

Clustering Health Trajectories Using Hidden Markov Models

Shima Ghassempour^{*}
shima.ghassempour@gmail.com

Federico Girosi[†]
F.Girosi@uws.edu.au

ABSTRACT

In this paper we describe an algorithm for clustering multivariate time series with variables taking both categorical and continuous values. Time series of this type are frequent in health care, where they represent the health trajectories of individuals. The problem is challenging because categorical variables make it difficult to define a meaningful distance across trajectories.

We propose an approach based on Hidden Markov Models (HMMs), where we first map each trajectory into an HMM, then define a suitable distance between HMMs and finally proceed to cluster the HMMs with a method based on a distance matrix. We test our approach on a simulated, but realistic, data set of 1,200 trajectories of individuals of age 45 and over. In addition to our preferred definition of distance between HMMs, based on the Kullback-Leibler divergence, we test an alternative definition and obtain a very similar cluster structure. The proposed method can be implemented quite simply with standard tools and packages in R and `Matlab` and may be a good candidate for solving the difficult problem of clustering multivariate time series with categorical variables.

Categories and Subject Descriptors

H.3.3 [Information Search and Retrieval]: Clustering

Keywords

Health Trajectory, Hidden Markov Model, Clustering

1. INTRODUCTION

The interaction of a patient with the health care system takes place at different points in space and time. This implies that in many cases the natural unit of observation for health care research is the entire trajectory of a patient. As linked data and personal health records become more easily

^{*}School of Computing, Engineering and Mathematics & Centre for Health Research, University of Western Sydney, New South Wales, Australia

[†]Centre for Health Research, School of Medicine, University of Western Sydney, New South Wales, Australia

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

KDD-DMH'13, August 11, 2013, Chicago, Illinois, USA
Copyright 2013 ACM 978-1-4503-2174-7/13/08 ...\$15.00.

available we expect that both researchers and health care stakeholders will have an increasing need for tools that can be used to analyze health trajectories.

A very common type of analysis one may need to perform on health trajectories is clustering. For example an insurer may wish to cluster claims trajectories in order to better account for risk categories, and a health regulator may need to cluster administrative data trajectories for the purpose of defining appropriate groups for activity based funding.

Clustering trajectories requires to define a proper notion of distance and to preserve the information about the ordering of the events, which are both challenging tasks. Traditional clustering techniques do not extend naturally to the domain of trajectories, and a fair amount of research has been performed on this subject. Most of the literature focuses on continuous valued time-series (possibly multi-dimensional) [16], which are much easier to analyze (and visualize) than time series of categorical variables.

Unfortunately, the time series we expect to find in health care related research are likely to contain a mix of categorical and continuous variables. Categorical variables may be used to denote a health condition (such as breast cancer, or diabetes), or a risk factor (such as smoking), the use of medication and the administration of a procedure or of a laboratory test. Continuous variables may arise in conjunctions with claims, costs and results of laboratory tests (such as glucose or cholesterol levels).

As a consequence, in order to make our research relevant for health care applications, in this paper we focus on the issue of clustering trajectories with a method that can handle both continuous and categorical variables at the same time.

Our approach is conceptually simple: since time series of continuous and categorical variables are difficult objects to deal with, we replace each time series with a model that is likely to generate it, and then perform the clustering in the space of models. As generative models we use Hidden Markov Models (HMMs) [22], extended to allow for covariates [2, 25]. Clustering HMMs is possible because to each HMMs we can attach the likelihood of generating any given trajectory, and the likelihoods can be used to define distances between HMMs. In turns this allows us to use any clustering technique which is based on a distance matrix.

The approach taken in this paper is similar in spirit to the one taken by Ramoni and collaborators [18, 19], which is based on Markov Chains rather than HMMs, and we will argue in the next section that for applications in health care HMMs with covariates might be a more appropriate starting point than Markov Chains.

2. HIDDEN MARKOV MODELS

Hidden Markov Models (HMMs) are a class of probabilistic models that were introduced in late 60s [4] and proved to be extremely useful in a variety of disciplines, including speech recognition, weather prediction, financial time series analysis, robotics, detecting protein homologies and computer vision [13, 22, 25]. The feature that distinguishes HMMs from the more conventional Markov chains is the fact in HMMs the state of the system is not observed. What is observed is the collection of values taken by a number of variables, whose probability distribution depends on the hidden state.

For the sake of mathematical and computational tractability the following assumptions are usually made: 1) the hidden states form a Markov chain, and 2) the probability distribution of an observation only depends on the current state. Therefore an HMM can be represented by a triple of parameters: the probability distribution of the initial state, the transition matrix for the hidden states, and the probabilities of the observations conditional on the hidden state.

HMMs are particularly appealing in the health care setting because they capture the notion that the health state of an individual is not a well defined quantity, and that the observations available on an individual only capture certain dimensions of health, and do not necessarily get to underlying construct of health state.

Therefore it seems reasonable to assume that a person’s health state remains unobservable, and the only quantities that we can observe are certain “manifestations” of the hidden state. This is very much in line with the approach taken, usually in a static setting, by Latent Class Analysis (LCA). Based on the success of LCA in a wide range of applications [12], it seems likely that HMMs will prove to be a valuable tool for the analysis of dynamic data sets in health care. In addition, recent advances in HMM theory [25] allow to incorporate easily time-dependent covariates into the model, so that unrealistic stationarity assumption do not need to be made. In the next section we will discuss how we plan to use HMMs to solve the problem of clustering health trajectories.

3. METHOD

In this section we describe our method in fairly general terms, leaving some of the details for section 4, that discusses the experimental results. Our starting point is a set of N health trajectories T_i corresponding to N distinct individuals, where each trajectory is a matrix with d columns. Each column is the time series of a categorical or continuous variable. The d time series will be in general correlated, and we refer to the variables they represents as the “observables”. The lengths of the trajectories do not have to be the same across individuals, but the observables must be defined for all individuals, meaning that we are taking the same measurements for all individuals in the data set.

Clustering the trajectories T_i is difficult because there is no obvious notion of similarity between trajectories. If all the observables were continuous-valued variables the problem would be easier, since in that case several notions of distance are available (for example, the Frechet distance [1]).

Since in most applications to health care we expect that many of the observables will be categorical we accept the fact that there is no obvious metric structure on the set of trajectories and take a different approach: we replace each trajectory with a probabilistic generative model, that

is a set of probabilistic rules that are likely to generate that trajectory. Then, rather than comparing two trajectories we compare the models that generated them.

This approach is attractive because, as we will discuss in more details in section 3.1, there is a natural way to compare probabilistic models that generate random trajectories: two models are similar if the probability distributions of the trajectories they generate are similar. Therefore the ill-defined problem of comparing two trajectories has been replaced by the well-defined (although not necessarily easy) problem of comparing two probability distributions.

As explained in section 2 our preferred class of probabilistic models is HMMs. This means that each trajectory T_i is mapped into an HMM λ_i . In order to avoid making stationarity assumption we use HMMs that allow the model probabilities to depend on a specific set of covariates [2, 25]. In order to fit an HMM model λ_i to the trajectory T_i one needs to decide the number of hidden states. This need not to be the same for all trajectories (for example if the length of the trajectories has high variance), and can be done using a variety of methods [20, 25].

Once the N HMMs models λ_i have been fit one can apply the preferred notion of distance $D(\cdot; \cdot)$ between two HMMs to define a distance matrix $D_{ij} \equiv D(\lambda_i; \lambda_j)$. This effectively defines the distance between two trajectories as $D(T_i; T_j) \equiv D(\lambda_i; \lambda_j)$, which was the most difficult obstacle to clustering. Armed with a distance matrix one can then apply one of several clustering algorithms, such as Partition Around Medoids (PAM) [14], spectral clustering [17] and hierarchical clustering [10], whose sole input is a distance matrix, and do not require manipulations of the objects underlying that matrix, as in the standard k -means algorithm. Therefore, once the distance matrix D_{ij} has been determined the problem can be considered solved, since standard methods can then be applied.

In the next section we describe in more details how a distance between HMMs can be computed and the choices we made in practice.

3.1 Definition of distance between HMMs

The need to compare different HMMs through an appropriate distance measure is not new, and has arisen in a variety of contexts such as speech recognition [9, 21, 22], document and image classification [5], time series prediction [7] and of course in the literature on clustering HMMs [6, 11].

In order to define a meaningful distance between two HMMs researchers have used the fact that it is easy to compute, by means of the forward-backward algorithm [21], the likelihood $P_{ij} \equiv P(T_i | \lambda_j)$ of observing a trajectory T_i given a model λ_j , and that likelihoods can be used to define distances.

The key observation is that the likelihood of a trajectory T given an HMM model λ , $P(T|\lambda)$, can be seen as a probability density on the space of trajectories, and a standard notion of distance between probability densities already exists, being the Kullback-Leibler (KL) divergence [15]. Therefore the distance between models λ_i and λ_j could be measured as the KL distance between $P(T|\lambda_i)$ and $P(T|\lambda_j)$, which is:

$$D_{KL}(P(T|\lambda_i); P(T|\lambda_j)) \equiv \int dT P(T|\lambda_i) \log \frac{P(T|\lambda_i)}{P(T|\lambda_j)} \quad (1)$$

The expression above is quite difficult to compute since it requires to integrate over the space of trajectories [9, 11]. Therefore we settle for an approximation of it: we replace

the integral over the space of trajectories with a sum over the space of trajectories that have been actually observed, under the assumption that they are sufficiently representative of the bigger space.

Operationally, for each model λ_j we create a likelihood vector $P(T_k|\lambda_j), k = 1, \dots, N$, that we normalize to one, obtaining a discrete probability distribution P_j^* over the set of observed trajectories. The KL distance between two discrete distributions p and q is easy to compute, since it is

$$D_{KL}^{disc}(p; q) \equiv \sum_{\alpha} p_{\alpha} \log \frac{p_{\alpha}}{q_{\alpha}}$$

Therefore we define the distance between model λ_i and model λ_j as

$$D_{KL}(\lambda_i; \lambda_j) \equiv \frac{1}{2} \left(D_{KL}^{disc}(P_i^*; P_j^*) + D_{KL}^{disc}(P_j^*; P_i^*) \right) \quad (2)$$

where we have “symmetrized” the expression to account for the fact that the KL distance is not symmetric.

While it is not the purpose of this paper to perform a comparison among distance measures for HMMs, as was done in [11], we do want to demonstrate that our method is not unstable, and that a change in the distance definition does not lead to a very different result. Therefore we adopt a second definition of distance, fully aware that we are using the word “distance” not in the strictly mathematical sense.

The second definition is motivated by the observation that, intuitively, two models are different if they differ in the probability of generating the same trajectory [22]. For example, defining the log-likelihoods $L_{ij} \equiv \log P_{ij}$, we expect the terms $L_{ii} - L_{ij}$ and $L_{jj} - L_{ji}$ to be large if λ_i and λ_j are different. This led Yin and Yang [24] to define the following distance

$$D_{YY}(\lambda_i, \lambda_j) = |L_{ii} - L_{ij} + L_{jj} - L_{ji}| \quad (3)$$

Porikli [20] introduced a similar distance using likelihoods instead of log-likelihoods. However, we prefer the Yin and Yang (YY) distance because the log-likelihood has better information theoretical properties [22]. In addition, the term $L_{ii} - L_{ij}$ and its symmetric counterpart in equation 3 can be seen as a very crude approximation of the KL distance. In fact, if we assume that the conditional probability $P(T|\lambda_i)$ in equation 1 is highly concentrated around the trajectory T_i , then the integral in that expression can be estimated by the “one point approximation” $\log \frac{P(T_i|\lambda_i)}{P(T_i|\lambda_j)} = L_{ii} - L_{ij}$.

To summarize, while our preferred distance is the KL distance because it makes use of more information, we consider in this paper also the YY distance, so that we can study the stability of the results to a change in distance metric.

4. AN EXPERIMENT WITH REALISTIC, SIMULATED DATA

We begin with a description of the simulated data set we used for the experiment.

4.1 Data

The data for our study come from a simulation based on the 45 and Up Study dataset [3]. The 45 and Up Study is based in the population of the state of New South Wales (NSW), Australia. Prospective participants were randomly sampled from the enrolment database of Medicare Australia, which provides near complete coverage of the population. People 80+ years of age and residents of rural and remote areas were oversampled. A total of 267,153 participants

joined the Study by completing a baseline questionnaire (between Jan 2006 and April 2009) and giving signed consent for follow-up and linkage of their information to routine health databases. About 18% of those invited participated and participants included about 11% of the NSW population aged 45 years and over.

A subset of 60,000 participants was interviewed two to three years after baseline, as part of the Social, Economic and Environmental Factors (SEEF) study. The longitudinal structure of this data set allowed us to estimate transition probabilities between health states over a two years interval. Applying repeatedly the transition probabilities to the original 45 and Up data we obtained a Markov model that we used to generate over 260,000 health trajectories. The trajectory of an individual stops when the individual dies, where the probability of dying was estimated using data from the Australian Bureau of Statistics mortality dataset linked to the 45 and Up data.

The health state of an individual was defined by a vector of binary and categorical variables associated to the following chronic conditions and risk factors: heart disease, diabetes, stroke, hypertension, cancer, BMI status and smoking status (current smoker, quit smoker, not a smoker).

All the transition probabilities of the model were estimated using probit regressions that included as covariates the previous health state, age, gender, income, education, and insurance status. The error terms in the probit regressions are correlated, so that the time series of different health conditions are correlated and reflect the observed patterns of co-morbidities.

We underscore that the model underlying these data is much more complex than the HMM used in our experiments. In addition, there is no guarantee that we will find any clusters, since we have not artificially introduced any clustering structure, and therefore we consider this model quite realistic.

4.2 Experimental Results

In order to test the feasibility of the method that we proposed we selected a subset of the 260,000 synthetic trajectories generated as described in section 4.1. Our goal was to use a set that was sufficiently complex and that exhibited a high degree of variability. We wanted to avoid using “trivial” trajectories, in which an individual never develops any disease or only develops a condition prior to death, since those are easy to differentiate from the others. Therefore we picked a set of 1,200 trajectories associated with the individuals who developed *all* the three health conditions of heart disease, stroke and diabetes at some point during their life, and included the Body Mass Index (BMI) as a time varying covariate. We did not include any continuous valued variable because none were available, but it is important to note that HMMs handle continuous and categorical variables equally well, and therefore this is not a concern.

The average age of the cohort is approximately 60, and the length of trajectories is about 18 on average, although it does vary between 4 to 20 time steps, where a time step corresponds to a period of two years of life.

The trajectory for individual i is therefore represented by 4 time series ($H_{it}, D_{it}, S_{it}, BMI_{it}$), which correspond to the variables for heart disease (H), diabetes (D), stroke (S) and BMI. It is important to underscore that the simulated data are realistic and therefore the four time series for each individual are correlated. Methods that treat the time series as

independent (such as the one in [18]) would not be appropriate in this context, and probably not appropriate in most health care applications.

For each of the 1,200 trajectories T_i we estimate a corresponding HMM λ_i using BMI as a time varying covariate. The estimation is performed in R using the package `dep-mixS4` [23]. We performed a preliminary analysis and found that most of the times an HMMs with three hidden states was preferable, although for short trajectories two hidden states would suffice. Since the purpose of the paper is to demonstrate the validity of the method and not solve a particular problem optimally we opted for simplicity and used three hidden states for all the trajectories.

Once we estimated the HMMs we used the standard forward-backward algorithm [21] to compute the likelihoods $P(T_i|\lambda_j)$, that were then used as input for the algorithms described in section 4.1 to compute KL and YY distance matrices.

Clustering is performed, on each of the matrices, using the Partition Around Medoids (PAM) algorithm [14]. The PAM algorithm takes as input the number K of clusters. Since that is unknown, we tried values of K from 2 to 7, and then used the Dunn Index [8] to select the optimal number of clusters. The Dunn Index is defined in terms of the intra-cluster distance, which measures the mean distance between all pairs of points in a cluster k , and the inter-cluster distance, which measures the mean of the pairwise distances between points in cluster C_i and C_j . A higher value of the Dunn index denotes a better clustering.

In order to remove dependence from the initial conditions the PAM algorithm was run 100 times for each value of K , and the clustering with the best Dunn Index was chosen as the result corresponding to K clusters.

In Figure 1 we show the Dunn Index as a function of the number of clusters K for both the KL and the YY distance. For the YY distance the Dunn Index is maximum for two clusters, while for the KL distance it achieves its maximum for three clusters. However, the Dunn Index for the YY distance takes very similar values at $K = 2$ and $K = 3$, and therefore in order to compare the results between the YY and the KL distance we choose $K = 3$ as our preferred number of clusters.

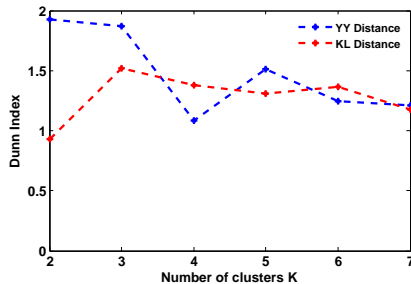
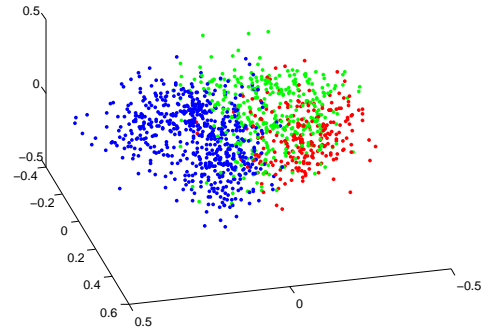
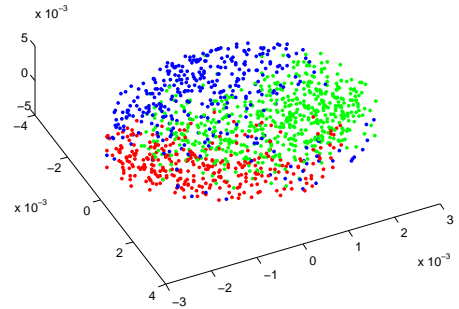


Figure 1: The Dunn Index for both the YY and the KL distance plotted as a function of the number of clusters K . In order to produce this plot, for each fixed number of clusters K the PAM algorithm was run 100 times and only the result with the best Dunn Index was plotted for that K .

In order to convince ourselves that three is a reasonable number of clusters for these data we use Multidimensional Scaling (MDS) to visualize the cluster structure. MDS works by finding a set of points in an d -dimensional Euclidean



(a) KL Distance



(b) YY Distance

Figure 2: MDS in three dimensions applied to the KL (a) and the YY (b) distance matrices. There are 1,200 points in each plot, one for each of the trajectories. We used the PAM algorithm to define three clusters, and we color-coded the points according to which of the three cluster they belong. The color coding is performed independently for the two sets of points, so points of one color in figure (a) cannot be matched to points of the same color in figure (b): this is just a depiction of the metric structure of the data and of the results of the clustering algorithm.

space such that the distance matrix of these points is as similar as possible to a given distance matrix D . By choosing d equal to two or three one can plot the location of these points and obtain an intuitive representations of how the points that generated the original distance matrix D relate to each other.

In our case we choose d equal to three, and for each of the distance matrices associated to the KL and to the YY distances we find a corresponding set of points in three dimensions. We then plot these points and assign them a unique color depending on which cluster they are in. We manually rotate the plotting view until we find one that best shows the cluster structure. The results of this procedure are shown in Figure 2.

While figure 2 does suggest that three clusters exists in the data, MDS is only a visual help and does not tell us anything about what is the content of these three clusters, or whether the grouping that we found is sensible. In order to interpret the meaning of the clusters we define, for each trajectory, a set of continuous valued features that characterize it, and then compare the means of these features across the three clusters, looking for some meaningful patterns.

Since a common measure used in health care research is the amount of time spent with a certain health condition, we compute for each trajectory the percentage of time that the individual spends with specific combinations of the three health conditions we have considered. We consider all the possible combinations (although they are not independent), that we label H, D, S, HD, HS, SD, HSD, where H stands for heart disease, D for diabetes and S for stroke. We also compute the percentage of the time spent in one of the three BMI states: normal, overweight and obese. We purposefully neglect the underweight state because it is relatively rare in this sample. As a result, each trajectory can be associated with a “profile” of 10 features, where each feature represents the percentage of the trajectory spent with a certain health condition. It is important to underscore that the features are not exclusive: so “time spent with heart disease” does not mean “time spent exclusively with heart disease”.

In Figure 3 (a) and (b) we show average, cluster-specific profiles for the three clusters obtained using the KL distance and the YY distance respectively. More specifically, for each cluster and for each distance definition we show the average percentage of trajectory spent with 10 different health conditions. From this figure we can draw the following conclusions:

- *the three clusters are composed of individuals with significantly different health trajectories.* This is clear from the fact that the profiles corresponding to the different clusters are sufficiently different.

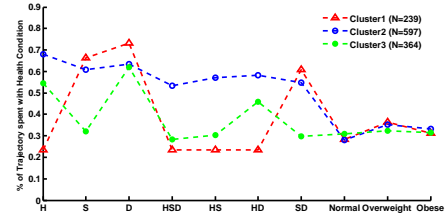
- *the profiles are easily interpretable and show meaningful groupings.* For example, looking at panel (a) we notice that people in cluster 2 (the blue profile) spent more than half of their trajectories with heart disease, diabetes and stroke at the same time. Individuals in cluster 1 and 3 are different, because they only spent 20 to 30 percent of their trajectories in such a state. What differentiates cluster 1 from cluster 3, however, is the time spent with two conditions: people in cluster 1 spent considerable time with diabetes and stroke, but not with heart disease, while people in cluster 3 spend considerable time with heart disease and diabetes, but not stroke. A similar interpretation is valid for panel (b).

- *the KL distance and the YY distance find three clusters with qualitatively similar profiles.* If we “read” panel (b) we find that it tells a very similar story to the one of panel (a). In fact, we were able to match the clusters obtained with the two distances, and color coded them in the same way. While the heights of the profiles are not exactly the same, and for example the profile of cluster 3 in panel (b) is about 10 percentage points too low compared to its partner in panel (a), the profiles have very similar shapes. The sizes of the clusters are also quite similar, although the size of cluster 1 in panel (a) is small compared the one in panel (b).

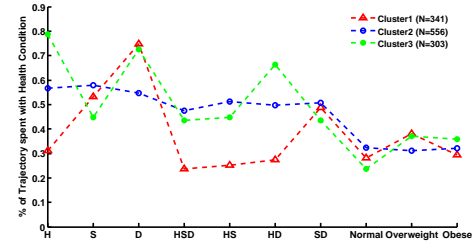
In order to confirm that individuals in the three clusters are indeed different we also looked at some baseline characteristics of the individuals in the three clusters, measured at the beginning of their trajectory. While these variables do not enter the HMMs, they were used to generate the simulated data, and therefore we expect them to be correlated with the health trajectories.

In Table (1) we show the distribution of some socio-economic variables and risk factors for people in the three clusters found with the KL distance (results for the YY distance are

similar). Without discussing each table entry, we merely point out that the tables show, for many variables, significant differences across clusters. In a real world application a table of this type would be used to understand the composition of the clusters that have been discovered.



(a) KL distance



(b) YY distance

Figure 3: Profiles of the three clusters in the feature space, for the KL distance (panel a) and the YY distance (panel b). The features correspond to the percentage of the trajectory time spent with different combinations of health conditions (H for heart disease, S for stroke and D for diabetes) and with different BMI status. For example, in panel (a), the height of the profile of cluster 3 (green) for feature “H” is 55%. This means that, on average, individuals in cluster 3 (defined using the KL distance) spent 55% of their trajectory time with heart disease.

5. CONCLUSION

While there is significant literature on the problem of clustering time series of continuous variables [16], much less is known when the time series consist of categorical variables. In this paper we have described a clustering method that is indifferent to the distinction of continuous versus categorical.

Our goal was to devise a method that was sound but also easy to implement and accessible to a large group of researchers, possibly in the field of health care, where we expect the problem of clustering health trajectories will become more and more relevant. The method is sound because it is based on the well-tested approach of embedding a time series into a dynamic model. The crucial step of the method is in the approximate computation of the KL distance between two HMMs. The fact that when using the alternative YY distance (which utilizes less information) we obtain similar results suggests that the method is robust to the choice of distance. Whether the approximate KL distance is better than the YY distance remains to be seen, and may depend on the specific problem at hand.

The data set on which we tested the algorithm was simulated but realistic, because it was generated by a model

Table 1: Distributions of some socio-economic variables and risk factors for the individuals in the three clusters found using the KL distance. The table shows that individuals in the three clusters are different in meaningful ways. For example, people in cluster 2 have the highest prevalence of hypertension (84%). This is sensible since Figure 3 already showed that these people are the sickest group.

Trajectory Length	Mean		
Cluster1	17.80		
Cluster2	18.04		
Cluster3	18.20		
Age	Mean	Std	
Cluster1	61.28	8.38	
Cluster2	59.58	7.81	
Cluster3	57.95	7.62	
Hypertension	No	Yes	
Cluster1	0.18	0.82	
Cluster2	0.16	0.84	
Cluster3	0.24	0.76	
Gender	Female		Male
Cluster1	0.56		0.44
Cluster2	0.55		0.45
Cluster3	0.61		0.39
Income	Less than 20K	20-40K	More than 40K
Cluster1	0.28	0.21	0.51
Cluster2	0.21	0.21	0.58
Cluster3	0.25	0.19	0.56
Smoking	Not smoking	Quit smoking	Smoking
Cluster1	0.63	0.32	0.05
Cluster2	0.57	0.35	0.08
Cluster3	0.56	0.34	0.10

based on longitudinal survey data. The problem was challenging because the data contained trajectories of different lengths corresponding to an heterogeneous population, and it was not designed to contain any specific clustering structure. The algorithm discovered three clusters of significantly different individuals, and it was easy to interpret their composition. The method seems promising, and we plan to test it on increasingly more difficult problems.

Acknowledgment

We are grateful to Anthony Maeder for providing useful suggestions and insight and to Louisa Jorm for her constant support.

6. REFERENCES

- [1] H. Alt and M. Godau. Computing the Frechet distance between two polygonal curves. *International Journal of Computational Geometry & Applications*, 05(01n02):75–91, 1995.
- [2] K. Banachewicz, A. Lucas, and A. Van Der Vaart. Modelling portfolio defaults using Hidden Markov Models with covariates. *Econometrics Journal*, 11(1):155–171, 2008.
- [3] E. Banks et al. Cohort profile: the 45 and up study. *Int J Epidemiol*, 37(5):941–947, 2008.
- [4] L. Baum and T. Petrie. Statistical Inference for Probabilistic Functions of Finite State Markov Chains. *The Annals of Mathematical Statistics*, 37(6):1554–1563, 1966.
- [5] M. Bicego and V. Murino. Investigating Hidden Markov Models’ capabilities in 2D shape classification. *IEEE Transactions on PAMI*, 26(2):281–286, 2004.
- [6] E. Coviello, A. Chan, and G. Lanckriet. Clustering hidden Markov models with variational HEM. *Computing Research Repository*, 1210.6707, 2012.
- [7] J. Duan, W. Wang, J. Zeng, D. Zhang, and B. Shi. A prediction algorithm for time series based on adaptive model selection. *Expert Systems with Applications*, 36(2):1308–1314, 2009.
- [8] J. C. Dunn. A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*, 3(3):32–57, 1973.
- [9] M. Falkhausen, H. Reininger, and D. Wolf. Calculation of distance measures between hidden Markov models. In *Proc. Eurospeech*, pages 1487–1490, 1995.
- [10] C. Fraley. Algorithms for model-based gaussian hierarchical clustering. *SIAM Journal on Scientific Computing*, 20(1):270–281, 1998.
- [11] D. García-García, E. P. Hernández, and F. Diaz de Maria. A new distance measure for model-based sequence clustering. *PAMI*, 31(7):1325–1331, 2009.
- [12] J. Hagenaaars and A. McCutcheon. *Applied Latent Class Analysis*. Cambridge University Press, 2002.
- [13] T. Jaakkola, M. Diekhans, and D. Haussler. Using the Fisher kernel method to detect remote protein homologies. In *Proceedings of the Seventh International Conference on ISMB*, volume 149, page 158, 1999.
- [14] L. Kaufman and P. J. Rousseeuw. *Finding groups in data: an introduction to cluster analysis*, volume 344. Wiley-Interscience, 2009.
- [15] S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.
- [16] T. W. Liao. Clustering of time series data: a survey. *Pattern Recognition*, 38(11):1857 – 1874, 2005.
- [17] U. Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.
- [18] P. S. M. Ramoni and P. Cohen. Multivariate clustering by dynamics. In *Proceedings of The National Conference on Artificial Intelligence*, pages 633–638, 2000.
- [19] P. S. M. Ramoni and P. Cohen. Bayesian clustering by dynamics. *Machine Learning*, 47(1):91–121, 2002.
- [20] F. Porikli. Trajectory distance metric using Hidden Markov Model based representation. In *IEEE ECCV, PETS Workshop*, volume 3, 2004.
- [21] L. Rabiner. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.
- [22] L. Rabiner and B. Juang. An introduction to Hidden Markov Models. *IEEE ASSP Magazine*, 3(1):4–16, 1986.
- [23] I. Visser and M. Speekenbrink. depmixS4: An R package for Hidden Markov Models. *Journal of Statistical Software*, 36(7):1–21, 2010.
- [24] J. Yin and Q. Yang. Integrating hidden Markov models and spectral analysis for sensory time series clustering. In *Fifth IEEE ICDM*, pages 506–513, 2005.
- [25] W. Zucchini and I. MacDonald. *Hidden Markov Models for Time Series: An Introduction Using R*. Taylor & Francis, 2009.