest cluster. The algorithms then alternate between reassigning cluster memberships and updating cluster features until convergence. Under the proposed framework, the misclassification probability is bounded as long as the spectral density matrix estimator is consistent. The proposed algorithms also are shown to perform well in simulated examples and real data application.

The remainder of the paper is organized as follows. Section 2 provides definitions of the spectral envelope and optimal scalings and corresponding estimators. Section 3 introduces the components of the proposed clustering framework, including the frequency-domain distance measure and its theoretical properties, clustering algorithms for standard K-means and K-medoids clustering, sparse clustering via feature selection, fuzzy clustering, and simulated examples to demonstrate strong finite-sample performance. Section 4 details the application of the proposed clustering procedures to the analysis of sleep stage time series. Section 5 provides some closing comments and impactful extensions of this work.

## 2. FREQUENCY DOMAIN FEATURES FOR CATEGORICAL TIME SERIES

### 2.1 Spectral envelope and optimal scalings

Let $X_t$, for $t = 1, 2, \ldots$, be a categorical time series with finite state space $\mathcal{C} = \{c_1, c_2, \ldots, c_S\}$. Assume $X_t$ is stationary such that $\{X_1, X_2, \ldots, X_t\} \overset{d}{=} \{X_{1+h}, X_{2+h}, \ldots, X_{t+h}\}$ for $h \geq 0$ and $\inf_{s=1,2,\ldots,S} \mathrm{P}(X_t = c_s) > 0$ so that there are no absorbing states. We can explore the frequency domain properties of $X_t$ by assigning numerical values, or scalings, to categories, $\beta = (\beta_1, \beta_2, \ldots, \beta_S)' \in \mathbb{R}^S$ and exploring prominent oscillatory patterns of the real-valued time series $X_t(\beta) = \beta_s$ when $X_t = c_s$. Different sets of scalings are considered that maximally emphasize oscillatory patterns at different frequencies [41].

**Definition 1.** Assuming $X_t(\beta)$ has a continuous and bounded spectral density

$$f_x(\omega; \beta) = \sum_{h=-\infty}^{\infty} \mathrm{Cov}[X_t(\beta), X_{t+h}(\beta)] \exp(-2\pi i \omega h)$$

for $\omega \in \mathbb{R}$, the spectral envelope and optimal scalings for frequency $\omega$ are defined as

$$\begin{aligned} \lambda(\omega) &= \max_{\beta \in \mathbb{R}^S, \beta \not\propto 1_S} \frac{f_x(\omega; \beta)}{V_x(\beta)}, \\ B(\omega) &= \arg\max_{\beta \in \mathbb{R}^S, \beta \not\propto 1_S} \frac{f_x(\omega; \beta)}{V_x(\beta)}, \end{aligned} \tag{1}$$

where $V_x(\beta) = \mathrm{Var}[X_t(\beta)]$ and $1_S$ is an $S$-dimensional vector of ones.

Scalings proportional to $1_S$ assign the same value to all categories and are not considered. In addition to being uninteresting, the spectral envelope and optimal scalings are not well-defined for such scalings since $V_x(\beta) = 0$. The spectral envelope represents the maximal normalized spectral density at frequency $\omega$ for different possible scalings such that $f_x(\omega, \beta) \leq \lambda(\omega)\ \forall \beta \in \mathbb{R}^S \not\propto \{1_S\}$, and the optimal scalings represent the particular transformation that attains this bound. Taken together, these interpretable features characterize dominant oscillatory patterns in categorical time series parsimoniously with minimal loss of information [41] and offer a good foundation for clustering.

For illustration, consider the four categorical time series and their corresponding estimated spectral envelopes and optimal scalings presented in Figure 2. The first time series cycles through categories slowly compared to the second time series. This is captured in the spectral envelope for each series, which is dominated by low frequency power for the first series and high frequency power for the second series. Comparing the third and fourth time series, they have similar power across frequencies but differ in their traversals through categories. For example, the third time series consists of longer visits to category 3, which contributes to low frequency power for this series, and shorter stays in other categories. However, the fourth time series consists of longer visits to category 1 and shorter stays in other categories. These characteristics are captured in the optimal scalings over low frequencies for these series, which assign larger values to categories with longer visits contributing to low frequency power.

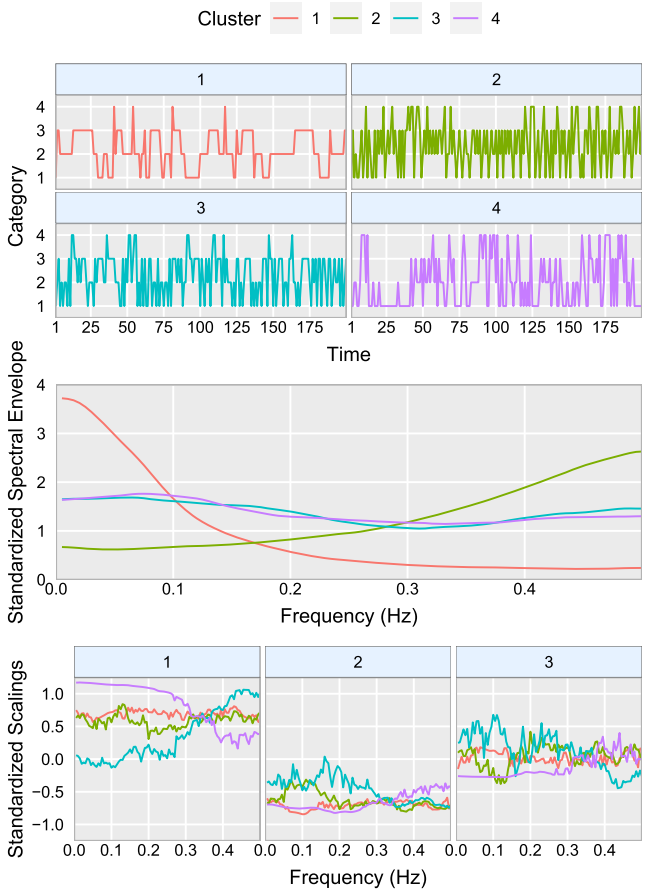


*Figure 2. Example 1 Visualization: 2 × 2 panel of simulated time series corresponding to four different clusters (top), estimated standardized spectral envelopes (middle), and estimated standardized optimal scalings for each category (bottom) for each cluster. This figure appears in color in the electronic version of this article, and any mention of color refers to that version.*

### 2.2 Computation

A multivariate point process representation of the categorical time series $X_t$ is used to facilitate computation of these frequency-domain features [41, 24, 25]. Let $Y_t$ be an $(S-1)$-dimensional vector such that the $s$th element of $Y_t$ equals one if $X_t = c_s$ for $s = 1, \ldots, S-1$ and zero elsewhere. Defining $Y_t$ in this way corresponds to setting the scaling for the $S$th category to zero, thus restricting the set of optimal scalings to a lower-dimensional space. This property is also

present in the original formulation of the spectral envelope and optimal scalings; interested readers can find more details in [41]. The assumption that $f_x(\omega, \beta)$ is continuous is necessary and sufficient for ensuring $Y_t$ has a continuous spectral density

$$f_y(\omega) = \sum_{h=-\infty}^{\infty} \text{Cov}[Y_t, Y_{t+h}] \exp(-2\pi i \omega h).$$

$f_y(\omega)$ is a positive definite Hermitian $(S-1)\times(S-1)$ matrix. Assuming $f_y(\omega)$ and $V_y = \text{Var}[Y_t]$ are non-singular for all $\omega \in \mathbb{R}$ [4], we can define the spectral envelope and optimal scalings via eigendecomposition.

**Definition 2.** For $\omega \in \mathbb{R}$, the spectral envelope, $\lambda(\omega)$, is defined as the largest eigenvalue of $h(\omega) = V_y^{-1/2} f_y(\omega) V_y^{-1/2}$. The $(S-1)$-variate vector of optimal scalings, $\gamma(\omega)$, is defined as the eigenvector associated with $\lambda(\omega)$.

Some remarks are in order. First, since the spectral density matrix is Hermitian with a skew symmetric imaginary component, for every $a \in \mathbb{R}^{S-1}$, $a' f_y(\omega) a = a' f_y^{re}(\omega) a$, where $f_y^{re}(\omega)$ is the real part of $f_y(\omega)$. Thus, the spectral envelope is equivalent to the largest eigenvalue of $h(\omega)^{re} = V_y^{-1/2} f_y^{re}(\omega) V_y^{-1/2}$. Second, the optimal scalings derived in this manner are connected with those of the original formulation introduced in (1) [24]. If $V_y^{1/2}\gamma(\omega)$ is an eigenvector of $h^{re}(\omega)$ associated with $\lambda(\omega)$, then

$$\begin{bmatrix} \gamma(\omega) \\ 0 \end{bmatrix} = \underset{\beta \in \mathbb{R}^S, \beta \not\propto 1_S}{\arg\max} \frac{f_x(\omega; \beta)}{V_x(\beta)}.$$

Furthermore, assuming the eigenvalues of $h^{re}(\omega)$ are distinct, there exists a unique $\gamma(\omega)$ such that $V_y^{1/2}\gamma(\omega)$ is an eigenvector of $h^{re}(\omega)$ associated with $\lambda(\omega)$ where $\gamma(\omega)' V_y \gamma(\omega) = 1$ and with the first nonzero entry of $V^{1/2}\gamma(\omega)$ to be positive.

## 2.3 Estimation

Consider finite realizations of $X_t$ and $Y_t$ denoted as $\{X_t\}_{t=1}^T$ and $\{Y_t\}_{t=1}^T$. A popular and computationally-efficient estimator of the spectral matrix $f_y(\omega)$ is the periodogram

$$I(\omega_j) = T^{-1} \left| \sum_{t=1}^{T} Y_t \exp(-2\pi i \omega_j t) \right|^2$$

where $\omega_j = j/T$ for $j = 1, \ldots, J = \lfloor (T-1)/2 \rfloor$ are the Fourier frequencies. The periodogram is well-known to be an asymptotically unbiased but inconsistent estimator of $f_y(\omega)$ and can be smoothed over frequencies to obtain a consistent estimator [4]. Consider the smoothed periodogram estimator

$$\hat{f}_y(\omega_j) = \sum_{\ell=-B_T}^{B_T} W_{B_T,\ell} I(\omega_j + \ell/T), \tag{2}$$

where $2B_T + 1$ is the smoothing span and $W_{B_T,\ell}$ are non-negative weights that satisfy the following conditions:

$$W_{B_T,\ell} = W_{B_T,-\ell}, \quad \sum_{\ell=-B_T}^{B_T} W_{B_T,\ell} = 1.$$

Generally, weights are chosen such that $W_{B_T,0}$ is a decreasing function of $B_T$, and many options are available in practice [40]. It is known that $\hat{f}_y(\omega_j)$ is consistent if $B_T \to \infty$ and $B_T T^{-1} \to 0$ as $T \to \infty$ [4]. In what follows, modified Daniell kernel weights are used with smoothing span $B_T = \lfloor \sqrt{T} \rfloor$. The literature on spectral matrix estimation is vast, and many other approaches [9, 38, 23] can be used here. However, kernel smoothing offers both computational efficiency and ease of theoretical exposition.

Following Definition 2, eigendecomposition of $\hat{h}(\omega)^{re} = \widehat{V}_y^{-1/2} \hat{f}_y^{re}(\omega) \widehat{V}_y^{-1/2}$ where $\hat{f}_y(\omega)$ is the smoothed periodogram estimator (2) and $\widehat{V}_y$ is the sample covariance matrix of $Y_t$ then provides both the estimated spectral envelope, $\hat{\lambda}(\omega)$, and optimal scalings, $\hat{\gamma}(\omega)$, as the largest eigenvalue and corresponding eigenvector respectively.

# 3. FEATURE-BASED CLUSTERING FOR CATEGORICAL TIME SERIES

Consider a population of categorical time series with common state space $\mathcal{C} = \{c_1, c_2, \ldots, c_S\}$ composed of $k = 1, 2, \ldots, K$ clusters with $K$ being fixed. Each cluster consists of a collection of independent stationary series with common spectral envelope, $\lambda^{(k)}(\omega)$, and $(S-1)$-variate optimal scalings, $\gamma^{(k)}(\omega) = \{\gamma_s^{(k)}(\omega_j)\}_{s=1}^{S-1}$. Suppose we have a collection of $N$ independent time series from this population, $X_{1t}, X_{2t}, \ldots, X_{Nt}$ and observe finite realizations, $\{X_{1t}\}_{t=1}^T, \{X_{2t}\}_{t=1}^T, \ldots, \{X_{Nt}\}_{t=1}^T$. Assuming each series belongs to one of the $K$ clusters, let $g_i \in \{1, 2, \ldots, K\}$ represent the unknown cluster membership for the $i$th series. For better illustration, consider the following example.

**Example 1. Four clusters with differences in particular features.** Following [13], realizations of categorical time series $X_{it}$ belonging to the $k$th cluster can be generated from the multinomial logit model

$$P(X_{it} = c_s) = \frac{\exp(\alpha_{k,s}' Y_{it-1})}{1 + \sum_{s=1}^{S-1} \exp(\alpha_{k,s}' Y_{it-1})}, \quad s = 1, \ldots, S-1,$$

and

$$P(X_{it} = c_S) = \frac{1}{1 + \sum_{s=1}^{S-1} \exp(\alpha_{k,s}' Y_{it-1})},$$

where $Y_{it}$ is the $(S-1)$-dimensional point process representation of $X_{it}$ introduced in Section 2.2 and $\alpha_{k,s}$ for $s = 1, \ldots, S-1$ are regression parameters for the $k$th cluster. This model satisfies $\sum_{s=1}^{S} P(X_{it} = c_s) = 1$ for $t = 1, 2, \ldots$

and incorporates a lag of order one. Let the number of categories $S = 4$, the number of clusters $K = 4$, and cluster-specific regression parameters

$$\begin{aligned}
\alpha_{1,1} &= (3,1,1)', \alpha_{1,2} = (1,3,1)', \alpha_{1,3} = (1,1,3)', \\
\alpha_{2,1} &= (-1,1,1)', \alpha_{2,2} = (1,-1,1)', \alpha_{2,3} = (1,1,-1)', \\
\alpha_{3,1} &= (-1,1,1)', \alpha_{3,2} = (1,1,1)', \alpha_{3,3} = (1,1,2)', \\
\alpha_{4,1} &= (2,1,1)', \alpha_{4,2} = (1,1,1)', \alpha_{4,3} = (1,1,-1)'.
\end{aligned}$$

Consider $N = 100$ realizations of length $T = 200$ composed of $N/K = 25$ independent realizations from each cluster. Figure 2 displays a single realization from each of the four clusters and the estimated spectral envelopes and optimal scalings for each cluster. This example is particularly interesting since all clusters are not well-separated along all features, which is likely to be the case in practice. Clusters 1 and 2 are different in their spectral envelopes but share common optimal scalings, and clusters 3 and 4 share a common spectral envelope but are different in their optimal scalings. In the following sections, this example will be used to evaluate performance of the proposed clustering algorithms.

## 3.1 A distance measure

Let $\hat{\lambda}_i(\omega_j)$ and $\hat{\gamma}_i(\omega_j) = \{\hat{\gamma}_{i,s}(\omega_j)\}_{s=1}^{S-1}$ represent the estimated spectral envelope and optimal scalings for $X_{it}$ where $\omega_j = j/T$ for $j = 1, \ldots, J = \lfloor (T-1)/2 \rfloor$ are the Fourier frequencies. To measure the distance from the $i$th series to the $k$th cluster, consider the following distance measure

$$\begin{aligned}
d_{i,k} = &\sum_{j=1}^{J} \left( \frac{\hat{\lambda}_i(\omega_j)}{||\hat{\lambda}_i||_2} - \frac{\lambda^{(k)}(\omega_j)}{||\lambda^{(k)}||_2} \right)^2 \\
&+ \sum_{j=1}^{J} \sum_{s=1}^{S-1} \left( \frac{\hat{\gamma}_{i,s}(\omega_j)}{||\hat{\gamma}_i||_2} - \frac{\gamma_s^{(k)}(\omega_j)}{||\gamma^{(k)}||_2} \right)^2
\end{aligned} \tag{3}$$

where

$$||\lambda||_2 = \sqrt{T^{-1} \sum_{j=1}^{J} |\lambda(\omega_j)|^2} \approx \sqrt{\int_0^{1/2} |\lambda(\omega)|^2 d\omega} \quad \text{and}$$

$$||\gamma||_2 = \sqrt{\sum_{s=1}^{S-1} T^{-1} \sum_{j=1}^{J} |\gamma_s(\omega)|^2} \approx \sqrt{\sum_{s=1}^{S-1} \int_0^{1/2} |\gamma_s(\omega)|^2 d\omega}.$$

Features are standardized to ensure (3) is similarly sensitive to differences in the spectral envelope and optimal scalings. In order to demonstrate that this distance measure offers consistent clustering, the following assumptions are needed.

**Assumption 1** (Mixing). *Let $Y_{it}$ be the multivariate point process representation of $X_{it}$. $Y_{it}$ is strictly stationary and the span of dependence is small such that cumulants of all orders exist and are absolutely summable [4, Assumption 2.6.1].*

**Assumption 2** (Smoothness). *Let $f_{iy}(\omega)$ be the $(S-1) \times (S-1)$ spectral density matrix of $Y_{it}$. Each element of $f_{iy}(\omega)$ has bounded and continuous first derivatives.*

**Assumption 3** (Separation). *Cluster features are well-separated such that*

$$\begin{aligned}
&\sum_{j=1}^{J} \left( \frac{\lambda^{(k)}(\omega_j)}{||\lambda^{(k)}||_2} - \frac{\lambda^{(k')}(\omega_j)}{||\lambda^{(k')}||_2} \right)^2 \\
&+ \sum_{j=1}^{J} \sum_{s=1}^{S-1} \left( \frac{\gamma_s^{(k)}(\omega_j)}{||\gamma^{(k)}||_2} - \frac{\gamma_s^{(k')}(\omega_j)}{||\gamma^{(k')}||_2} \right)^2 \geq CT
\end{aligned}$$

*for a positive constant $C$ $\forall k, k' \in \{1, 2, \ldots, K\}$.*

Under Assumptions 1 and 2, asymptotic consistency of the estimates $\hat{\lambda}_i(\omega)$ and $\hat{\gamma}_i(\omega)$ can be established, and the largest eigenvalue of the spectral density matrix is continuous and bounded from above. Assumption 3 implies that the cluster-specific features are well-separated in the sum-of-squares sense. It should be noted that this assumption does not require all features to be well-separated; it is possible that separation may be concentrated within certain frequency bands or occurring primarily in the spectral envelope or optimal scalings. With these assumptions, the following theorem states the consistency for using (3) for clustering.

**Theorem 1.** *Let $X_{it}$ be a time series belonging to cluster $k^*$. Under Assumptions 1-3, the probability of the distance between $X_{it}$ and cluster $k^*$ exceeding the distance between $X_{it}$ and a different cluster $k \neq k^*$ is bounded such that*

$$P(d_{i,k^*} > d_{i,k}) = O(B_T T^{-1}) \text{ for } \forall k \neq k^*.$$

*where $d_{i,k}$ is defined in (3).*

Proof for this theorem is provided in the Appendix.

## 3.2 K-means clustering

Using this distance measure, the standard K-means clustering framework [28, 29] can be adapted for categorical time series as follows. Let $\hat{\mathbf{g}} = (\hat{g}_1, \ldots, \hat{g}_N)$ be the collection of estimated cluster membership values, then

$$\hat{\mathbf{g}} = \arg\min_{\mathbf{g}} \sum_{i=1}^{N} \sum_{k=1}^{K} I(g_i = k) d_{i,k} \tag{4}$$

where $I(g_i = k) = 1$ if the $i$th series is assigned to the $k$th cluster and $I(g_i = k) = 0$ otherwise. In practice, cluster features are unknown and estimated from the data as

$$\begin{aligned}
\frac{\hat{\lambda}^{(k)}(\omega_j)}{||\hat{\lambda}^{(k)}||_2} &= \frac{\sum_{i=1}^{N} I(\hat{g}_i = k) \frac{\hat{\lambda}_i(\omega_j)}{||\hat{\lambda}_i||_2}}{\sum_{i=1}^{N} I(\hat{g}_i = k)} \\
\frac{\hat{\gamma}_s^{(k)}(\omega_j)}{||\hat{\gamma}^{(k)}||_2} &= \frac{\sum_{i=1}^{N} I(\hat{g}_i = k) \frac{\hat{\gamma}_{i,s}(\omega_j)}{||\hat{\gamma}_i||_2}}{\sum_{i=1}^{N} I(\hat{g}_i = k)}
\end{aligned}$$

for $k = 1, \ldots, K$. $K$ is also unknown and can be estimated from the data. Many data-driven methods are presented in the literature for selecting $K$ [32, 44]. The traditional scree plot is used in what follows. For a particular choice of $K$, the above optimization problem can be solved via the iterative algorithm presented in Algorithm 1. Multiple initializations should be considered to avoid settling at local optima.

**Result:** Cluster assignments, $\hat{\mathbf{g}} = (\hat{g}_1, \ldots, \hat{g}_N)$, and features, $\frac{\hat{\lambda}^{(k)}(\omega_j)}{||\hat{\lambda}^{(k)}||_2}, \frac{\hat{\gamma}_s^{(k)}(\omega_j)}{||\hat{\gamma}^{(k)}||_2}$ for $k = 1, \ldots, K$, $j = 1, \ldots, J$, $s = 1, \ldots, S-1$.

Initialize $\hat{\mathbf{g}}$ by independently drawing each $\hat{g}_i$ from $\{1, \ldots, K\}$ with equal probability.

stop $\leftarrow 0$

**while** *stop=0* **do**

Update cluster features

$$\frac{\hat{\lambda}^{(k)}(\omega_j)}{||\hat{\lambda}^{(k)}||_2} \leftarrow \frac{\sum_{i=1}^N I(\hat{g}_i = k)\frac{\hat{\lambda}_i(\omega_j)}{||\hat{\lambda}_i||_2}}{\sum_{i=1}^N I(\hat{g}_i = k)},$$

$$\frac{\hat{\gamma}_s^{(k)}(\omega_j)}{||\hat{\gamma}^{(k)}||_2} \leftarrow \frac{\sum_{i=1}^N I(\hat{g}_i = k)\frac{\hat{\gamma}_{i,s}(\omega_j)}{||\hat{\gamma}_i||_2}}{\sum_{i=1}^N I(\hat{g}_i = k)},$$

for $j = 1, \ldots, J$ , $k = 1, \ldots, K$, and $s = 1, \ldots, S-1$ where $I(\hat{g}_i = k) = 1$ if series $i$ is assigned to cluster $k$ (0 otherwise).

Update distances

$$d_{i,k} \leftarrow \sum_{j=1}^J \left( \frac{\hat{\lambda}_i(\omega_j)}{||\hat{\lambda}_i||_2} - \frac{\hat{\lambda}^{(k)}(\omega_j)}{||\hat{\lambda}^{(k)}||_2} \right)^2 + \sum_{j=1}^J \sum_{s=1}^{S-1} \left( \frac{\hat{\gamma}_{i,s}(\omega_j)}{||\hat{\gamma}_i||_2} - \frac{\hat{\gamma}_s^{(k)}(\omega_j)}{||\hat{\gamma}^{(k)}||_2} \right)^2$$

for $i = 1, \ldots, N, k = 1, \ldots, K$.

Update cluster assignments

$$\tilde{g}_i \leftarrow \arg\min_k d_{i,k}, \ i = 1, \ldots, N.$$

**if** $\tilde{\mathbf{g}} = \hat{\mathbf{g}}$ **then** stop $\leftarrow 1$

**else** $\hat{\mathbf{g}} \leftarrow \tilde{\mathbf{g}}$

**end**

**return** $\hat{\mathbf{g}}, \hat{\lambda}^{(k)}(\omega_j)/||\hat{\lambda}^{(k)}||_2, \hat{\gamma}_s^{(k)}(\omega_j)/||\hat{\gamma}^{(k)}||_2$

**Algorithm 1:** K-MEANS CLUSTERING

## 3.3 K-medoids clustering

In some cases, it is preferable to estimate cluster features, $\lambda^{(k)}(\omega_j)/||\lambda^{(k)}||_2$ and $\gamma_s^{(k)}(\omega_j)/||\gamma^{(k)}||_2$, using the estimated features for a particular time series belonging to the cluster, rather than the mean across all members of the cluster. This can be especially beneficial when clusters contain outliers that would impact the mean [21]. This section describes how the $K$ medoids clustering framework introduced by [21] can be adapted for categorical time series.

Let $\boldsymbol{\phi} = (\phi_1, \phi_2, \ldots, \phi_K)$ be the collection of indices for the time series serving as medoids for clusters $1, 2, \ldots, K$ and $\widehat{\boldsymbol{\phi}}$ be a corresponding estimate. Cluster features would then be estimated by directly using estimates from the medoid time series

$$\frac{\hat{\lambda}^{(k)}(\omega_j)}{||\hat{\lambda}^{(k)}||_2} = \frac{\hat{\lambda}_{\hat{\phi}_k}(\omega_j)}{||\hat{\lambda}_{\hat{\phi}_k}||_2}, \ \frac{\hat{\gamma}_s^{(k)}(\omega_j)}{||\hat{\gamma}^{(k)}||_2} = \frac{\hat{\gamma}_{\hat{\phi}_k,s}(\omega_j)}{||\hat{\gamma}_{\hat{\phi}_k}||_2}$$

for $k = 1, \ldots, K$. To further mitigate the impact of outliers, the $L_1$ distance is used

$$\tilde{d}_{i,k} = \sum_{j=1}^J \left| \frac{\hat{\lambda}_i(\omega_j)}{||\hat{\lambda}_i||_2} - \frac{\lambda^{(k)}(\omega_j)}{||\lambda^{(k)}||_2} \right| + \sum_{j=1}^J \sum_{s=1}^{S-1} \left| \frac{\hat{\gamma}_{i,s}(\omega_j)}{||\hat{\gamma}_i||_2} - \frac{\gamma_s^{(k)}(\omega_j)}{||\gamma^{(k)}||_2} \right|. \tag{5}$$

This distance measure can then be used in (4), and this optimization problem can be solved by the iterative algorithm presented in Algorithm 2. Similar to K-means, multiple initializations should be considered for the K-medoids approach as well to avoid settling at local optima. The following example illustrates a data setting where such an approach is advantageous.

**Example 2. Four clusters with outliers.** Realizations of categorical time series are again generated from the multinomial logit model [13], and let the number of categories $S = 4$, the number of clusters $K = 4$, and cluster-specific regression parameters, $\alpha_{s,k}$ $s = 1, \ldots, S$ and $k = 1, \ldots, K$ be the same as in Example 1.

Consider 40 realizations composed of 10 independent realizations from each cluster. For each cluster, 2 of the 10 realizations are considered as outliers and their cluster-specific regression parameters are altered such that $\widetilde{\alpha}_{s,k} = \alpha_{s,k} + Z$ where $Z \sim U(0, 1)$. This setting mimics practical situations in which a few outlier series exhibit features deviating from the typical behavior of the cluster and can negatively impact the performance of standard K-means clustering.

Finite sample performance for K-means and K-medoids approaches on this example are evaluated in Section 3.5.

## 3.4 Feature selection

In practice, it is of particular interest to identify the features most responsible for distinguishing clusters. Perhaps clusters are well-separated for certain frequency bands, or, as in Examples 1 and 2, some clusters are distinguishable across a particular subset of features while other clusters are distinguishable across an entirely different subset of features. In these cases, clustering accuracy could be improved by focusing on important features for which clusters are well-separated. Following [46], a sparse clustering algorithm for categorical time series is introduced in this section for simultaneous clustering and feature selection.

**Result:** Cluster assignments, $\hat{\mathbf{g}} = (\hat{g}_1, \ldots, \hat{g}_N)$, medoids $\widehat{\boldsymbol{\phi}} = (\phi_1, \ldots, \phi_K)$, and features, $\hat{\lambda}_{\hat{\phi}_k}(\omega_j)/||\hat{\lambda}_{\hat{\phi}_k}||_2, \hat{\gamma}_{\hat{\phi}_k,s}(\omega_j)/||\hat{\gamma}_{\hat{\phi}_k}||_2$ for $k = 1, \ldots, K,\ j = 1, \ldots, J,\ s = 1, \ldots, S-1$.

Initialize $\hat{\boldsymbol{\phi}}$ by independently drawing $K$ of the time series with equal probability and compute distances

$$\tilde{d}_{i,k} \leftarrow \sum_{j=1}^{J} \left| \frac{\hat{\lambda}_i(\omega_j)}{||\hat{\lambda}_i||_2} - \frac{\hat{\lambda}_{\hat{\phi}_k}(\omega_j)}{||\hat{\lambda}_{\hat{\phi}_k}||_2} \right| + \sum_{j=1}^{J} \sum_{s=1}^{S-1} \left| \frac{\hat{\gamma}_{i,s}(\omega_j)}{||\hat{\gamma}_i||_2} - \frac{\hat{\gamma}_{\hat{\phi}_k}(\omega_j)}{||\hat{\gamma}_{\hat{\phi}_k}||_2} \right|$$

for $i = 1, \ldots, N, k = 1, \ldots, K$.

Initialize cluster assignments $\hat{g}_i \leftarrow \arg\min_k \tilde{d}_{i,k}$, for $i = 1, \ldots, N$.

Initialize total cost $\widehat{D} = \sum_{k=1}^{K} \sum_{i=1}^{N} I(\hat{g}_i = k)\tilde{d}_{i,k}$

stop $\leftarrow 0$

**while** *stop=0* **do**

  $D_0 \leftarrow \widehat{D}$

  Evaluate swaps

  **for** $k = 1;\ k \leq K;\ k++$ **do**

    **for** $i = 1;\ i \leq N;\ i++$ **do**

      $\tilde{\boldsymbol{\phi}} \leftarrow \widehat{\boldsymbol{\phi}},\ \tilde{\phi}_k \leftarrow i$

      Obtain new distances

$$\tilde{d}_{i,k} \leftarrow \sum_{j=1}^{J} \left| \frac{\hat{\lambda}_i(\omega_j)}{||\hat{\lambda}_i||_2} - \frac{\hat{\lambda}_{\tilde{\phi}_k}(\omega_j)}{||\hat{\lambda}_{\tilde{\phi}_k}||_2} \right| + \sum_{j=1}^{J} \sum_{s=1}^{S-1} \left| \frac{\hat{\gamma}_{i,s}(\omega_j)}{||\hat{\gamma}_i||_2} - \frac{\hat{\gamma}_{\tilde{\phi}_k}(\omega_j)}{||\hat{\gamma}_{\tilde{\phi}_k}||_2} \right|$$

      for $i = 1, \ldots, N, k = 1, \ldots, K$.

      Obtain new cluster assignments

      $\tilde{g}_i \leftarrow \arg\min_k \tilde{d}_{i,k}$, for $i = 1, \ldots, N$.

      Compute new total cost

      $\widetilde{D} = \sum_{k=1}^{K} \sum_{i=1}^{N} I(\tilde{g}_i = k)\tilde{d}_{i,k}$

      **if** $\widetilde{D} < \widehat{D}$ **then**

      $\widehat{D} \leftarrow \widetilde{D}, \widehat{\boldsymbol{\phi}}_{new} \leftarrow \tilde{\boldsymbol{\phi}}, \hat{\mathbf{g}}_{new} \leftarrow \tilde{\mathbf{g}}$

    **end**

  **end**

  **if** $\widehat{D} = D_0$ **then** stop $\leftarrow 1$

  **else** $\widehat{\boldsymbol{\phi}} \leftarrow \widehat{\boldsymbol{\phi}}_{new}, \hat{\mathbf{g}} \leftarrow \hat{\mathbf{g}}_{new}$

**end**

**return** $\hat{\mathbf{g}}, \widehat{\boldsymbol{\phi}}, \hat{\lambda}_{\hat{\phi}_k}(\omega_j)/||\hat{\lambda}_{\hat{\phi}_k}||_2, \hat{\gamma}_{\hat{\phi}_k,s}(\omega_j)/||\hat{\gamma}_{\hat{\phi}_k}||_2$

**Algorithm 2:** K-MEDOIDS CLUSTERING

Let

$$W_{J\times S} = \left[ \{w_j^{(\lambda)}\}_{j=1}^{J}, \{w_{j,1}^{(\gamma)}\}_{j=1}^{J}, \ldots, \{w_{j,S-1}^{(\gamma)}\}_{j=1}^{J} \right]$$

be a collection of non-negative feature-specific weights. Given a fixed set of weights, we can construct a weighted version of the previous distance measure (3)

$$d_{i,k}(W) = \sum_{j=1}^{J} w_j^{(\lambda)} \left( \frac{\hat{\lambda}_i(\omega_j)}{||\hat{\lambda}_i||_2} - \frac{\lambda^{(k)}(\omega_j)}{||\lambda^{(k)}||_2} \right)^2 + \sum_{j=1}^{J} \sum_{s=1}^{S-1} w_{j,s}^{(\gamma)} \left( \frac{\hat{\gamma}_{i,s}(\omega_j)}{||\hat{\gamma}_i||_2} - \frac{\gamma_s^{(k)}(\omega_j)}{||\gamma^{(k)}||_2} \right)^2. \tag{6}$$

This enhancement allows (6) to be more sensitive to differences in features with larger weights while reducing or eliminating sensitivity to features with smaller or zero weights respectively. Similar theoretical results can be proved given a slightly modified assumption regarding cluster separation.

**Assumption 4** (Weighted Separation)**.** *Given a set of weights, cluster features are well-separated such that*

$$\sum_{j=1}^{J} w_{\lambda j} \left( \frac{\lambda^{(k)}(\omega_j)}{||\lambda^{(k)}||_2} - \frac{\lambda^{(k')}(\omega_j)}{||\lambda^{(k')}||_2} \right)^2 + \sum_{j=1}^{J} \sum_{s=1}^{S-1} w_{\gamma sj} \left( \frac{\gamma_s^{(k)}(\omega_j)}{||\gamma^{(k)}||_2} - \frac{\gamma_s^{(k')}(\omega_j)}{||\gamma^{(k')}||_2} \right)^2 \geq CT$$

*for a positive constant* $C$ $\forall k, k' \in \{1, 2, \ldots, K\}$.

Assumption 4 implies that cluster features must be well-separated specifically along features with non-zero weights. This is generally a reasonable assumption, but it depends on how weights are determined and may be violated in situations where weights are unduly sparse. The next result states the consistency for using (6) for clustering.

**Corollary 1.** *Let* $X_{it}$ *be a time series belonging to cluster* $k^*$*. Under Assumptions 1, 2, and 4, the probability of the weighted distance between* $X_{it}$ *and cluster* $k^*$ *exceeding the weighted distance between* $X_{it}$ *and a different cluster* $k \neq k^*$ *is bounded such that*

$$P(d_{i,k^*}(W) > d_{i,k}(W)) = O(B_T T^{-1}) \text{ for } \forall k \neq k^*.$$

*where* $d_{i,k}(W)$ *is defined in* (6)*.*

It remains to introduce a data-driven framework for determining appropriate feature weights. Following Section 3.2, replacing $d_{i,k}$ with $d_{i,k}(W)$ in the standard K-means formulation (4) and imposing sparsity-inducing regularization constraints on the weights [46] yields

$$\begin{aligned} (\hat{\mathbf{g}}, \hat{W}) = \arg\min_{(\mathbf{g},W)} & \sum_{i=1}^{N} \sum_{k=1}^{K} I(g_i = k) d_{i,k}(W) \\ \text{subject to } & ||\operatorname{vec} W||_2^2 \leq 1, ||\operatorname{vec} W||_1 \leq r, \text{ and} \\ & w_j^{(\lambda)} \geq 0, w_{j,s}^{(\gamma)} \geq 0\ \forall j, s, \end{aligned} \tag{7}$$

where $r$ is a tuning parameter that determines the level of sparsity in the weights. However, the objective function

in (7) is minimized by setting all weights to zero, which is not an interesting solution. Instead, recognizing that traditional K-means clustering (4) seeks to maximize between-cluster sum-of-squares [46], the problem can be reformulated as

(8)
$$(\hat{\mathbf{g}}, \hat{W}) = \arg\max_{(\mathbf{g},W)} \sum_{j=1}^{J} w_j^{(\lambda)} \left[ N^{-1} \sum_{i=1}^{N} \sum_{i'=1}^{N} d_{i,i',j}^{(\lambda)} - \sum_{k=1}^{K} N_k^{-1} \sum_{g_i=g_i'=k} d_{i,i',j}^{(\lambda)} \right] + \sum_{j=1}^{J} \sum_{s=1}^{S-1} w_{j,s}^{(\gamma)} \left[ N^{-1} \sum_{i=1}^{N} \sum_{i'=1}^{N} d_{i,i',j,s}^{(\gamma)} - \sum_{k=1}^{K} N_k^{-1} \sum_{g_i=g_i'=k} d_{i,i',j,s}^{(\gamma)} \right]$$
subject to $||\operatorname{vec} W||_2^2 \leq 1, ||\operatorname{vec} W||_1 \leq r,$ and
$w_j^{(\lambda)} \geq 0, w_{j,s}^{(\gamma)} \geq 0 \; \forall j, s,$

where $N_k = \sum_{i=1}^{N} I(g_i = k)$ and

$$d_{i,i',j}^{(\lambda)} = \left( \frac{\hat{\lambda}_i(\omega_j)}{||\hat{\lambda}_i||_2} - \frac{\hat{\lambda}_{i'}(\omega_j)}{||\hat{\lambda}_{i'}||_2} \right)^2,$$
$$d_{i,i',j,s}^{(\gamma)} = \left( \frac{\hat{\gamma}_{i,s}(\omega_j)}{||\hat{\gamma}_i||_2} - \frac{\hat{\gamma}_{i',s}(\omega_j)}{||\hat{\gamma}_{i'}||_2} \right)^2.$$

Solutions to (8) will assign weight to each feature depending on its contribution towards the overall between-cluster sum-of-squares and may result in zero weights for some features. Optimizing (8) can be done in an iterative fashion by alternating between two steps until convergence.

1. Given fixed weights $W$, optimize the objective function in (8) with respect to $\mathbf{g}$ via standard K-means clustering with weighted distance (6).
2. Given fixed cluster centers and fixed cluster membership $\mathbf{g}$, optimize the objective function in (8) with respect to $W$ by assigning more weight to features with larger between-cluster sum-of-squares.

The solution to the convex optimization problem in the second step can be easily obtained via soft thresholding [46]. Algorithm 3 presents the complete iterative algorithm for sparse clustering of categorical time series. It is important to note that this framework can be further generalized. For example, K-medoids clustering can also be used in the first step above along with different distance measures, including a weighted version of (5), so long as the distance measure is additive over features [46].

In order to select the value of sparsity tuning parameter $r$, the permutation-based gap statistic of [46] can be easily adapted for this setting. The steps are outlined as follows.

1. Obtain $B$ permuted data sets by independently permuting observations within each feature.
2. For each candidate tuning parameter $r$, calculate $\text{Gap}(r) = \log O(r) - B^{-1} \sum_{b=1}^{B} \log O_b(r)$ where $O(r)$ is the value of the objective function in (8) for the original data set and $O_b(r)$ is the objective function in (8) for the $b$th permuted data set.
3. Choose $r^* = \arg\max_r \text{Gap}(r)$ or as the smallest $r$ such that $\text{Gap}(r)$ is within a standard deviation of $\log O_b(r)$ of $\text{Gap}(r^*)$.

**Result:** Cluster assignments, $\hat{\mathbf{g}} = (\hat{g}_1, \ldots, \hat{g}_N)$, features, $\frac{\hat{\lambda}^{(k)}(\omega_j)}{||\hat{\lambda}^{(k)}||_2}, \frac{\hat{\gamma}_s^{(k)}(\omega_j)}{||\hat{\gamma}^{(k)}||_2}$ and weights, $\hat{W}_{J\times S}$ for $k = 1, \ldots, K$, $j = 1, \ldots, J$, $s = 1, \ldots, S-1$, .
Initialize $\hat{g}$ by independently drawing each $\hat{g}_i$ from $\{1, \ldots, K\}$ with equal probability.
Initialize $\widehat{W}$ as $1/\sqrt{JS}$ for all $J \times S$ features.
stop $\leftarrow 0$
**while** *stop=0* **do**
  Given $\hat{W}$ fixed, update $\hat{\mathbf{g}}$, $\frac{\hat{\lambda}^{(k)}(\omega_j)}{||\hat{\lambda}^{(k)}||_2}$, and $\frac{\hat{\gamma}_s^{(k)}(\omega_j)}{||\hat{\gamma}^{(k)}||_2}$ via k-means clustering (Algorithm 1) using $d_{i,k}(\hat{W})$ defined in (6).
  Given $\hat{\mathbf{g}}$ fixed, update $\hat{W}$ via soft thresholding
  $$\operatorname{vec} \widetilde{W} = \frac{S[\mathbf{a}_+, \Delta]}{||S[\mathbf{a}_+, \Delta]||_2}$$
  where $x_+$ denotes the positive part of $x$, $S[x, c] = \text{sign}(x)(|x| - c)_+$, and $\mathbf{a} = [\mathbf{a}_\lambda', \mathbf{a}_{\gamma 1}', \ldots, \mathbf{a}_{\gamma S-1}']'$ where
  $$a_{\lambda j} = \frac{1}{N} \sum_{i=1}^{N} \sum_{i'=1}^{N} d_{i,i',j}^{(\lambda)} - \sum_{k=1}^{K} \frac{\sum_{\hat{g}_i=\hat{g}_{i'}=k} d_{i,i',j}^{(\lambda)}}{\sum_{i=1}^{N} I(\hat{g}_i = k)},$$
  $$a_{\gamma sj} = \frac{1}{N} \sum_{i=1}^{N} \sum_{i'=1}^{N} d_{i,i',j,s}^{(\gamma)} - \sum_{k=1}^{K} \frac{\sum_{\hat{g}_i=\hat{g}_{i'}=k} d_{i,i',j,s}^{(\gamma)}}{\sum_{i=1}^{N} I(\hat{g}_i = k)},$$
  for $j = 1, \ldots, J$ and $s = 1, \ldots, S-1$. Let $\Delta = 0$ if that results in $||\operatorname{vec} \widetilde{W}||_1 < r$ where $r$ is a sparsity tuning parameter. Otherwise, choose $\Delta > 0$ such that $||\operatorname{vec} \widetilde{W}||_1 = r$.
  **if** $||\operatorname{vec} \widetilde{W} - \operatorname{vec} \widehat{W}||_1 / ||\operatorname{vec} \widehat{W}||_1 < \epsilon$ *where* $\epsilon > 0$ *is small* **then** stop $\leftarrow 1$
  **else** $\hat{W} \leftarrow \tilde{W}$
**end**
**return** $\hat{\mathbf{g}}, \hat{W}, \hat{\lambda}^{(k)}(\omega_j)/||\hat{\lambda}^{(k)}||_2, \hat{\gamma}_s^{(k)}(\omega_j)/||\hat{\gamma}^{(k)}||_2$

**Algorithm 3:** Sparse k-means clustering

## 3.5 Finite sample performance

We now return to Example 1 to evaluate the performance of the proposed clustering and feature selection algorithms on this simulated example. Recall that this example consists of $N = 100$ time series of length $T = 200$ each belonging to one of $K = 4$ equally-sized clusters such that $N_k = 25$ for $k = 1, 2, 3, 4$. Figure 3 displays the relevant fit and performance details. The scree plot indicates a decrease in the rate of decline in within-cluster sum-of-squares after $K = 4$,

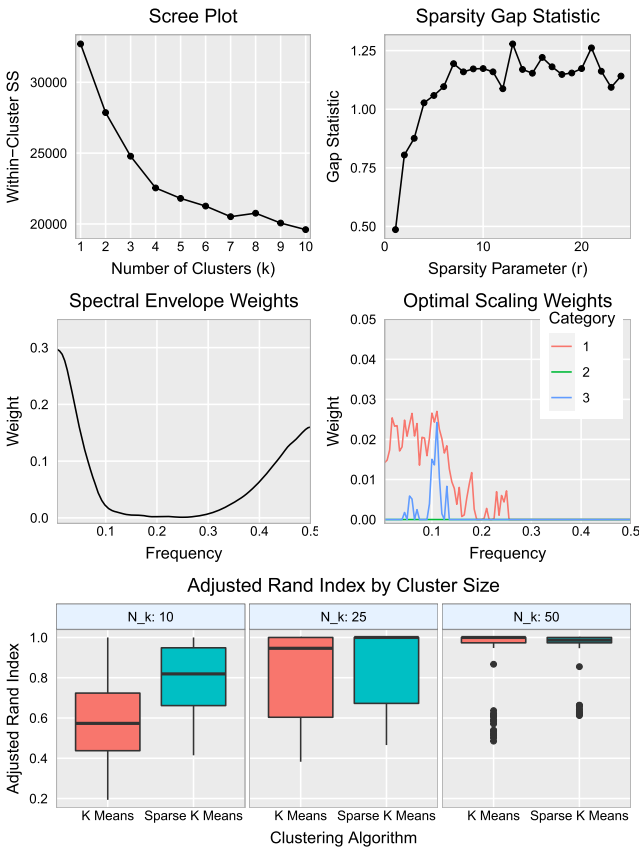


*Figure 3. Example 1 Performance: Scree plot and gap statistic plot for selecting $K$ and $r$ respectively (top), feature weights for the spectral envelope and optimal scalings (middle), and adjusted Rand index values for 100 replications of the data setting in Example 1 for different cluster sizes $N_k = 10, 25, 50$ (bottom). This figure appears in color in the electronic version of this article, and any mention of color refers to that version.*

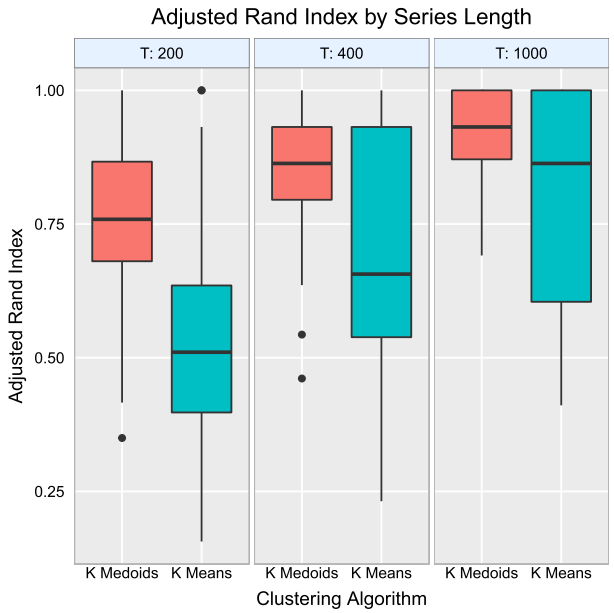


*Figure 4. Example 2 Performance: Adjusted Rand index values for 100 replications of the data setting in Example 2 for different time series lengths $T = 200, 400, 1000$ with 10 time series for each cluster. K-means and K-medoids are both used to evaluate cluster membership. This figure appears in color in the electronic version of this article, and any mention of color refers to that version.*

which matches our simulation setting. Using the gap statistic to consider various tuning parameter values between 1.1 and 25, $r = 7$ is the smallest $r$ with a gap statistic within one standard deviation of the maximum gap statistic and is used to determine feature weights via sparse clustering. With this choice of $r$, a larger proportion of the total weight is assigned to the spectral envelope (89.5%) compared to the optimal scalings (10.5%). The algorithm correctly assigns more weight to low ($< 0.1Hz$) and high ($> 0.3Hz$) frequency bands in the spectral envelope, which are largely responsible for distinguishing clusters 1 and 2. Also, the algorithm correctly assigns more weight to differences in the scalings for category 1, which are largely responsible for distinguishing clusters 3 and 4, while weights for other categories are mostly wiped out. This shows that the algorithm can still perform well even when all clusters are not well-separated along all features.

Finally, to evaluate clustering accuracy, both the standard and sparse K-means algorithms are applied to 100 replications of this data setting for different cluster sizes $N_k \in \{10, 25, 50\}$. The adjusted Rand index [36, 20] is used to assess clustering accuracy. Values of the adjusted Rand index close to 0 indicate poor estimation of the true cluster memberships and values close to 1 indicate good estimation of the true cluster memberships. The distributions of adjusted Rand index values are presented at the bottom of Figure 3. The mean adjusted Rand index values are 0.59, 0.82, and 0.92 for the proposed standard K-means algorithm for $N_k = 10, 25, 50$, and the mean adjusted Rand index values are 0.81, 0.90, and 0.94 for the proposed sparse K-means algorithm for $N_k = 10, 25, 50$. Notice that for $N_k = 10, 25$, the sparse K-means algorithm significantly outperforms the standard K-means algorithm in terms of the adjusted Rand index (p-values for Wilcoxon rank-sum tests of $<0.0001$ and 0.0002 respectively). This demonstrates the potential for the sparse K-means algorithm to exhibit superior finite-sample performance due to its simultaneous clustering and feature selection ability.

Next we turn to Example 2 to compare the proposed K-means and K-medoids algorithms in clustering categorical time series with outliers. Recall that this example consists of $K = 4$ clusters each of size $N_k = 10$ in which 2 of the 10 realizations are outliers generated by adding standard uniform noise to their cluster-specific regression parameters. Figure 4 displays the adjusted Rand index

values over 100 replications using the proposed K-means and K-medoids algorithms. Different lengths of time series $T = 200, 400, 1000$ are considered. The K-medoids algorithm significantly outperforms K-means for this data setting (p-values for Wilcoxon rank-sum tests below 0.004 for $T = 200, 400, 1000$) which demonstrates the improved finite-sample performance of K-medoids when outliers are present.

## 3.6 Clustering under fuzzy membership

In real data applications, it is entirely possible that some time series may exhibit characteristics that resemble multiple clusters. In this case, it can be helpful to relax the requirement of strict membership to a single cluster for each time series and allow for partial membership to multiple clusters, which is known as fuzzy clustering [1]. This phenomenon can be due to dynamic drifting or switching behavior over time that results in features with some resemblance to multiple clusters. If this behavior corresponds to a particular trajectory through time, methods that characterize nonstationary time series [5, 26, 17] would be more appropriate for use in clustering such series. However in many cases, this dynamic behavior may be vague and not tied to a particular course in time [31] and can be naturally treated via fuzzy clustering. For better illustration, consider the following example.

**Example 3. Two clusters with switching series.** Realizations of categorical time series are again generated from the multinomial logit model [13] as in Example 1. Let the number of categories $S = 4$, the number of clusters $K = 2$, and cluster-specific regression parameters

$$\alpha_{1,1} = (2.5, 1, 1)', \alpha_{1,2} = (1, 1, 1)', \alpha_{1,3} = (1, 1, 2, 5)',$$
$$\alpha_{2,1} = (1, 1, 2)', \alpha_{2,2} = (1, 1, 1)', \alpha_{2,3} = (2, 1, 1)'.$$

Consider 50 realizations of length $T = 400$ composed of 25 independent realizations from each cluster and an additional 5 realizations of switching time series such that $\alpha_1 = (2.5, 1, 1)'$, $\alpha_2 = (1, 1, 1)'$, $\alpha_3 = (1, 1, 2, 5)'$ for $t = 1, \ldots, 300$ and $\alpha_1 = (1, 1, 2)', \alpha_2 = (1, 1, 1)', \alpha_3 = (2, 1, 1)'$ for $t = 301, \ldots, 400$. This setting mimics practical situations for which most series can be mapped to a single cluster, but some series exhibit similarities to multiple clusters. Figure 5 displays a single realization from each of the two clusters and the switching group along with estimated spectral envelopes and optimal scalings.

Let $U$ be an $N \times K$ membership matrix such that each element, $u_{i,k} \in [0, 1]$, represents the degree of membership of the $i$th time series to the $k$th cluster for $i = 1, \ldots, N$ and $k = 1, \ldots, K$. Then the fuzzy clustering solution is

$$\hat{U} = \arg\min_{U} \sum_{i=1}^{N} \sum_{k=1}^{K} u_{i,k}^{m} d_{i,k}$$
$$\text{subject to } \sum_{k=1}^{K} u_{i,k} = 1 \; \forall \, i, u_{i,k} \geq 0 \; \forall \, i, k, \tag{9}$$

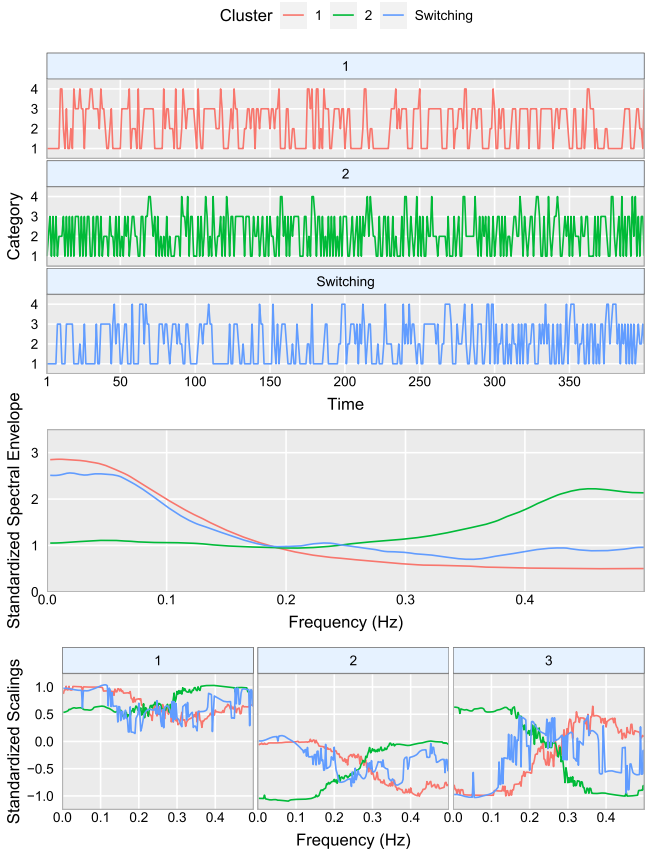


*Figure 5. Example 3 Visualization: $3 \times 1$ panel of simulated time series corresponding to two different clusters and switching series (top), estimated standardized spectral envelopes (middle), and estimated standardized optimal scalings for each category (bottom) for each cluster and switching group. This figure appears in color in the electronic version of this article, and any mention of color refers to that version.*

where $m > 1$ is a tuning parameter controlling the fuzziness of the cluster membership matrix such that $m$ close to 1 results in membership values, $u_{i,k}$, close to 0 and 1 and $u_{i,k} \to 1/K$ as $m \to \infty$. Following [31], this optimization problem can be solved using Lagrangian multipliers resulting in an iterative solution that updates cluster features and the membership matrix in turn until convergence. See Algorithm 4 for a detailed implementation of this approach. A fuzzy K-medoids algorithm can also be developed in a similar fashion, which is left to future work.

We now return to Example 3 to demonstrate how the fuzzy clustering algorithm performs for this simulated example. Figure 6 displays the scree plot and gap statistic plot for selecting $K$ and $m$ respectively. The scree plot indicates a decrease in the rate of decline in within-cluster sum-of-squares after $K = 2$, which matches our simulation setting. Using the gap statistic to consider various tuning parameter values between 1.02 and 7, $m = 1.7675$ is the

**Result:** Fuzzy cluster membership, $\widehat{U} = \{\hat{u}_{i,k}\}_{N\times K}$, and features, $\frac{\hat{\lambda}^{(k)}(\omega_j)}{||\hat{\lambda}^{(k)}||_2}, \frac{\hat{\gamma}_s^{(k)}(\omega_j)}{||\hat{\gamma}^{(k)}||_2}$ for $k = 1,\ldots,K$, $j = 1,\ldots,J$, $s = 1,\ldots,S-1$.

Initialize $\widehat{U}$ by independently drawing each $\hat{u}_{i,k}$ from a standard uniform random variable and dividing by the sum over $k = 1,\ldots,K$ for each $i = 1\ldots,N$.

stop $\leftarrow 0$

**while** *stop=0* **do**

Update cluster centers

$$\frac{\hat{\lambda}^{(k)}(\omega_j)}{||\hat{\lambda}^{(k)}||_2} \leftarrow \frac{\sum_{i=1}^N \hat{u}_{i,k}^m \frac{\hat{\lambda}_i(\omega_j)}{||\hat{\lambda}_i||_2}}{\sum_{i=1}^N \hat{u}_{i,k}^m},$$

$$\frac{\hat{\gamma}_s^{(k)}(\omega_j)}{||\hat{\gamma}^{(k)}||_2} \leftarrow \frac{\sum_{i=1}^N \hat{u}_{i,k}^m \frac{\hat{\gamma}_{i,s}(\omega_j)}{||\hat{\gamma}_i||_2}}{\sum_{i=1}^N \hat{u}_{i,k}^m},$$

for $j = 1,\ldots,J$ , $k = 1,\ldots,K$, and $s = 1,\ldots,S-1$ where $m$ is a fuzziness tuning parameter.

Update distances

$$d_{i,k} \leftarrow \sum_{j=1}^J \left(\frac{\hat{\lambda}_i(\omega_j)}{||\hat{\lambda}_i||_2} - \frac{\hat{\lambda}^{(k)}(\omega_j)}{||\hat{\lambda}^{(k)}||_2}\right)^2 + \sum_{j=1}^J \sum_{s=1}^{S-1} \left(\frac{\hat{\gamma}_{i,s}(\omega_j)}{||\hat{\gamma}_i||_2} - \frac{\hat{\gamma}_s^{(k)}(\omega_j)}{||\hat{\gamma}^{(k)}||_2}\right)^2$$

for $i = 1,\ldots,N, k = 1,\ldots,K$.

Update fuzzy cluster membership

$$\tilde{u}_{i,k} = \left[\sum_{k'=1}^K \left(\frac{d_{i,k}}{d_{i,k'}}\right)^{(S-1)^{-1}}\right]^{-1},$$

for $i = 1,\ldots,N, k = 1,\ldots,K$.

**if** $||\operatorname{vec}\widetilde{U} - \operatorname{vec}\widehat{U}||_1/||\operatorname{vec}\widehat{U}||_1 < \epsilon$ *where* $\epsilon > 0$ *is small* **then** stop $\leftarrow 1$

**else** $\widehat{U} \leftarrow \widetilde{U}$

**end**

**return** $\widehat{U}, \hat{\lambda}^{(k)}(\omega_j)/||\hat{\lambda}^{(k)}||_2, \hat{\gamma}_s^{(k)}(\omega_j)/||\hat{\gamma}^{(k)}||_2$

**Algorithm 4:** Fuzzy k-means clustering

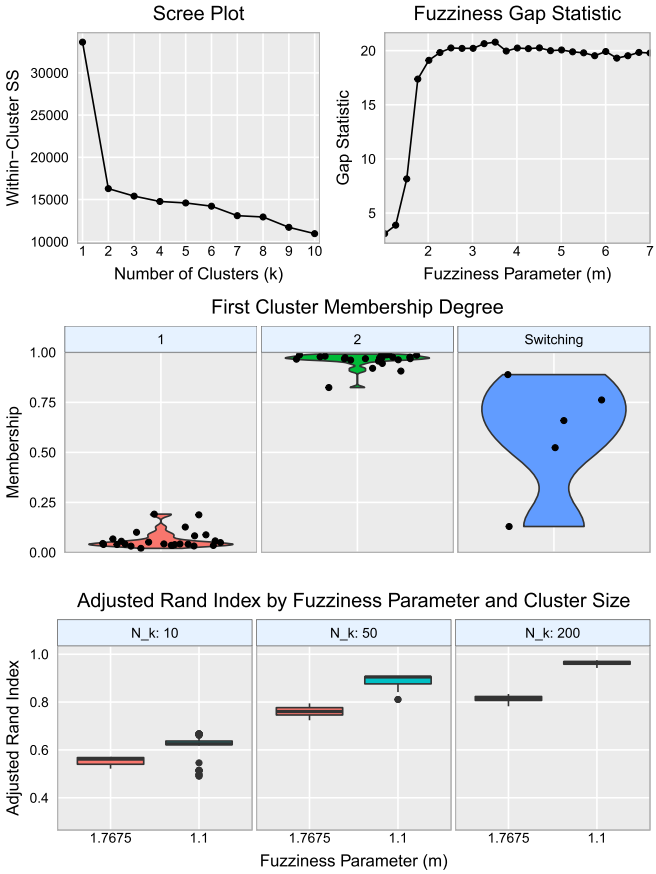


*Figure 6. Example 3 Performance: Scree plot and gap statistic plot for selecting $K$ and $m$ (top), first cluster membership degree by true membership to cluster 1, 2, and the switching group (middle), and adjusted Rand index for 100 replicates varying cluster size ($N_k$) and fuzziness parameter ($m$) while keeping the number of switching series the same (5) (bottom). This figure appears in color in the electronic version of this article, and any mention of color refers to that version.*

smallest $m$ with a gap statistic within one standard deviation of the maximum gap statistic and is used to allow for some fuzziness in the clustering results. Figure 6 also displays the membership degrees for the first cluster, $\hat{u}_{i,1}$, for observations from the first true cluster, second true cluster, and switching group. Observations generated from the two dominant clusters are correctly clustered together with membership values close to 0 and 1 while the switching series have membership values between 0.12 and 0.89, correctly indicating partial membership to both clusters.

A version of the adjusted Rand index for fuzzy clustering [7] is used to evaluate the effectiveness of the estimated fuzzy cluster memberships in recovering the true cluster assignments over 100 replications for different fuzziness parameters ($m = 1.7675, 1.1$) and cluster sizes ($N_k = 10, 50, 100$ and 5 switching series). As expected, performance improves as more time series are observed and a smaller fuzziness parameter is used, which reduces the fuzziness in the cluster assignments.

## 4. CLUSTER ANALYSIS OF SLEEP STAGE TIME SERIES

During a full night of sleep, the body cycles through two primary sleep stages, rapid eye movement (REM) sleep, in which dreaming typically occurs, and non-rapid eye movement (NREM) sleep, which consists of four stages representing light sleep (S1,S2) and deep sleep (S3,S4). These sleep stages are associated with specific physiological behaviors that are essential to the rejuvenating properties of sleep, and disruptions to typical cyclical behavior have been found to be associated with many sleep disorders [33]. It is of interest to determine if common profiles of sleep stage time

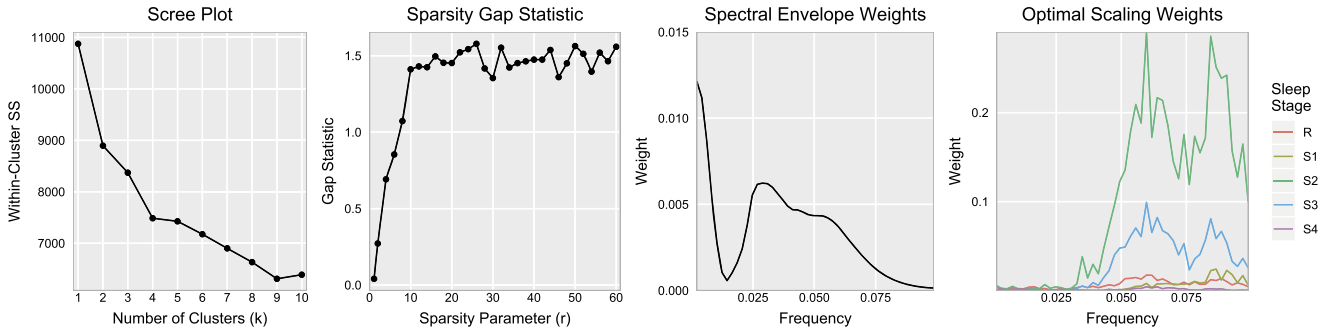


*Figure 7. Scree plot and gap statistic plot for selecting $K$ and $r$ respectively and feature weights for the spectral envelope and optimal scalings using sparse clustering for the application to sleep stage time series. This figure appears in color in the electronic version of this article, and any mention of color refers to that version.*

series exist within the population of sleep disorder patients that demonstrate distinct patterns of sleep disruption. The proposed clustering algorithms can provide tools for characterizing particular patterns indicative of sleep disorder, which can, in turn, aide in diagnosing sleep disorder and evaluating treatment efficacy.

The data for this analysis was collected through a study of various sleep-related disorders [43] and is publicly available via `Physionet` [16]. All participants were monitored during a full night of sleep and their sleep stages were annotated by experienced technicians every 30 seconds according to well-established sleep staging criteria [37]. The current analysis considers sleep stage time series from $N = 80$ participants: 38 patients with nocturnal frontal lobe epilepsy (NFLE), 18 patients with REM behavior disorder (RBD), 9 patients with periodic leg movements (PLM), 6 patients with insomnia (INS), and 9 control patients that did not present any neurological disorder. Time series consist of $S = 6$ sleep stages (REM, S1, S2, S3, S4, and Wake/Movement) with Wake/Movement used as the reference category. Examples are provided in Figure 1.

Since time series can exhibit nonstationary behavior associated with falling asleep at the beginning of the night and awakening at the end of the night, clustering was performed on subsets of the full night time series beginning at the 20th percentile of total sleep time and ending at the 90th percentile of total sleep time for each participant. Since sleep stage time series can vary in length, we follow [6, 30] and interpolate periodogram ordinates at the Fourier frequencies associated with the shortest time series in order to estimate the spectral envelope and optimal scalings. 89.5% of the power in the spectral envelope is contained in low frequencies ($< 0.1$) corresponding to sleep cycles lasting 5 minutes and longer, so only these frequencies are considered for clustering.

## 4.1 Sparse K-means clustering

Figure 7 contains the scree and gap statistic plots for choosing $K$ and $r$ as well as the feature weights determined by fitting the proposed sparse clustering algorithm to the data. Figure 8 displays the estimated spectral envelope and optimal scalings for each cluster. Table 1 illustrates the com-

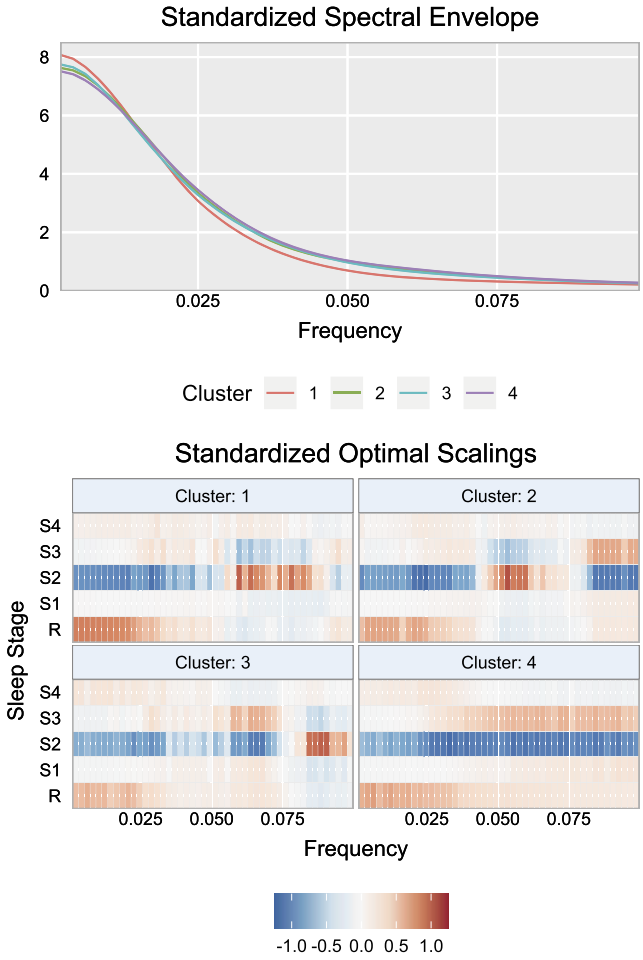


*Figure 8. Estimated standardized spectral envelope and optimal scalings for each cluster using sparse clustering for the application to sleep stage time series. This figure appears in color in the electronic version of this article, and any mention of color refers to that version.*

*Table 1. Cluster composition by sleep disorder for the application to sleep stage time series.*

| | Cluster | | | | |
|---|---|---|---|---|---|
| Group | 1 | 2 | 3 | 4 | Total |
| Control | 2 (0.18) | 3 (0.25) | 1 (0.06) | 3 (0.07) | 9 (0.11) |
| Insomnia | 2 (0.18) | 1 (0.08) | 1 (0.06) | 2 (0.05) | 6 (0.07) |
| NFLE | 4 (0.36) | 4 (0.33) | 9 (0.56) | 21 (0.51) | 38 (0.48) |
| PLM | 2 (0.18) | 0 (0.00) | 1 (0.06) | 6 (0.15) | 9 (0.11) |
| RBD | 1 (0.09) | 4 (0.33) | 4 (0.25) | 9 (0.22) | 18 (0.22) |
| Total | 11 (1.00) | 12 (1.00) | 16 (1.00) | 41 (1.00) | 80 (1.00) |

position of each cluster according to sleep disorder. While the clustering algorithm does not take into account sleep disorder types, it is of interest to see if particular types of sleep disorders dominate particular clusters. Classifying categorical time series using the spectral envelope and optimal scalings is an interesting direction of work with preliminary results available in [25].

The scree plot indicates a slight decrease in the rate of decline in within-cluster sum-of-squares after $K = 4$, so four clusters are used to characterize the population of sleep disorder patients. Using the gap statistic to consider various tuning parameter values between 1.1 and 60, $r = 10$ is the smallest $r$ with a gap statistic within one standard deviation of the maximum gap statistic and is used to determine feature weights via sparse clustering.

With this choice of $r$, a larger proportion of the total weight is assigned to the optimal scalings (97.7%) compared to the spectral envelope (2.3%), indicating that clusters exhibit more variability between clusters in their traversals through categories rather than in power across frequencies. This is expected as sleep disorders tend to disturb typical sleep cycles rather than fundamentally restructure sleep. Sleep is dominated by low frequency cycles lasting between 70 to 120 minutes [33]. Disruptions interrupt these dominant cycles, which shifts some power from lower frequencies to higher frequencies, but do not drastically attenuate low frequency power. This is further supported by the clustering results for which the estimated power in low frequencies ($< 0.02$), representing cycles lasting 25 minutes or longer, is dominant across all clusters and increases as the number of REM-behavior disorder patients, which typically experience more severe sleep disruption, in the cluster decreases (see Figure 8 and Table 1). On the other hand, sleep disruptions can have a major impact on cyclical traversals through categories depending on the nature of the disruption. This can result in significantly different scalings. For example, disruptions to REM sleep are common in patients with REM-behavior disorder (RBD) due to dream-enacting behavior [39], which can cause RBD patients to wake up abruptly [14]. However, patients with nocturnal frontal lobe epilepsy (NFLE) tend to experience seizures more often during NREM sleep [45] and are less likely to be awakened following a seizure [14]. These two examples illustrate different types of sleep disruptions to different sleep stages that can cause optimal scalings to take on different values.

Turning to Figure 8 and Table 1, we can investigate cluster profiles to better understand their distinguishing features and composition by sleep disorder types. According to the spectral envelope weights (see Figure 7), clusters are more distinguished in power for two frequency bands: frequencies less than or equal to 0.0167 (i.e. cycles lasting 30 minutes or longer) and frequencies between 0.0167 and 0.1 (i.e. cycles between 5 and 30 minutes). Looking to the cluster spectral envelopes in Figure 8, relatively less power in the spectral envelope for one band corresponds to relatively more power in the other. This means that clusters differ with respect to how much power in the spectral envelope is shifted from the low frequency band to the high frequency band due to sleep disturbances.

Turning to the cluster optimal scalings (Figure 8), the low frequency band is generally characterized by cycling between light sleep (S2) and REM sleep (R), which have larger scalings than other categories. This is expected as the observed time series spend the most time in these stages, with 40.6% and 20.4% of total sleep time on average being spent in light sleep (S2) and REM sleep (R) respectively. However, subtle differences exist across clusters. For example, the third cluster exhibits similar low frequency scalings for REM sleep (R) and deep sleep (S4). This can be attributed to more frequent disruptions to REM sleep for series belonging to this cluster, resulting in the lowest amount of time spent in REM sleep on average for this cluster (18.2%) compared to other clusters (ranging from 19.4% to 22.6%) and the most time spent on average in Wake/Movement (17.3%) compared to other clusters (ranging from 9.5% to 15.0%). This is not completely unexpected since this cluster has the lowest proportion of healthy controls (6%) and the highest proportion of patients with known sleep disorders (94%) compared to other clusters.

On the other hand, the high frequency band has significant differences in scalings across clusters, largely for light sleep (S2) and deep sleep (S3) categories according to the feature weights (see Figure 7). For example, the fourth cluster scalings for frequencies between 0.05 and 0.1, representing cycles lasting between 5 and 10 minutes, are associated with cycling between light sleep (S2) and either deep sleep (S3), lighter sleep (S1) or REM sleep (R). This cluster is

composed largely of NFLE patients (51%) for which nocturnal seizures may occur and typically disrupt NREM sleep. For comparison, the second cluster contains a larger proportion of healthy controls (25%), and scalings over this range consist of two sub-bands with different behavior. For frequencies between 0.05 and 0.075, representing cycles lasting between 6.67 and 10 minutes, scalings indicate dominant cycling between light sleep (S2) and deep sleep (S3). For frequencies between 0.075 and 0.1, representing cycles lasting between 5 and 6.67 minutes, scalings are similar to those of the fourth cluster and indicate dominant cycling between light sleep (S2) and either deep sleep (S3), lighter sleep (S1) or REM sleep (R). This suggests that nocturnal seizures associated with NFLE can disrupt the typical cycling between light and deep sleep for frequencies between 0.05 and 0.075. This is not completely unexpected as nocturnal seizures are more prevalent during NREM sleep [45], which may interfere with typical cyclical behavior. It is important to note that the third cluster also contains a large proportion (56%) of NFLE patients and exhibits different behavior from the fourth cluster, which is also majority NFLE (51%). This suggests potential heterogeneity within the population of NFLE patients and could be due to different seizure types, partial and generalized, affecting NFLE patients in different ways during the night [35].

### 4.2 Fuzzy K-means clustering

When clusters exhibit fairly similar behavior, which seems to be a reasonable assumption for this application, fuzzy clustering has been shown to have better adaptivity in singling out cluster structures [31]. Figure 9 displays the results of applying the proposed fuzzy clustering algorithm with $K = 4$ clusters to the sleep stage time series data described above. Using the gap statistic to consider various tuning parameter values between 1.02 and 10, $m = 1.39$ is the smallest $m$ with a gap statistic within one standard deviation of the maximum gap statistic and is used to allow for some fuzziness in the clustering results. Figure 9 also displays the membership degrees for each of the four clusters, $\hat{u}_{i,k}$, by sleep disorder type. Many membership degrees can be observed between 0.25 and 0.75, indicating partial membership to multiple clusters and supporting the notion that fuzzy clustering may be applicable for this dataset. The estimated spectral envelope and optimal scalings for each cluster exhibit many of the same characteristics already noted in the previous section and are thus omitted.

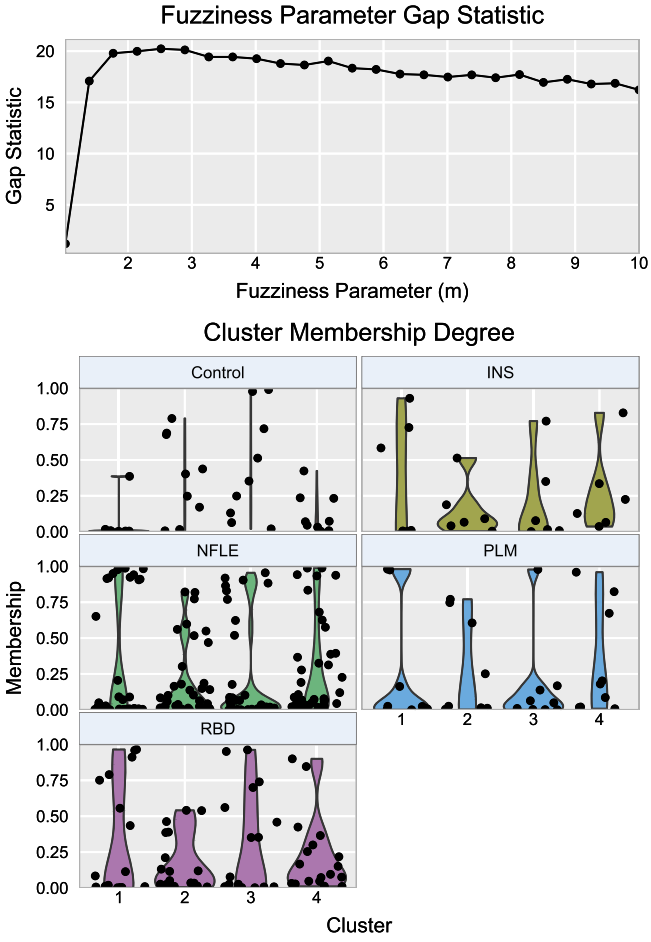


*Figure 9. Gap statistic plot for selecting $m$ and cluster membership degree by sleep disorder type for fuzzy clustering on the application to sleep stage time series. This figure appears in color in the electronic version of this article, and any mention of color refers to that version.*

## 5. CONCLUSION

This article presents a novel approach to clustering and feature selection for categorical time series via interpretable frequency-domain features. The proposed distance measure based on the spectral envelope and optimal scalings is shown to provide consistent clustering, and four algorithms are presented that tackle common challenges arising in practice, namely identifying important features and possible fuzziness in the clustering partition. All `R` code needed to reproduce the simulation and application results contained herein are available at the following GitHub repository https://github.com/sbruce23/envclust.

However, the proposed methods are not without limitations. First, these methods assume time series are stationary. However, in some applications, time series could be nonstationary, which would require time-varying extensions of the spectral envelope and optimal scalings for proper characterization. Incorporating nonstationarity may also further improve clustering accuracy. Second, the proposed methods assume time series within clusters share common cyclical patterns. However, extra variability may be present in some applications [22]. A topic of future research would be to incorporate within-group variability into the clustering framework. Finally, hierarchical clustering methods [10] could be developed to supplement the partitional clustering methods proposed in this work.

# APPENDIX A. PROOFS

## A.1 Preliminaries

To prove Theorem 1, the following lemmas are used.

**Lemma 1.** *Under Assumptions 1 and 2 and assume that $h(\omega)^{re}$ has distinct eigenvalues. Let $\lambda(\omega)$ and $\gamma(\omega)$ be the largest eigenvalue and corresponding eigenvector of $h(\omega)^{re}$. If $B_T \to \infty$ and $T \to \infty$ with $B_T T^{-1} \to 0$, then,*

$$\hat{\lambda}(\omega) - \lambda(\omega) = O(B_T T^{-1}),$$
$$\hat{\gamma}(\omega) - \gamma(\omega) = O(B_T T^{-1}).$$

Lemma 1 follows directly from [4, Theorems 9.4.1, 9.4.3].

**Lemma 2.** *Under Assumptions 1 and 2 and assume that $h(\omega)^{re}$ has distinct eigenvalues. Let $\lambda(\omega)$ and $\gamma(\omega)$ be the largest eigenvalue and corresponding eigenvector of $h(\omega)^{re}$ for $\omega \in (-1/2, 1/2]$. If $B_T \to \infty$ and $T \to \infty$ with $B_T T^{-1} \to 0$, then,*

$$||\hat{\lambda}||_2 - ||\lambda||_2 = O(B_T^{1/2} T^{-1/2}),$$
$$||\hat{\gamma}||_2 - ||\gamma||_2 = O(B_T^{1/2} T^{-1/2}).$$

*Proof.* Using of the result from Lemma 1, we have

$$\begin{aligned}
&||\hat{\lambda}||_2 \\
&= \sqrt{T^{-1} \sum_{j=1}^{J} |\hat{\lambda}(\omega_j)|^2} \\
&= \sqrt{T^{-1} \sum_{j=1}^{J} |\lambda(\omega_j) + O(B_T T^{-1})|^2} \\
&\leq \sqrt{T^{-1} \sum_{j=1}^{J} |\lambda(\omega_j)|^2 + O(B_T^{1/2} T^{-1/2})} \\
&\to \sqrt{\int_0^{1/2} |\lambda(\omega)|^2 d\omega + O(B_T^{1/2} T^{-1/2})} \\
&= ||\lambda||_2 + O(B_T^{1/2} T^{-1/2}).
\end{aligned}$$

Proof for $||\hat{\gamma}||_2$ follows similar steps and is thus omitted. □

**Lemma 3.** *Under Assumptions 1 and 2 and assume that $h(\omega)^{re}$ has distinct eigenvalues. Let $\lambda(\omega)$ and $\gamma(\omega)$ be the largest eigenvalue and corresponding eigenvector of $h(\omega)^{re}$ for $\omega \in (-1/2, 1/2]$. If $B_T \to \infty$ and $T \to \infty$ with $B_T T^{-1} \to 0$, then,*

$$\begin{aligned}
\frac{\hat{\lambda}(\omega)}{||\hat{\lambda}||_2} - \frac{\lambda(\omega)}{||\lambda||_2} &= O(B_T^{1/2} T^{-1/2}), \\
\frac{\hat{\gamma}_s(\omega)}{||\hat{\gamma}_s||_2} - \frac{\gamma_s(\omega)}{||\gamma||_2} &= O(B_T^{1/2} T^{-1/2}).
\end{aligned}$$

*Proof.* Using of the result from Lemmas 1 and 2, we have

$$\begin{aligned}
\frac{\hat{\lambda}(\omega)}{||\hat{\lambda}||_2} &= \frac{\lambda(\omega) + O(B_T T^{-1})}{||\lambda||_2 + O(B_T^{1/2} T^{-1/2})} \\
&= \frac{[\lambda(\omega) + O(B_T T^{-1})]/||\lambda||_2}{[||\lambda||_2 + O(B_T^{1/2} T^{-1/2})]/||\lambda||_2} \\
&= \frac{\lambda(\omega)/||\lambda||_2 + O(B_T T^{-1})}{1 + O(B_T^{1/2} T^{-1/2})} \\
&= [\lambda(\omega)/||\lambda||_2 + O(B_T T^{-1})][1 + O(B_T^{1/2} T^{-1/2})] \\
&= \frac{\lambda(\omega)}{||\lambda||_2} + O(B_T^{1/2} T^{-1/2}).
\end{aligned}$$

Since $|B_T^{1/2} T^{-1/2}| < 1$ for suitable $B_T$ (e.g. $B_T = \lfloor T^{1/2} \rfloor$). Proof for $\hat{\gamma}_s(\omega)/||\hat{\gamma}||_2$ follows similar steps and is thus omitted. □

## A.2 Proof of Theorem 1

Now that the asymptotic behavior of the standardized estimates of the spectral envelope and optimal scalings has been determined, the proof of this theorem uses Lemmas 1–3 and follows similar steps to those in the proofs of [25, Theorems 1–3].

*Proof.* Let $\hat{\lambda}(\omega_j)/||\hat{\lambda}||_2$ and $\hat{\gamma}_s(\omega_j)/||\hat{\gamma}||_2$ where $\omega_j = j/T$ for $j = 1, \ldots, J, J = \lfloor (T-1)/2 \rfloor$, and $s = 1, \ldots, S-1$ be the standardized estimates of the spectral envelope and optimal scalings respectively for time series $X_t$. Without loss of generality, assume $X_t$ belongs to the first cluster with standardized spectral envelope and optimal scalings features $\lambda^{(1)}(\omega_j)/||\lambda^{(1)}||_2$ and $\gamma_s^{(1)}(\omega_j)/||\gamma^{(1)}||_2$ respectively. Consider the distance measure

$$\begin{aligned}
&d_k^2 \\
&= \sum_{j=1}^{J} \left( \frac{\hat{\lambda}(\omega_j)}{||\hat{\lambda}||_2} - \frac{\lambda^{(k)}(\omega_j)}{||\lambda^{(k)}||_2} \right)^2 + \sum_{j=1}^{J} \sum_{s=1}^{S-1} \left( \frac{\hat{\gamma}_s(\omega_j)}{||\hat{\gamma}||_2} - \frac{\gamma_s^{(k)}(\omega_j)}{||\gamma^{(k)}||_2} \right)^2
\end{aligned}$$

for $k = 1, \ldots, K$ where $K =$ the number of clusters and $K$ is fixed. Consider the difference in the distance to the first cluster and the distance to a different cluster $k' \in \{2, \ldots, K\}$. It can be shown that

$$\begin{aligned}
&d_1^2 - d_{k'}^2 \\
&= -2 \sum_{j=1}^{J} \left( \frac{\hat{\lambda}(\omega_j)}{||\hat{\lambda}||_2} - \frac{\lambda^{(1)}(\omega_j)}{||\lambda^{(1)}||_2} \right) \left( \frac{\lambda^{(1)}(\omega_j)}{||\lambda^{(1)}||_2} - \frac{\lambda^{(k')}(\omega_j)}{||\lambda^{(k')}||_2} \right) \\
&\quad -2 \sum_{j=1}^{J} \sum_{s=1}^{S-1} \left( \frac{\hat{\gamma}_s(\omega_j)}{||\hat{\gamma}||_2} - \frac{\gamma_s^{(1)}(\omega_j)}{||\gamma^{(1)}||_2} \right) \left( \frac{\gamma_s^{(1)}(\omega_j)}{||\gamma^{(1)}||_2} - \frac{\gamma_s^{(k')}(\omega_j)}{||\gamma^{(k')}||_2} \right) \\
&\quad - \sum_{j=1}^{J} \left( \frac{\lambda^{(1)}(\omega_j)}{||\lambda^{(1)}||_2} - \frac{\lambda^{(k')}(\omega_j)}{||\lambda^{(k')}||_2} \right)^2
\end{aligned}$$

$$-\sum_{j=1}^{J}\sum_{s=1}^{S-1}\left(\frac{\gamma_s^{(1)}(\omega_j)}{||\gamma^{(1)}||_2}-\frac{\gamma_s^{(k')}(\omega_j)}{||\gamma^{(k')}||_2}\right)^2.$$

Using Chebyshev's inequality,

$$\begin{aligned}
&P(d_1^2-d_{k'}^2>0)\\
&\leq \mathrm{E}\left\{\left[-2\sum_{j=1}^{J}\left(\frac{\hat{\lambda}(\omega_j)}{||\hat{\lambda}||_2}-\frac{\lambda^{(1)}(\omega_j)}{||\lambda^{(1)}||_2}\right)\left(\frac{\lambda^{(1)}(\omega_j)}{||\lambda^{(1)}||_2}-\frac{\lambda^{(k')}(\omega_j)}{||\lambda^{(k')}||_2}\right)\right.\right.\\
&\left.\left.-2\sum_{j=1}^{J}\sum_{s=1}^{S-1}\left(\frac{\hat{\gamma}_s(\omega_j)}{||\hat{\gamma}||_2}-\frac{\gamma_s^{(1)}(\omega_j)}{||\gamma^{(1)}||_2}\right)\left(\frac{\gamma_s^{(1)}(\omega_j)}{||\gamma^{(1)}||_2}-\frac{\gamma_s^{(k')}(\omega_j)}{||\gamma^{(k')}||_2}\right)\right]^2\right\}\\
&\times\left\{\sum_{j=1}^{J}\left(\frac{\lambda^{(1)}(\omega_j)}{||\lambda^{(1)}||_2}-\frac{\lambda^{(k')}(\omega_j)}{||\lambda^{(k')}||_2}\right)^2\right.\\
&\left.+\sum_{j=1}^{J}\sum_{s=1}^{S-1}\left(\frac{\gamma_s^{(1)}(\omega_j)}{||\gamma^{(1)}||_2}-\frac{\gamma_s^{(k')}(\omega_j)}{||\gamma^{(k')}||_2}\right)^2\right\}^{-2}.
\end{aligned}$$

Using Lemma 3,

$$\begin{aligned}
&\mathrm{E}\left\{\left[-2\sum_{j=1}^{J}\left(\frac{\hat{\lambda}(\omega_j)}{||\hat{\lambda}||_2}-\frac{\lambda^{(1)}(\omega_j)}{||\lambda^{(1)}||_2}\right)\left(\frac{\lambda^{(1)}(\omega_j)}{||\lambda^{(1)}||_2}-\frac{\lambda^{(k')}(\omega_j)}{||\lambda^{(k')}||_2}\right)\right.\right.\\
&\left.\left.-2\sum_{j=1}^{J}\sum_{s=1}^{S-1}\left(\frac{\hat{\gamma}_s(\omega_j)}{||\hat{\gamma}||_2}-\frac{\gamma_s^{(1)}(\omega_j)}{||\gamma^{(1)}||_2}\right)\left(\frac{\gamma_s^{(1)}(\omega_j)}{||\gamma^{(1)}||_2}-\frac{\gamma_s^{(k')}(\omega_j)}{||\gamma^{(k')}||_2}\right)\right]^2\right\}\\
&=O(B_T T).
\end{aligned}$$

Assuming well-separated cluster features (Assumption 3),

$$\begin{aligned}
&\left\{\sum_{j=1}^{J}\left(\frac{\lambda^{(1)}(\omega_j)}{||\lambda^{(1)}||_2}-\frac{\lambda^{(k')}(\omega_j)}{||\lambda^{(k')}||_2}\right)^2\right.\\
&\left.+\sum_{j=1}^{J}\sum_{s=1}^{S-1}\left(\frac{\gamma_s^{(1)}(\omega_j)}{||\gamma^{(1)}||_2}-\frac{\gamma_s^{(k')}(\omega_j)}{||\gamma^{(k')}||_2}\right)^2\right\}^{-2}
\end{aligned}$$

is of the order $T^{-2}$ and thus $P(d_1^2>d_{k'}^2)=O(B_T T^{-1})$ for $k'\in\{2,\ldots,K\}$. □

## A.3 Proof of Corollary 1

*Proof.* Weights belong to the closed interval $[0,1]$. Since some weights can be exactly zero, Assumption 4 is needed to ensure cluster features are well-separated for the subset of components that have non-zero weights. Then, by following the steps of the proof for Theorem 1 and replacing the uniformly-weighted distance measure (3) with the weighted distance measure (6), we have the result. □




## ACKNOWLEDGEMENTS

Research reported in this publication was supported by the National Institute of General Medical Sciences of the National Institutes of Health (NIH) under Award Number R01GM140476. The content is solely the responsibility of the author and does not necessarily represent the official views of the NIH. The author thanks Dr. Pramita Bagchi and Dr. Zeda Li for their helpful discussions.

*Received 18 July 2021*


## REFERENCES


[1] Bezdek, J. C. (1981). Objective function clustering. In *Pattern Recognition with Fuzzy Objective Function Algorithms* 43–93. Springer. MR0631231

[2] Billingsley, P. (1961). *Statistical Inference for Markov Processes*. University of Chicago Press. MR0123419

[3] Brenčič, M. (2016). Statistical analysis of categorical time series of atmospheric elementary circulation mechanisms – Dzerdzeevski Classification for the northern hemisphere. *PLOS ONE* **11** 1–24.

[4] Brillinger, D. R. (2002). *Time Series: Data Analysis and Theory*. Philadelphia: SIAM. MR1853554

[5] Bruce, S. A., Hall, M. H., Buysse, D. J. and Krafty, R. T. (2018). Conditional adaptive Bayesian spectral analysis of nonstationary biomedical time series. *Biometrics* **74** 260–269. MR3777946

[6] Caiado, J., Crato, N. and Peña, D. (2009). Comparison of times series with unequal length in the frequency domain. *Communications in Statistics—Simulation and Computation* **38** 527–540. MR2649563

[7] Campello, R. J. (2007). A fuzzy extension of the Rand index and other related indexes for clustering and classification assessment. *Pattern Recognition Letters* **28** 833–841.

[8] Cardot, H., Lecuelle, G., Schlich, P. and Visalli, M. (2019). Estimating finite mixtures of semi-Markov chains: An application to the segmentation of temporal sensory data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **68** 1281–1303. MR4022813

[9] Dai, M. and Guo, W. (2004). Multivariate spectral analysis using Cholesky decomposition. *Biometrika* **91** 629–643. MR2090627

[10] Euán, C., Ombao, H. and Ortega, J. (2018). The hierarchical spectral merger algorithm: A new time series clustering procedure. *Journal of Classification* **35** 71–99. MR3790113

[11] Fahrmeir, L. and Kauifmann, H. (1987). Regression models for nonstationary categorical time series. *Journal of Time Series Analysis* **8** 147–160. MR0886137

[12] Fokianos, K. and Kedem, B. (1998). Prediction and classification of nonstationary categorical time series. *Journal of Multivariate Analysis* **67** 277–296. MR1659164

[13] Fokianos, K. and Kedem, B. (2003). Regression theory for categorical time series. *Statistical Science* **18** 357–376. MR2061915

[14] Foldvary-Schaefer, N. and Alsheikhtaha, Z. (2013). Complex nocturnal behaviors: Nocturnal seizures and parasomnias. *Continuum: Lifelong Learning in Neurology* **19** 104–131.

[15] García-Magariños, M. and Vilar, J. A. (2015). A framework for dissimilarity-based partitioning clustering of categorical time series. *Data Mining and Knowledge Discovery* **29** 466–502. MR3312468

[16] Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P., Mark, R., Mietus, J., Moody, G., Peng, C.-K. and Stanley, H. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation* **101** e215–e220.

[17] Hadj-Amar, B., Rand, B. F., Fiecas, M., Lévi, F. and Huckstepp, R. (2020). Bayesian model search for nonstationary peri-

odic time series. *Journal of the American Statistical Association* **115** 1320–1335. MR4143468
[18] Heiner, M. and Kottas, A. (2019). Estimation and selection for high-order Markov chains with Bayesian mixture transition distribution models. *ArXiv*. MR4387214
[19] Holan, S. H. and Ravishanker, N. (2018). Time series clustering and classification via frequency domain methods. *Wiley Interdisciplinary Reviews: Computational Statistics* **10** e1444. MR3873674
[20] Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification* **2** 193–218.
[21] Kaufman, L. and Rousseeuw, P. J. (1987). Clustering by means of medoids. *Statistical Data Analysis Based on the L1 Norm and Related Methods* 405–416.
[22] Krafty, R. T. (2016). Discriminant analysis of time series in the presence of within-group spectral variability. *Journal of Time Series Analysis* **37** 435–450. MR3512965
[23] Krafty, R. T. and Collinge, W. O. (2013). Penalized multivariate Whittle likelihood for power spectrum estimation. *Biometrika* **100** 447–458. MR3068445
[24] Krafty, R. T., Xiong, S., Stoffer, D. S., Buysse, D. J. and Hall, M. (2012). Enveloping spectral surfaces: Covariate dependent spectral analysis of categorical time series. *Journal of Time Series Analysis* **33** 797–806. MR2969912
[25] Li, Z., Bruce, S. A. and Cai, T. (2021). Classification of categorical time series using the spectral envelope and optimal scalings. *ArXiv*.
[26] Li, Z. and Krafty, R. T. (2019). Adaptive Bayesian time–frequency analysis of multivariate time series. *Journal of the American Statistical Association* **114** 453–465. MR3941268
[27] Liao, T. W. (2005). Clustering of time series data—a survey. *Pattern recognition* **38** 1857–1874.
[28] Lloyd, S. (1957). Least squares quantization in PCM Technical Report No. RR-5497, Bell Lab.
[29] MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* **1** 281–297. Oakland, CA, USA. MR0214227
[30] Maharaj, E. A., D'Urso, P. and Caiado, J. (2019). *Time Series Clustering and Classification*. CRC Press. MR3644715
[31] Maharaj, E. A. and D'Urso, P. (2011). Fuzzy clustering of time series in the frequency domain. *Information Sciences* **181** 1187–1211. MR2563306
[32] Milligan, G. W. and Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika* **50** 159–179.
[33] Institute of Medicine (2006). *Sleep Disorders and Sleep Deprivation: An Unmet Public Health Problem*. The National Academies Press, Washington, DC.
[34] Pamminger, C., Frühwirth-Schnatter, S. et al. (2010). Model-based clustering of categorical time series. *Bayesian Analysis* **5** 345–368. MR2719656
[35] Passouant, P. (1991). Historical aspects of sleep and epilepsy. In *Epilepsy, Sleep, and Sleep Deprivation. (Epilepsy Research Supplement)* 2nd ed. (R. Degen and E. A. Rodin, eds.) 19–22. Amsterdam: Elsevier Scientific Publishers, Amsterdam.
[36] Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association* **66** 846–850.
[37] Rechtschaffen, A. and Kales, A. (1968). *A Manual of Standardized Terminology, Techniques and Scoring System for Sleep Stages of Human Subjects*. US Government Printing Office, Washington DC.
[38] Rosen, O. and Stoffer, D. (2007). Automatic estimation of multivariate spectra via smoothing splines. *Biometrika* **94** 335–345. MR2331489
[39] Schenck, C. H., Bundlie, S. R., Ettinger, M. G. and Mahowald, M. W. (1986). Chronic behavioral disorders of human REM sleep: A new category of parasomnia. *Sleep* **9** 293–308.
[40] Shumway, R. H. and Stoffer, D. (2016). *Time Series Analysis and its Applications*, 4th ed. Springer: New York. MR3642322
[41] Stoffer, D., Tyler, D. and McDougall, A. J. (1993). Spectral analysis for categorical time series: Scaling and the spectral envelope. *Biometrika* **80** 611–632. MR1248026
[42] Stoffer, D. S., Tyler, D. E. and Wendt, D. A. (2000). The spectral envelope and its applications. *Statist. Sci.* **15** 224–253. MR1820769
[43] Terzano, M., Parrino, L., Sherieri, A., Chervin, R., Chokroverty, S., Guilleminault, C., Hirshkowitz, M., Mahowald, M., Moldofsky, H., Rosa, A., Thomas, R. and Walters, A. (2001). Atlas, rules, and recording techniques for the scoring of cyclic alternating pattern (CAP) in human sleep. *Sleep Med* **2** 537–553.
[44] Tibshirani, R., Walther, G. and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **63** 411–423. MR1841503
[45] Touchon, J., Baldy-Moulinier, M., Billiard, M., Besset, A. and Cadilhac, J. (1991). Sleep organization and epilepsy. In *Epilepsy, Sleep, and Sleep Deprivation. (Epilepsy Research Supplement)* 2nd ed. (R. Degen and E. A. Rodin, eds.) 73–81. Amsterdam: Elsevier Scientific Publishers, Amsterdam.
[46] Witten, D. M. and Tibshirani, R. (2010). A framework for feature selection in clustering. *Journal of the American Statistical Association* **105** 713–726. MR2724855

Scott A. Bruce
Texas A&M University
Department of Statistics
3143 TAMU
College Station, TX 77843, USA
E-mail address: sabruce@stat.tamu.edu