

A New Distance Measure for Model-Based Sequence Clustering

Darío García-García, Emilio Parrado-Hernández, *Member, IEEE*,
and Fernando Díaz-de-María, *Member, IEEE*

Abstract—We review the existing alternatives for defining model-based distances for clustering sequences and propose a new one based on the Kullback-Leibler divergence. This distance is shown to be especially useful in combination with spectral clustering. For better performance in real-world scenarios, a model selection scheme is also proposed.

Index Terms—



1 INTRODUCTION

ONE of the most usual assumptions in Machine Learning is that the observation vectors are independent and identically distributed (i.i.d.). This is reasonably true in a wide range of scenarios and provides useful simplifications that enable the development of efficient learning algorithms. Nonetheless, there are lots of problems where this assumption is far from being valid. For example, sometimes the useful information is encoded in the way data vectors evolve along time, so emphasis is required in modeling the system dynamics. Clearly, this can not be optimally done from an i.i.d. perspective: it requires a sequential approach, where the minimal meaningful unit is not a data vector but a sequence of vectors. Moreover, each of these sequences can have a different length, this being an additional difficulty for the traditional machine learning methods, which mostly rely on comparing patterns living in the same vector space.

A first step towards the development of efficient machine learning techniques to address these problems is obtaining an adequate modeling that enables pattern comparison. There has been a lot of research in generative models for sequential data, some of the most well-known and successful paradigms being the hidden Markov models (HMM) [1] and their extensions: hierarchical HMMs [2], buried Markov models [3], etc. They offer a good trade-off between computational complexity and expressive power, while at the same time being adequate models for lots of real-life processes.

In this work we address the problem of clustering sequential data. Clustering is one of the most common and useful tasks in machine learning, so it is a well-studied problem. Many efficient algorithms exist for the usual case of equal-length feature vectors, and amongst them Spectral clustering [4] stands out as a state-of-the-art non-parametric technique that allows for unsupervised classification without making any assumption about the

underlying distribution of data. Hierarchical clustering [5] is another widely used technique, especially when real hierarchical relations exist in the data. That being the case, results obtained by this clustering procedure can be highly descriptive and informative.

In order to apply these well-known clustering methods to the sequential data scenario we must face the difficulty of defining a meaningful distance measure for sequences. A popular framework is to first generate adequate models for the individual sequences in the dataset, and then use these models to obtain likelihood-based distances between sequences [6]. Based on this work, several other distance measures based on a likelihood matrix have been proposed [7], [8], [9], all of them being very similar in their philosophy.

We propose to explore a different approach and define a distance measure between sequences under the aforementioned framework by looking at the likelihood matrix from a probabilistic perspective. We regard the patterns created by the likelihoods of each of the sequences under the trained models as samples from the conditional likelihoods of the models given the data. This point of view differs largely from the existing distances. One of its differentiating properties is that it embeds information from the whole dataset or just a subset of it into each pairwise distance between sequences. This gives rise to highly structured distance matrices, which can be exploited by spectral methods to give a very high performance in comparison with previous proposals. Moreover, we also tackle the issue of selecting an adequate representative subset of models, proposing a simple method for that purpose when using spectral clustering. This greatly increase the quality of the clustering in those scenarios where the underlying dynamics of the sequences do not adhere well to the employed models.

The rest of this paper is organized as follows: In Section 2 we review the general framework for clustering sequential data, together with the most employed tools within that framework, namely HMMs as generative

• The authors are with the Signal Theory and Communications Department, University Carlos III of Madrid, Leganés, Spain.

models and hierarchical and spectral clustering, whose main characteristics are briefly outlined. The existing algorithms under this framework are also reviewed. Section 3 introduces our proposal of a new distance measure between sequences. Performance comparisons are carried out in Section 4, using both synthetic and real-world data. Finally, Section 5 collects the main conclusions of this work and sketches some promising lines for future research.

2 A FRAMEWORK FOR CLUSTERING SEQUENTIAL DATA

The seminal work of Smyth [6] proposes a probabilistic model-based framework for sequence clustering. Given a dataset $\mathcal{S} = \{S_1, \dots, S_N\}$ of N sequences, it assumes that each of them is generated by a single model from a discrete pool. The main idea behind this framework is to model the individual sequences and then use the resulting models to obtain a length-normalized log-likelihood matrix \mathbf{L} , whose $i^{j\text{th}}$ element l_{ij} is defined as

$$l_{ij} = \log p_{ij} = \frac{1}{\text{length}(S_j)} \log f_{\mathbf{S}}(S_j; \theta_i), \quad 1 \leq i, j \leq N, \quad (1)$$

where S_j is the j^{th} sequence, θ_i is the model trained for the i^{th} sequence and $f_{\mathbf{S}}(\cdot; \theta_i)$ is the probability density function (pdf) over sequences according to model θ_i . Based on this matrix, a distance matrix \mathbf{D} can be obtained so that the original variable-length sequence clustering problem is reduced to a typical similarity-based one.

The following subsections will describe the most usual tools under this framework: HMMs as the individual sequence models and hierarchical and spectral clustering for the actual partitioning of the dataset. Then, the existing algorithms in the literature under this framework will be briefly addressed.

2.1 Hidden Markov Models

Hidden Markov models (HMMs) [1] are a kind of parametric discrete state-space models which are widely employed in signal processing and pattern recognition. Their success comes mainly from their relative low complexity compared with their expressive power and their ability to model naturally occurring phenomena. Its main field of application has traditionally been speech recognition [1], but they have also found success in a wide range of areas, from bioinformatics [10] to video analysis [11].

In an HMM, the (possibly multidimensional) observation \mathbf{y}_t at a given time instant t (living in a space \mathbf{Y}) is generated according to a conditional pdf $f_{\mathbf{Y}}(\mathbf{y}_t|q_t)$, with q_t being the hidden state at time t . These states follow a time-homogeneous first-order Markov chain, so that $P(q_t|q_{t-1}, q_{t-2}, \dots, q_0) = P(q_t|q_{t-1})$. Bearing this in mind, an HMM θ can be completely defined by the following parameters:

- The discrete and finite set of the K possible states $\mathcal{X} = \{x_1, x_2, \dots, x_K\}$
- The state transition matrix $\mathbf{A} = \{a_{ij}\}$, where each a_{ij} represents the probability of a transition between two states: $a_{ij} = P(q_{t+1} = x_j | q_t = x_i)$, $1 \leq i, j \leq K$
- The emission pdf $f_{\mathbf{Y}}(\mathbf{y}_t|q_t)$
- The initial probabilities vector $\pi = \{\pi_i\}$, where $1 \leq i \leq K$ and $\pi_i = P(q_0 = x_i)$

The parameters of an HMM are traditionally learnt using the Baum-Welch algorithm [1], which represents a particularization of the well-known Expectation-Maximization (EM) algorithm [12]. Its complexity is $O(K^2T)$ per iteration, with T being the length of the training sequence. A hidden Markov model can be seen as a simple Dynamic Bayesian Network (DBN) [13], an interpretation which provides an alternative way of training this kind of models by applying the standard algorithms for DBNs. This allows for a unified way of inference in HMMs and their generalizations.

2.2 Hierarchical clustering

Hierarchical clustering (HC) [5] algorithms try to organize the data into a hierarchical structure according to a distance matrix. This organization happens in an iterative fashion; at each step of the algorithm, existing clusters are either fused (agglomerative methods) or partitioned (divisive methods) according to a particular heuristic. The resulting structure can be depicted as a binary tree or dendrogram which can be cut at a certain level to produce the desired number of clusters.

Agglomerative methods start by assigning each single datum to a different cluster and then progressively proceed to refine this partition by merging similar clusters attending to the distances amongst them. On the other hand, divisive methods initially consider the whole dataset as a unique cluster and then recursively partition it in a way such that the resulting clusters are as distant as possible.

2.3 Spectral clustering

Spectral clustering (SC) [4] is a concept that includes various techniques which make use of the spectrum of an adjacency matrix W in order to cluster data. Its roots lie within the graph theory and involves obtaining an (in some sense) optimum partition of the weighted graph whose nodes represent the data instances and whose edges are given by W . This way, w_{ij} is the adjacency between the i^{th} and j^{th} data vectors.

Originally, this optimum partition was selected as the one that minimized the value of the cut, that is, the sum of the connections that are removed by the partition [4]. In [14] the authors proposed the use of the normalized cut, instead of just the cut, as a measure of the goodness of the partition. The normalized cut takes into account the ratio between the cut of a partition and the total connections of the generated clusters. This has a regularizing

effect, penalizing the creation of small isolated clusters and thus promoting more uniform partitions. Obtaining such a clustering is an NP-complete problem, but it can be relaxed and simplified to a generalized eigenvalue (GEV) problem based on the Laplacian matrix, followed by k -means or any other simple clustering algorithm on the resulting eigenvectors.

The spectral techniques have the additional advantage of providing a clear and well-founded way of determining the optimal number of clusters for a dataset, based on the eigengap of the similarity matrix [15].

2.4 Existing algorithms

The pioneering proposal for sequential data clustering of [6] aims at fitting a single generative model to the whole set \mathcal{S} of sequences. The clustering itself is part of the initialization procedure of the model. In this initialization step, each sequence S_i is modeled with a HMM θ_i . Then, the distance between two sequences S_i and S_j is defined based on the log-likelihood of each sequence given the model generated for the other sequence:

$$d_{SYM}^{ij} = \frac{1}{2} (l_{ij} + l_{ji}), \quad (2)$$

where l_{ij} , as in equation (1), represents the (length-normalized) log-likelihood of sequence S_j under model θ_i . This is actually the symmetrized distance previously proposed in [16]. Given these distances, the data is partitioned using agglomerative hierarchical clustering with the ‘‘furthest-neighbor’’ merging heuristic.

The work in [7] inherits this framework for sequence clustering but introducing a new dissimilarity measure, the BP metric:

$$d_{BP}^{ij} = \frac{1}{2} \left\{ \frac{l_{ij} - l_{ii}}{l_{ii}} + \frac{l_{ji} - l_{jj}}{l_{jj}} \right\}. \quad (3)$$

The BP metric takes into account how well a model represents the sequence it has been trained for, so it is expected to work better than the symmetrized distance in those cases where the quality of the models may vary along the different sequences.

Yet another alternative distance within this framework is proposed in [8], namely

$$d_{POR}^{ij} = |p_{ij} + p_{ji} - p_{ii} - p_{jj}|, \quad (4)$$

with p_{ij} as defined in eq. (1).

Recently, the popularity of spectral clustering has motivated some works in which this kind of techniques are applied to the clustering of sequences. In [9] the authors propose a distance measure resembling the BP metric

$$d_{YY}^{ij} = |l_{ii} + l_{jj} - l_{ij} - l_{ji}|, \quad (5)$$

and then apply spectral clustering on a similarity matrix derived from the distance matrix by means of a Gaussian kernel. This method has reported good results in comparison with traditional parametric methods using initializations such as the proposed in [6] or Dynamic Time Warping (DTW) [17].

Another example of applying spectral clustering to sequential data can be found in [18], where a novel approach involving similarity between the probability distributions defined by the different HMMs as measured by means of a probability product kernel (PPK) is presented. This way, the clustering is done directly in parameter space instead of in the usual likelihood space, so this method falls out of the scope of the present paper.

3 PROPOSED ALGORITHM

Our proposal is based on the observation that the aforementioned methods define the distance between two sequences S_i and S_j based solely on the probabilities of that sequences under the models trained with them (θ_i and θ_j). We consider that a better performance can be expected if we add into the distance some global characteristics of the dataset. Moreover, since distances under this framework are obtained from a likelihood matrix, it seems a good intuition to take the probabilistic nature of this matrix into account when selecting adequate distance measures.

Bearing this in mind, we propose a novel sequence distance measure based on the Kullback-Leibler (KL) divergence [19], which is a standard measure of the similarity between probability density functions.

The first step of our algorithm involves obtaining the likelihood matrix \mathbf{L} as in eq. (1) (we will assume at first that an HMM is trained for each sequence). The i^{th} column of \mathbf{L} represents the likelihood of the sequence S_i under each of the trained models. These models can be regarded as a set of ‘‘intelligently’’ sampled points from the model space, in the sense that they have been obtained according to the sequences in the dataset. This way, they are expected to lie in the area of the model space θ surrounding the HMMs that actually span the data space. Therefore, these trained models become a good discrete approximation $\hat{\theta} = \{\theta_1, \dots, \theta_N\}$ to the model subspace of interest. If we normalize the likelihood matrix so that each column adds up to 1, we get a new matrix \mathbf{L}_N whose columns can be seen as the probability density functions over the approximated model space conditioned on each of the individual sequences:

$$\mathbf{L}_N = \left[f_{\hat{\theta}}^{S_1}(\theta), \dots, f_{\hat{\theta}}^{S_N}(\theta) \right].$$

This interpretation leads to the familiar notion of dissimilarity measurement between probability density functions, the KL divergence being a natural choice for this purpose. Its formulation for the discrete case is as follows:

$$D_{KL}(f_P||f_Q) = \sum_i f_P(i) \log \frac{f_P(i)}{f_Q(i)}, \quad (6)$$

where f_P and f_Q are two discrete pdfs. Since the KL divergence is not a proper distance because of its asymmetry, a symmetrized version is used

$$D_{KLSYM}(f_P||f_Q) = \frac{1}{2} [D_{KL}(f_P||f_Q) + D_{KL}(f_Q||f_P)]. \quad (7)$$

This way, the distance between the sequences S_i and S_j can be defined simply as

$$d^{ij} = D_{KLSYM} \left(f_{\tilde{\theta}}^{S_i} || f_{\tilde{\theta}}^{S_j} \right). \quad (8)$$

This implies a change of focus from the probability of the sequences under the models to the likelihood of the models given the sequences. Distances defined this way are obtained according to the patterns created by each sequence in the probability space spanned by the different models. With this approach, the distance measure between two sequences S_i and S_j somehow involves information related to the rest of the data sequences, represented by their corresponding models.

This redundancy can be used to define a representative subset $\mathcal{Q} \subseteq \mathcal{S}$ of the sequences, so that $\tilde{\theta} = \{\theta_{Q_1}, \dots, \theta_{Q_P}\}$, $P \leq N$. This way, only the models trained with sequences belonging to \mathcal{Q} will be taken into account in the calculation of the distances, instead of using the whole dataset for that purpose. The advantage of defining such a subset is twofold: on the one hand, computational load can be reduced since the number of models to train is reduced to P and posterior probabilities calculations also drop from $N \times N$ to $P \times N$. On the other hand, if the dataset is prone to outliers or the models suffer from overfitting, the stability of the distance measures and thus the clustering performance can be improved if \mathcal{Q} is carefully chosen. Examples of both of these approaches will be shown in the experiments included in Section 4.

Finally, before applying a spectral clustering, the distance matrix $\mathbf{D} = \{d_{ij}\}$ must be transformed into a similarity matrix \mathbf{A} . A commonly used procedure is to apply a Gaussian kernel so that $a_{ij} = \exp\left(\frac{-d_{ij}^2}{2\sigma^2}\right)$, with σ being a free parameter representing the kernel width. Then, a standard normalized-cut algorithm is applied to matrix \mathbf{A} , resulting in the actual clustering of the sequences in the dataset. In the sequel, we will refer to this combination of our proposed KL-based distance and spectral clustering as KL+SC.

4 EXPERIMENTAL RESULTS

This section presents some experimental results concerning several synthetic and real-world sequence clustering problems. Synthetic data experiments aim at illustrating the performance of the different sequence clustering algorithms under tough separability conditions but fulfilling the assumption that the sequences are generated by hidden Markov models. This way, we focus the analysis on the impact of the distance measures as we isolate the adequateness of the modeling (except for overfitting). Besides, we also use real-world speech data to show a sample application of sequence clustering on a field where HMMs have been typically used as rough approximate generative models.

The different methods to be compared are the following:

SYM	Symmetrized distance (eq. (2))
BP	BP distance (eq. (3))
POR	Porikli distance (eq. (4))
YY	Yin - Yang distance (eq. (5))
KL	Proposed KL distance (eq. (refeq:KL))

All of them will be paired with both a normalized-cut spectral clustering algorithm and an agglomerative hierarchical one using the furthest-neighbor merging heuristic, as in [6]. For the spectral clustering algorithm, the value of parameter σ of the Gaussian kernel is selected empirically in a completely unsupervised fashion as the one that maximizes the eigengap for each distance measure in each case, as proposed in [15]. It is also remarkable that the k -means part of the spectral clustering algorithm, due to its strong dependence on the initialization, is run 10 times at each iteration, and we choose as the most adequate partition the one with the minimal intra-cluster distortion, defined as:

$$D_{cluster} = \sum_{k=1}^K \sum_{i \in \mathcal{C}_k} \|\mathbf{x}_i - \mathbf{c}_k\|^2,$$

where K is the number of clusters, \mathcal{C}_k is the set of the indexes of points belonging to the k^{th} cluster, \mathbf{c}_k is the centroid of that cluster and \mathbf{x}_i is the i^{th} data point. This distortion can be seen as the ‘‘tightness’’ of the resulting clusters, and it is also well known that this minimum distortion criterion implies a maximum separation amongst centroids [20].

Both the code and the datasets for the following experiments can be found at the author’s website¹.

4.1 Synthetic data

The first scenario under which the comparison is carried out is the original example from [6]: each sequence in the dataset is generated with equal probability by one of two possible HMMs θ_1 and θ_2 , each one of them having two hidden states ($m = 2$). Transition matrices for the generating HMMs are given by

$$\mathbf{A}_1 = \begin{pmatrix} 0.6 & 0.4 \\ 0.4 & 0.6 \end{pmatrix} \quad \mathbf{A}_2 = \begin{pmatrix} 0.4 & 0.6 \\ 0.6 & 0.4 \end{pmatrix}.$$

Initial states are equiprobables and emission probabilities are the same in both models, specifically $N(0, 1)$ in the first state and $N(3, 1)$ in the second. This scenario represents a very appropriate testbed for sequence clustering techniques, since the only way to differentiate sequences generated by each model is to attend to their dynamical characteristics. These, in turn, are very similar, making this a hard clustering task. The length of each individual sequence is obtained by sampling a uniform pdf in the range $[\mu_L(1 - V/100) \quad \mu_L(1 + V/100)]$, where μ_L is the sequence’s mean length and V is a parameter which we will refer to as the percentage of variation in the length. All the given results are averaged over 50 randomly generated datasets.

1. <http://www.tsc.uc3m.es/~dggarcia>

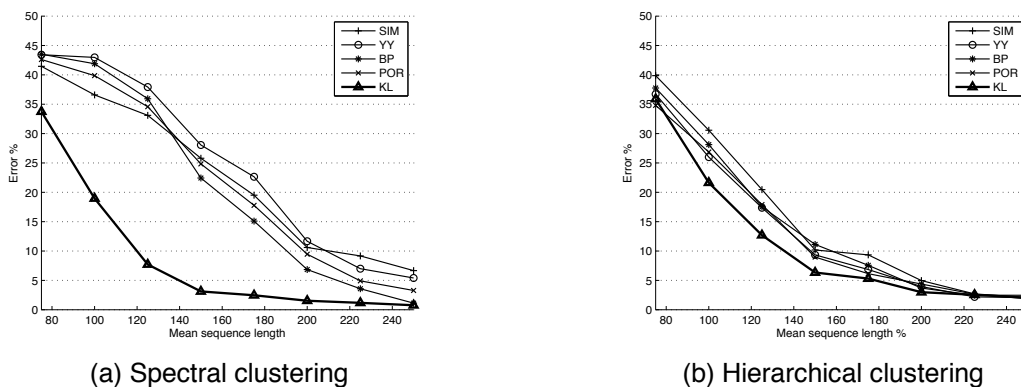


Fig. 1. Clustering error percentage achieved by the compared methods against different mean sequence lengths ($V = 40\%$, $N = 80$)

Figure 1 shows the results of the performance comparison of the different distance measures and clustering methods against variations of the mean length μ_L of the data sequences for a fixed length variation V of 40% in a dataset comprised of $N = 80$ sequences. It can be seen that, as expected, the longer the sequences the more accurate the clustering. It is also clear that our proposed distance measure outperforms the previous proposals under both hierarchical and spectral clustering, attaining specially good results using the latter technique. Specifically, the proposed KL+SC method yields the best performance for every mean sequence length, showing consistent improvements which are more dramatic for short mean sequence lengths ($\mu_L < 200$). Models trained with such short sequences suffer from severe overfitting, not being able to adequately capture the underlying dynamics and thus giving unrealistic results when evaluated using the sequences in the dataset. This results into incoherent distance matrices using the typical methods which renders the use of spectral clustering algorithms unproductive. Nonetheless, our proposal is more resilient against this issue since it takes a global view on the dataset that allows for the correct clustering of sequences even if the models are rather poor. Evaluating the sequences on a large enough number of individually inadequate models can generate patterns that our distance measure can capture, which translates into a consistent distance matrix very suitable for applying spectral methods. This shows that our approach is efficient even when the models are poor so they cannot be expected to correctly sample the model space. In these scenarios, the probabilistic interpretation of the proposed distance is not clear and it takes more of a pattern matching role.

Agglomerative hierarchical clustering is more forgiving of loosely structured distance matrix, since it merges clusters based on pairwise distance comparisons instead of taking a global view. This way, it seems more suitable than spectral clustering methods for its use with the previously proposed model-based sequence distances.

On the other hand, it also implies that it cannot benefit from the use of our proposed distance as much as spectral techniques can.

Figure 2 displays the evolution of the error along the number of sequences in the dataset. As more sequences are present in the dataset, the aforementioned problems of the previous proposals in combination with spectral clustering become clearer, while our method manages to improve its performance. Using hierarchical clustering, all the distances achieve stable results no matter the number of sequences, but once again this comes at the expense of an inferior performance compared to the KL+SC combination.

Figure 3 shows the results for a multi-class clustering with $K = 3$ classes. The sequences being clustered were generated using the two previously employed HMMs (θ_1 and θ_2) and a third one θ_3 that only differs from them in the transition matrix. Specifically,

$$\mathbf{A}_3 = \begin{pmatrix} 0.7 & 0.3 \\ 0.3 & 0.7 \end{pmatrix}.$$

The additional class makes this a harder problem than the two-class scenario, so it is logical that lengthier sequences are required to achieve comparable results. Nonetheless, the use of our proposed distance still shows significant improvements over the rest of the distances, all of which give almost identical results.

4.2 Real-world data

Now, the different sequential-data clustering algorithms will be evaluated using real sample problems. The chosen scenario is speaker clustering: we are given a set of audio files, each one of them containing speech from a single speaker, and the task is to group together files coming from the same speaker (two speakers per experiment). Two different databases are employed, namely:

- **AHUMADA** [21] A database specially tailored for speaker identification. We use a free subset² consisting of 25 speakers, and choose the isolated

2. <http://atvs.ii.uam.es/databases.jsp>

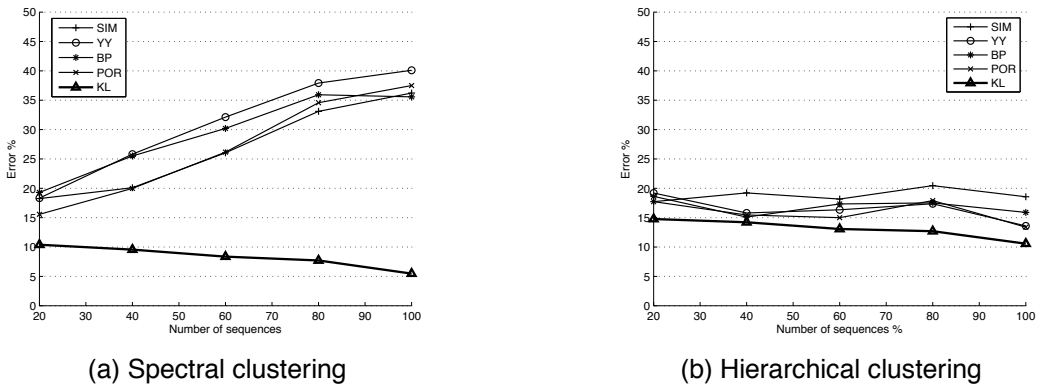


Fig. 2. Clustering error against number of sequences in the dataset ($\mu_L=125, V=40\%$)

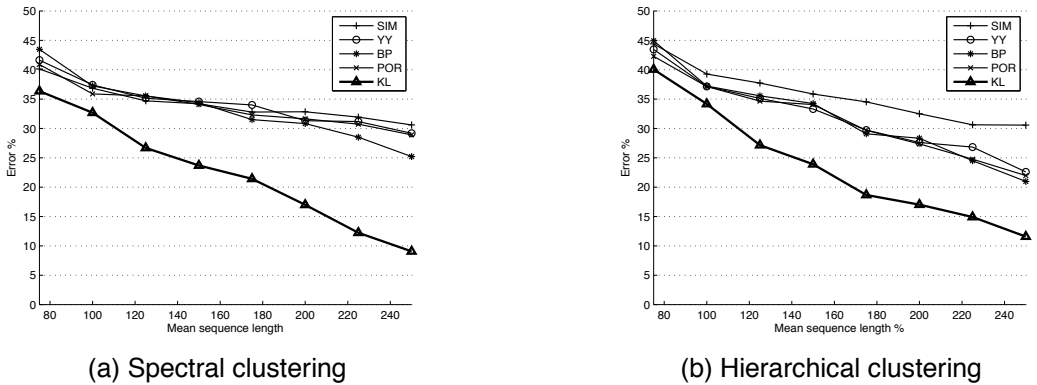


Fig. 3. Performance in a multiclass ($K = 3$) clustering task against different mean sequence lengths ($V=40\%, N=100$)

digits task: each speaker records 24 digits, which are further concatenated in groups of two, giving 12 sequences per user with a mean length of 0.7 seconds.

- **GPM-UC3M** A database recorded at the Multimedia Processing Group of the University Carlos III of Madrid using a PDA. It consists of 30 speakers with 50 isolated words for each one of them. Every single word is considered an individual sequence, and its mean length is around 1.3 seconds.

The audio files were processed using the freely available HTK software³, using a standard parametrization consisting of 12 Mel-frequency cepstral coefficients (MFCCs), an energy term and the increments (δ) of that coefficients, giving a total of 26 parameters. These parameters were obtained every 10ms with a 25ms analysis windows. The resulting 26-dimensional sequences were fed into the different clustering algorithms without any further processing.

Since these datasets are inherently noisy and the coefficient sequences don't perfectly fit a markovian generative model, the property of embedding information about all the sequences in each pairwise distance can become performance-degrading. Thus, it becomes inter-

esting to select only an adequate subset of the models for obtaining the distance matrix.

For this purpose, we propose a simple method to determine which models to include in the KL+SC method: firstly, since models coming from lengthier sequences are expected to be less influenced by outliers and to provide more information about the underlying processes, the general heuristic is to keep these models. The remaining question is, how many models should be considered? To answer this, the percentage of modeled sequences is swept and at each step a heuristic h is obtained as

$$h = \frac{\sum_{j=1}^K \lambda_j}{\sum_{j'=K+1}^N \lambda_{j'}}, \quad (9)$$

where λ_j is the j^{th} eigenvalue of the spectral clustering GEV problem, sorted by decreasing magnitude. This value can be intuitively seen as representing a normalized measure of the energy preserved by taking only the first K eigenvalues, and thus can be regarded as a measure of the clustering quality. It is then natural to select as the appropriate percentage of sequences to model the one that maximizes h . Similar approaches can be found in the PCA literature for selecting the optimum number of principal components to retain [22]. As previously stated, this is a simple method with no aspirations

3. <http://htk.eng.cam.ac.uk>

to be optimum but developed just for illustrating that an adequate selection of models can be advantageous if not necessary to obtain good clustering results. It is also worth noting that, using this method, the model selection is carried out based on a likelihood matrix obtained using all the sequences in the dataset. We refer to the KL+SC method coupled with this model selection scheme as KL+SC+MS.

Table 1 shows results (averaged along 15 iterations) using spectral and hierarchical clustering⁴, with the former performing considerably better. Agglomerative methods fails because in this scenarios because the relationships amongst the data that must be exploited in order to obtain an adequate partition are impossible to capture in a pairwise fashion.

All in all, the KL+SC+MS combination noticeably outperforms the rest of the alternatives. The use of KL+SC or KL+HC without model selection does not work very well because in this scenario there is no such thing as “true” HMMs generating the dataset. This way, the interpretation of the likelihood matrix that gives birth to our proposal loses much of its strength. However, it can be seen that if the KL+SC combination is coupled with the proposed model selection method it produces great results even in this otherwise adverse conditions.

A remarkable fact is that the previously proposed distances combined with spectral clustering suffer a severe performance loss as the number of hidden states increases. This is due to the models overfitting the individual sequences which, as previously stated, gives poorly structured distance matrices. The use of our proposed KL distance gives more stable results along the number of states, resulting in an improved performance relative to the other distances as the overfitting becomes more noticeable. This robustness is a very useful property of our proposal since, in practice, it is usually hard to determine the optimum model structure and overfitting is likely to occur.

It is also worth noting that the advantage of using our proposed methods is clearer in the GPM-UC3M dataset, because of the number of sequences considered in each clustering task being larger. This agrees with the conclusions drawn from the results obtained with synthetic data.

In Table 2 we show the number of models chosen for consideration by the model selection algorithm in each of the clustering tasks. The straightforward conclusion is that the algorithm takes into account a larger number of models as the overfitting increases. This is logical since the poorer the models, the more of them are required in order to give an adequate representation of the sequences for clustering purposes.

4. Results of the POR distance have been omitted from the table because of its extremely bad performance (around 50% error)

5 CONCLUSIONS AND FUTURE WORK

We have proposed a new distance measure for sequential data clustering, based on the Kullback-Leibler divergence. It embeds information of the whole dataset into each element of the distance matrix, introducing certain structure that makes it specially suitable for its use in combination with spectral clustering techniques. This measure also allows for the use of a reduced representative subset of models, which, if chosen properly, can give an increase in performance in real-world scenarios with potential outliers and misleading data.

A model selection scheme for this task has also been presented in the paper, with very positive results. This method works from a likelihood matrix L constructed using all the sequences in the dataset. If the selection could be done directly on the sequences, great computational cost reduction would be achieved by not having to fit a model for each sequence in the dataset.

The reported results have been obtained using HMMs as generative models for the individual sequences, although the proposed method is independent of this selection. In fact, exploring more expressive models is a straightforward and promising future line in order to successfully apply this clustering technique to a wider range of problems, such as video event detection, text mining, etc.

REFERENCES

- [1] L. Rabiner, “A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition,” *Proc. of the IEEE*, vol. 77, no. 2, pp. 257–286, February 1989.
- [2] S. Fine, Y. Singer, and N. Tishby, “The Hierarchical Hidden Markov Model: Analysis and Applications,” *Machine Learning*, vol. 32, no. 1, pp. 41–62, 1998.
- [3] J. Bilmes, “Buried Markov Models for Automatic Speech Recognition,” in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing*, Phoenix, AZ, March 1999.
- [4] Z. Wu and R. Leahy, “An Optimal Graph-Theoretic Approach to Data Clustering: Theory and its Application to Image Segmentation,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 15, no. 11, pp. 1101–1113, November 1993.
- [5] R. Xu and D. W. II, “Survey of Clustering Algorithms,” *IEEE Trans. Neural Networks*, vol. 16, no. 3, pp. 645–678, May 2005.
- [6] P. Smyth, “Clustering Sequences with Hidden Markov Models,” *Advances in Neural Information Processing Systems*, vol. 9, pp. 648–654, 1997.
- [7] A. Panuccio, M. Bicego, and V. Murino, “A Hidden Markov Model-Based Approach to Sequential Data Clustering,” in *Proc. of the Joint IAPR International Workshop on Structural, Syntactic and Statistical Pattern Recognition*, 2002, pp. 734–742.
- [8] F. Porikli, “Clustering Variable Length Sequences by Eigenvector Decomposition Using HMM,” in *Proc. International Workshop on Structural and Syntactic Pattern Recognition*, Lisbon, Portugal, 2004, pp. 352–360.
- [9] J. Yin and Q. Yang, “Integrating Hidden Markov Models and Spectral Analysis for Sensory Time Series Clustering,” in *Fifth IEEE International Conference on Data Mining*, November 2005.
- [10] P. Baldi, S. Brunak, and G. Stolovitzky, *Bioinformatics: The Machine Learning Approach*. MIT Press, 1998.
- [11] G. Jin, L. Tao, and G. Xu, “Hidden Markov Model Based Events Detection in Soccer Video,” in *Proc. International Conference on Image Analysis and Recognition*, 2004, pp. 605–612.
- [12] A. Dempster, N. Laird, and D. Rubin, “Maximum Likelihood from Incomplete Data via the EM Algorithm,” *Journal of the Royal Statistical Society*, vol. 39(1), pp. 1–38, 1977.

TABLE 1

Mean and standard deviation of clustering accuracy (%) on the AHUMADA isolated digits and GPM-UC3M isolated words datasets using HMMs with different number of hidden states and (a) spectral clustering (b) hierarchical clustering

Dataset	# of hidden states	SYM	BP	YY	KL	KL+MS
AHUMADA	m=2	79.08 (± 0.41)	82.72 (± 0.42)	78.15 (± 0.33)	77.65 (± 0.45)	82.36 (± 0.76)
	m=3	78.83 (± 0.35)	75.88 (± 0.51)	77.74 (± 0.44)	77.61 (± 0.42)	82.01 (± 0.48)
	m=4	76.80 (± 0.74)	71.42 (± 1.01)	75.75 (± 0.66)	75.85 (± 0.63)	81.02 (± 0.52)
GPM-UC3M	m=2	83.64 (± 4.01)	86.26 (± 4.53)	83.91 (± 4.45)	83.11 (± 4.87)	89.17 (± 4.00)
	m=3	73.05 (± 4.34)	75.93 (± 5.25)	74.89 (± 4.83)	76.13 (± 4.52)	90.35 (± 2.93)
	m=4	61.68 (± 2.73)	61.82 (± 3.44)	64.88 (± 4.33)	70.25 (± 3.76)	89.96 (± 2.50)

(a) Spectral clustering

Dataset	# of hidden states	SYM	BP	YY	KL
AHUMADA	m=2	61.30 (± 0.34)	57.61 (± 0.61)	58.06 (± 0.44)	57.95 (± 0.61)
	m=3	62.10 (± 0.83)	58.46 (± 0.61)	59.12 (± 0.68)	58.93 (± 0.58)
	m=4	62.67 (± 0.57)	59.55 (± 0.45)	60.45 (± 0.30)	60.42 (± 0.45)
GPM-UC3M	m=2	68.23 (± 2.49)	62.61 (± 1.93)	64.42 (± 1.56)	63.2 (± 1.61)
	m=3	58.74 (± 1.82)	56.61 (± 1.67)	57.41 (± 2.01)	61.55 (± 2.25)
	m=4	55.27 (± 1.63)	54.48 (± 1.70)	54.82 (± 1.38)	62.25 (± 2.58)

(b) Hierarchical clustering

TABLE 2

Optimal number of models chosen by the model selection algorithm

Dataset	m=2	m=3	m=4
AHUMADA	66.63%	72.28%	75.3%
GPM-UC3M	43.1%	58.61%	63.20%

- [13] K. Murphy, "Dynamic Bayesian Networks: Representation, Inference and Learning," Ph.D. dissertation, UC Berkeley, Computer Science Division, July 2002.
- [14] J. Shi and J. Malik, "Normalized Cuts and Image Segmentation," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888–905, August 2000.
- [15] A. Ng, M. Jordan, and Y. Weiss, "On Spectral Clustering: Analysis and an Algorithm," in *Advances in Neural Information Processing Systems*, 2002.
- [16] B. Juang and L. Rabiner, "A Probabilistic Distance Measure for Hidden Markov Models," *AT&T Technical Journal*, vol. 64, no. 2, pp. 391–408, February 1985.
- [17] T. Oates, L. Firoiu, and P. R. Cohen, "Using Dynamic Time Warping to Bootstrap HMM-Based Clustering of Time Series," in *Sequence Learning - Paradigms, Algorithms, and Applications*. London, UK: Springer-Verlag, 2001, pp. 35–52.
- [18] T. Jebara, Y. Song, and K. Thadani, "Spectral Clustering and Embedding with Hidden Markov Models," in *Proc. of the 18th European Conference on Machine Learning (ECML)*, Warsaw, Poland, September 2007.
- [19] S. Kullback and R. Leibler, "On Information and Sufficiency," *Annals of Mathematical Statistics*, vol. 22, pp. 79–86, 1951.
- [20] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.
- [21] J. Ortega-Garcia, J. Gonzalez-Rodriguez, and V. Marrero-Aguilar, "Ahumada: A large speech corpus in spanish for speaker characterization and identification," *Speech Communication*, vol. 31, pp. 255–264, 2000.
- [22] I. Jolliffe, *Principal Component Analysis*, 2nd ed. Springer, 2002.

Darío García-García Biography text here.

Emilio Parrado-Hernández Biography text here.

Fernando Díaz-de-María Biography text here.