Learning similarities between irregularly sampled short multivariate time series from EHRs

Karl Øyvind Mikalsen[a,*], Filippo Maria Bianchi[a], Cristina Soguero-Ruiz[a,b], Stein-Olav Skrøvseth[c], Rolv-Ole Lindsetmo[c], Arthur Revhaug[c], Robert Jenssen[a,c]

[a] *Machine Learning Group, UiT – The Arctic University of Norway*
[b] *Dept. of Signal Theory and Comm., Telematics and Computing, Universidad Rey Juan Carlos*
[c] *University Hospital of North Norway*
[*] *Corresponding author.* Email: *karl.o.mikalsen@uit.no*

***Abstract*—A large fraction of the Electronic Health Records consists of clinical multivariate time series. Building models for extracting information from these is important for improving the understanding of diseases, patient care and treatment. Such time series are oftentimes particularly challenging since they are characterized by multiple, possibly dependent variables, length variability and irregular samples. To deal with these issues when such data are processed we propose a probabilistic approach for learning pairwise similarities between the time series. These similarities constitute a kernel matrix that can be used for many different purposes. In this work it is used for clustering and data characterization. We consider two different multivariate time series datasets, one of them consisting of physiological measurements from the Department of Gastrointestinal Surgery at The University Hospital of North Norway and we show the proposed method's robustness and ability of dealing with missing data. Finally we give a clinical interpretation of the clustering results.**

I. Introduction and background

The digitalization of the healthcare systems has lead to enormous opportunities for developing data-driven systems for extracting useful information from Electronic Health Records (EHRs) that can be used to improve the daily care and treatment of the patients. However, the fact that the very nature of the data is uniquely complex, has forced researchers to not merely re-use methods that have been shown to work well in other application domains, but to develop novel approaches specially designed for this particular kind of data [1].

A large fraction of the data in the EHRs consists of measurements over time such as blood tests and other physiological variables. The blood tests are important for understanding the patients' health status and the dynamics of their disease. Even small deviations over time can be indicators of serious underlying complications or diseases. To manually identify such changes is both a cumbersome and time consuming task that can be very difficult as well. Therefore, a reliable system that automatically discovers irregularities will benefit both the patients and the care givers, since complications could be identified at an earlier stage, and human resources could be allocated more efficiently.

The task of interpreting data relative to blood measurements could be thought of as comparing how similar or dissimilar patients are to each other. If the clinician identifies a pattern deviating from nominal conditions, an action is immediately taken. In the machine learning field, many methods require to evaluate a suitable similarity metric on the data. In the time series domain, by providing a dissimilarity measure that is meaningful for the application at hand, one can, in theory, apply several kinds of classical clustering procedures, such as e.g. *k-means*, *hierarchical clustering* or *spectral clustering* [2].

To compare time series, *Dynamic time warping* (DTW) [3] is one of the most famous methods and it has become the state-of-the-art dissimilarity measure in most applications. Many other approaches have been proposed in the literature, which, however, are limited to deal with univariate time series, where the values are regularly sampled and have no missing data.

However, the time series from the are often more complex. They can be characterized by multiple possibly interdependent variables, length variability (patients are monitored for different time periods), short time duration, and the time series are usually irregularly sampled over time [4], [5].

The DTW is designed to deal with the issues of length variability and different sampling rates, since one simply can compute the dissimilarity between time series of different length. There also exist several formulations for extending the DTW to the multidimensional setting [6]. However, in the case of multivariate time series with missing data no solution that works well in every circumstance has been proposed [7].

There have been some other attempts to deal with the aforementioned issues. Cruz et al. introduced a method based on fuzzy clustering where first order derivatives are accounted as features and imputation is used to deal with missing data in univariate time series [8]. Ghassemi et al. proposed to transform the irregularly sampled time series into a new latent space using *multi-task Gaussian processes* (MTGP) models to learn a similarity metric [9]. However, this was done in a supervised setting. Lasko et al. proposed a method for phenotype discovery using Gaussian process regression and deep learning [10]. Marlin et al. introduced a probabilistic approach to deal with the problem of missing data [11]. Specifically, they modeled the data using a mixture of diagonal covariance normal distributions. Under a *missing at random* (MAR) assumption [12], this method can effectively deal with missing data. Furthermore, to create robustness against sparsely-sampled data they put priors on the parameters and used a *maximum a posteriori* (MAP) algorithm to estimate the model. This yields estimates of the mean, which are smoother than ordinary maximum likelihood *expectation maximization* (EM) estimation [13]. They showed that even though the MAR assumption could be violated, the method still works and manages to discover interesting patterns in EHR data. However, a downside is the selection of three different hyper-parameters, whose tuning is not obvious. In particular, modeling the covariance in time is difficult; choosing a too small hyper-parameter leads to a degenerate covariance matrix, which cannot be inverted. On the other hand, a too large choice will basically remove the covariance such that the prior knowledge is not incorporated. In addition to the problem of choosing hyper-parameters, the method does not provide an obvious way of determining the number of clusters, which has to be set a-priori.

The recently proposed *probabilistic cluster kernel* (PCK) [14] is a promising approach that tackles the hyper-parameter dependency problems by introducing consensus clustering [15] into the *Gaussian mixture model* (GMM) framework for learning a parameter–free kernel. This kernel is supposed to account for probabilistic similarities at different resolutions. Furthermore, one obtains a similarity measure, embedded into a kernel matrix, that can be employed by kernel methods for e.g. feature extraction, classification or regression. So far, the PCK framework has been applied in the classical case of regular multivariate time-independent data [14].

In this work, we introduce a kernel approach that captures similarities between time series and is able to deal with the aforementioned real-data issues. The idea is to fuse the best from two approaches by using the probabilistic approach introduced by Marlin as a basis in the PCK framework. More specifically, we perform the MAP-EM for the diagonal covariance GMMs multiple times with different hyper-parameter choices and varying number of Gaussians. By doing so, we learn a probabilistic cluster kernel, which is a robust similarity measure for irregularly sampled multivariate time series. After having learned the kernel, we apply standard spectral clustering. We evaluate the proposed method both to simulated time series, with and without missing data, and to real-world medical data. In order to assess the effectiveness of the novel method in a comparative study, we propose several extensions of the DTW framework and to compare the performance of these approaches with the proposed method and the MAP-EM diagonal covariance GMM.

II. Methods

This section consists of two different parts that can be read separately. First, we explain the diagonal covariance Gaussian mixture model framework. Thereafter, the PCK framework is described. Due to space limitations we do not describe the DTW framework here, but for the interested reader we refer to [3], [6].

A. MAP-EM diagonal covariance GMM augmented with empirical prior.

1) Notations and model description: Assume that there are N multivariate time series, each of them consisting of V variables that are observed over T time intervals. We define a tensor X consisting of the entries x_{nvt}, which are the realizations of the stochastic variable X_{nvt} relative to variable v at time t in the n-th time series. We let R be the tensor with entries $r_{nvt} = 0$ if the realization x_{nvt} is missing and $r_{nvt} = 1$ if observed. In the following, if nothing else is specified, letters that appear with and without indices refer to tensors and their entries, respectively.

Given a multivariate time series n, described by the matrix x_n, we want to assign it to one of the G Gaussian distributions. With $Z_n = g$, we denote the assignment of x_n to cluster g and let θ_g be the parameter of the discrete prior distribution over the clusters. The density for the g-th cluster is completely described by the first two moments, μ_{gvt} and σ^2_{gv}, with the assumption that the variance is assumed to be time independent.

The basic model is described by the equations

$$P(Z_n = g) = \theta_g, \tag{1}$$

$$P(X_{nvt} = x_{nvt} \mid \mu_{gvt}, \sigma_{gv}) = \mathcal{N}(x_{nvt};\, \mu_{gvt}, \sigma^2_{gv}). \tag{2}$$

Under the MAR assumption, the missing data can be ignored in the likelihood computation [11], [12] and the posterior can be evaluated as

$$\begin{aligned}\pi_{ng} &\equiv P(Z_n = g \mid x_n,\, r_n,\, \theta,\, \mu,\, \sigma) \\ &= \frac{\theta_g \prod_{v=1}^{V} \prod_{t=1}^{T} \mathcal{N}(x_{nvt};\, \mu_{gvt}, \sigma^2_{gv})^{r_{nvt}}}{\sum_{k=1}^{G} \theta_k \prod_{v=1}^{V} \prod_{t=1}^{T} \mathcal{N}(x_{nvt};\, \mu_{kvt}, \sigma^2_{kv})^{r_{nvt}}}\end{aligned} \tag{3}$$

2) Parameter estimation using MAP-EM: To be able to deal with large amounts of missing data one can introduce informative priors for the parameters and estimate them using MAP-EM. This will ensure that it is possible to obtain a smooth mean over time in each cluster and that in clusters containing few time series their parameters are similar to the overall mean and covariance.

A kernel based Gaussian prior on the mean enforces smoothness,

$$P(\mu_{gv} \mid m_v,\, S_v) = \mathcal{N}(\mu_{gv};\, m_v,\, S_v), \tag{4}$$

where m_v is the empirical mean and the prior covariance matrix, S_v, is defined via the kernel,

$$K_{tt'} = b_0 \exp(-a_0(t - t')^2), \tag{5}$$

and the empirical standard deviation s_v as

$$S_v = s_v K_{tt'}, \tag{6}$$

where a_0, b_0 are user-defined hyper-parameters. On the standard deviation σ_{gv} we put an inverse Gamma distribution prior,

$$P(\sigma_{gv} \mid N_0,\, s_v) \propto \frac{1}{\sigma_{gv}^{N_0}} \exp\left(-\frac{N_0 s_v}{2\sigma^2_{gv}}\right), \tag{7}$$

Algorithm 1 MAP-EM diagonal covariance GMM

```
for $i = 1$ to $I$ do
    for $n = 1$ to $N$, $g = 1$ to $G$ do
        $\pi_{ng} \leftarrow P(Z_n = g \mid x_n,\, r_n,\, \theta,\, \mu,\, \sigma)$
    end for
    for $g = 1$ to $G$, $v = 1$ to $V$ do
        $\theta_g \leftarrow N^{-1} \sum_{n=1}^{N} \pi_{ng}$
        $\sigma_{gv}^2 \leftarrow \dfrac{N_0 s_v^2 + \sum_{n=1}^{N} \sum_{t=1}^{T} r_{nvt} \pi_{ng} (x_{nvt} - \mu_{gvt})^2}{N_0 + \sum_{n=1}^{N} r_{nvt} \pi_{ng}}$
        $\mu_{gv} \leftarrow \big(S_v^{-1} + \sigma_{gv}^{-2} \sum_{n=1}^{N} \pi_{ng} \mathrm{diag}(r_{nv})\big)^{-1}$
            $\cdot \big(S_v^{-1} m_v + \sigma_{gv}^{-2} \sum_{n=1}^{N} \pi_{ng} \mathrm{diag}(r_{nv})\, x_n\big)$
    end for
end for
Output $\{\theta,\, \sigma^2,\, \mu\}$ and $\pi_n(q, G) = (\pi_{n1}, \ldots, \pi_{nG})^T$ for $n = 1 : N$, where $q$ represent the initialization.
```

where N_0 is a user-defined hyper-parameter.

The parameters are now estimated using MAP EM as shown in Algorithm 1. The E-step is exactly the same as for maximum-likelihood EM, however the M-step differs because of the priors on the mean and variance. The speed of convergence depends on the initial cluster assignments. Here, we propose to kick-start by initially performing a single round of k-means where missing data are replaced by mean values.

B. *Probabilistic cluster kernel*

MAP EM outcome heavily depends on the initial choices that have to be made; hyper-parameters a_0, b_0, N_0, number of clusters K, initial mean vectors (and standard deviation) μ_{kvt} and σ_{kv}^2. To address this problem we propose to exploit a more robust consensus framework, i.e. the probabilistic cluster kernel. In this work, the framework is applied for the first time to time series containing missing data.

The PCK similarity matrix, K, is built by fitting diagonal covariance Gaussian mixture models to the multivariate time series for a range of number of mixture components. By generating partitions at different resolutions, one can capture both the local and global structure of the data. For each resolution (number of components) the model is estimated using MAP-EM over a range of random initializations and randomly chosen hyper-parameters. The posterior distribution relative to the cluster assignment, that is computed for every time series, is then used to build the PCK matrix following a consensus clustering strategy, where one adds up the contribution from every hyper-parameter configuration and initialization [15]. Algorithm 2 delivers the details of the method.

After having learned the probabilistic cluster kernel K, we apply spectral clustering [2]. Note that we perform clustering twice, first to learn K, thereafter to generate the final partition according to K. A justification for this is given in [14].

Algorithm 2 PCK

```
Input Dataset $X$, $Q$ initializations, $C$ number of clusters.
for $q = 1$ to $Q$, $c = 2$ to $C$ do
    MAP-EM diagonal covariance GMM with $c$ clusters and initialization $q$ over $X \rightarrow \pi_n(q, c)$.
    for $n = 1 : N$, $m = 1 : N$ do
        $K_{nm} \leftarrow K_{nm} + \pi_n(q, c)^T \pi_m(q, c)$
    end for
end for
Output $K$ PCK kernel matrix, MAP-EM diagonal covariance GMM clustering parameters.
```

III. Experiments and results

In this section we apply the proposed method to two different datasets. First, we test our method on a synthetic two-variate time series. Then, we present a real-world dataset consisting of blood tests of patients at the University Hospital of North-Norway (UNN) and we discuss the results obtained.

A. *Simulated two-variate time series generated from a vector autoregressive model*

We generate a two-variate time series dataset from a first-order vector autoregressive model [16], VAR(1), given by

$$\begin{pmatrix} X_t \\ Y_t \end{pmatrix} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} + \begin{pmatrix} \rho_x & 0 \\ 0 & \rho_y \end{pmatrix} \begin{pmatrix} X_{t-1} \\ Y_{t-1} \end{pmatrix} + \begin{pmatrix} u_t \\ v_t \end{pmatrix} \qquad (8)$$

It is easily shown that, if the noise $(u_t,\, v_t)^T$ has zero mean, the α-constants and the mean of the process are related by

$$\begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} = \begin{pmatrix} 1 - \rho_x & 0 \\ 0 & 1 - \rho_y \end{pmatrix} E \begin{pmatrix} X_t \\ Y_t \end{pmatrix}. \qquad (9)$$

To make X_t and Y_t correlated with $corr(X_t, Y_t) = \rho$ it can also be shown that we must choose the noise term such that

$$corr(u_t, v_t) = \rho\, (1 - \rho_x \rho_y)\, [(1 - \rho_x^2)(1 - \rho_y^2)]^{-1} \qquad (10)$$

We simulate 100 two-variate time series of length 50 from the VAR(1)-model with parameters $\rho = \rho_x = \rho_y = 0.8$, $E[(X_t, Y_t)^T] = (0.5, -0.5)^T$. Furthermore, we simulate 100 time series using the parameters $\rho = -0.8$, $\rho_x = \rho_y = 0.6$, $E[(X_t, Y_t)^T] = (0, 0)^T$. Hence, the dataset consists of two clusters.

We apply the MAP EM and the PCK to the simulated VAR(1) dataset with different fractions of missing data. To ensure the MAR assumption is satisfied, we sample uniformly the time intervals and the features that are removed. We choose to discard 0, 10, 20, 30, 40, 50 and 60 % of the values in the dataset.

In our experience the most sensitive hyper-parameter is the bandwidth of the time-kernel $K_{tt'}$. For the MAP EM we use three different choices for a_0: 0.01, 0.1 and 1. We let $b_0 = 0.1$ and $n_0 = 0.01$. We run maximally 50 iterations of the E- and M-steps and report the mean accuracy of 100 runs of the algorithm.

For the PCK we use $g = 2, \ldots, 30$ number of clusters and 30 different, randomly chosen, initializations and sets of hyper-parameters (a_0, b_0, N_0). The initial cluster assignments are made by running one round of k-means on one randomly chosen variable. a_0 is sampled with a uniform distribution from $(0.001, 1)$, b_0 from $(0.005, 0.2)$ and n_0 from $(0.001, 0.2)$. We run 10 iterations of the MAP EM algorithm, varying each time initialization and hyper-parameter configuration. If the reciprocal condition number for S_v is lower than 10^{-15} we draw a new value for a_0.

We compare the proposed similarity measure to different versions of DTW, which extend the basic formulation in order to account for multivariate time series and missing data: *dependent* (d) and *independent* (i) DTW with imputation of the mean (I), imputation using linear interpolation (II), and no imputation but instead using time series of different length (III). DTW returns a dissimilarity matrix, D, from which we compute $K = \exp(-D^2/\sigma)$. To guarantee that this procedure is not affected by a poorly chosen bandwidth, we select σ by performing a grid search and picking the one that yields the best accuracy. Finally, we apply spectral clustering to K.

TABLE I
ACCURACY ON SIMULATED VAR(1) DATASET OBTAINED USING THREE DIFFERENT METHODS; PCK FOLLOWED BY SPECTRAL CLUSTERING, DIAGONAL COVARIANCE GMMS ESTIMATED USING MAP EM AND DTW FOLLOWED BY SPECTRAL CLUSTERING.

% missing	0	10	20	30	40	50	60
PCK	0.94	0.94	0.91	0.92	0.90	0.89	0.87
MAP, $a_0 = 0.01$	Covariance matrix not invertible						
MAP, $a_0 = 0.1$	0.79	0.80	0.78	0.79	0.79	0.77	0.77
MAP, $a_0 = 1$	0.80	0.82	0.80	0.82	0.82	0.81	0.83
DTW (d) I	0.76	0.75	0.75	0.73	0.73	0.70	0.66
DTW (d) II	0.76	0.75	0.74	0.75	0.74	0.74	0.73
DTW (d) III	0.76	0.74	0.73	0.72	0.70	0.65	0.65
DTW (i) I	0.64	0.74	0.75	0.69	0.80	0.55	0.67
DTW (i) II	0.64	0.69	0.71	0.72	0.74	0.71	0.69
DTW (i) III	0.64	0.65	0.55	0.59	0.62	0.57	0.52

The 6 last rows in Table I shows the accuracy obtained using the different variants of DTW for the different fractions of missing data. The first thing to notice is that the 6 approaches behave quite differently. Hence, it will be difficult to choose the best approach for a new dataset. We also notice that the independent DTW provides unstable results. Dependent DTW with linear interpolation gives the best results and a stable accuracy as function of missing ratio. However, compared to PCK and diagonal covariance GMM, we see that this method gives the worst results. The accuracies obtained using diagonal covariance GMM MAP EM remain stable as the fraction of missing data increase and are, in general, better than the results for DTW. However, the table also reveals a problem with the MAP EM. If the bandwidth for the time-kernel is chosen to small the (reciprocal) condition number of covariance matrix becomes very small. On the other hand, if the bandwidth is chosen to large there will not be any covariance in time.

The results obtained using the PCK are shown in the first row in Table I and are much better than the other two methods. The accuracy is also pretty stable as the fraction of missing data increases. Figure 1 shows the mean in the two different clusters for both variates in the simulated VAR(1) dataset with no missing data obtained using the PCK. The figure also shows examples of one time series from each of the two clusters.

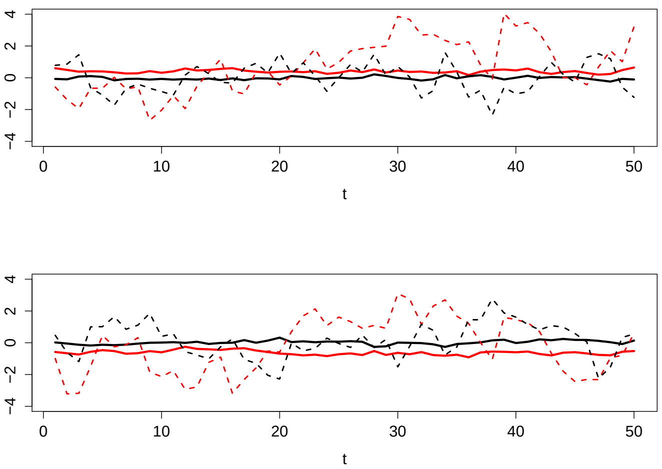

Fig. 1. Mean of the two variables (first variable in the upper plot) in the simulated VAR(1) dataset with 0 % missing data for the two clusters identified by the proposed method. The dashed red lines correspond to one randomly selected time series from the red cluster (non-zero mean and positive correlation) and the dashed black line represents one randomly chosen times series from the cluster generated from the VAR(1) model with zero mean and negative correlation.

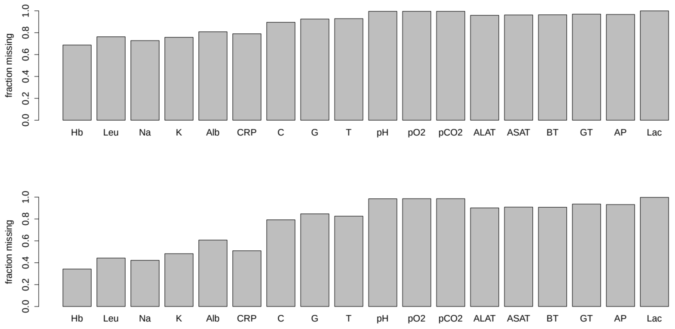

Fig. 2. Top: Fraction of missing data from 10 days before surgery to 20 days after for each blood test. Bottom: Fraction of missing data from the day of surgery to 10 days after for each blood test.

B. Clinical dataset

Through our close collaboration with UNN we have access to the EHR for patients at the department of Gastrointestinal Surgery from 2004 to 2012. In this work we consider physiological measurements and, in particular blood tests. Blood tests from the same database have earlier been used, either as the only source or as one of several sources, for predicting postoperative complications [17]. Here they are for the first time used for clustering.

We have access to all blood tests for 1138 patients that have undergone a major abdominal surgery. The dataset contains all blood tests performed in the period from 10 days before the surgery to 20 days after the surgery. Table II shows the 18 different blood tests available.

TABLE II
NAME OF THE 18 BLOOD MEASUREMENTS WE HAVE ACCESS TO.

Blood test	Abbr.	Unit
Hemoglobin	Hb	g/dl
Leukocytes	Leu	$10^9/l$
Sodium	Na	mmol/l
Potassium	K	mmol/l
Albumin	Alb	g/l
C-Reactive Protein	CRP	mg/l
Creatinine	C	
Glucose	G	
Trombocytes	T	
potentia Hydrogenii	pH	
Pressure of oxygen	pO_2	
Pressure of carbon dioxide	pCO_2	
Alanine aminotransferase	ALAT	
Aspartate aminotransferase	ASAT	
Bilrubin total	BT	
Gamma-glutamyl transferase	GT	
Alkaline phosphatase	AP	
Lactate	Lac	

In each case, we see that the data have the usual characteristics of clinical time series [4], namely *multiple variables* (18 different blood tests), *short sampling period*, *irregular samples* and *length variability*. In fact, measurements span from 10 days before to 20 days after surgery, but for most patients there are measurements only for a few of these days and not all blood tests are reported. This is illustrated in Figure 2, where we have plotted the fraction of missing data for the different blood tests in two different time frames. We see that for 6 blood tests, hemoglobin, leukocytes, sodium, potassium, albumin and CRP, the missing ratio is approximately 50 percent in the time frame from the day of surgery to 10 days after. For the rest of the blood tests, the missing ratio is higher than 80 percent. If we consider a larger time frame from 10 days before surgery to 20 days after the ratio of missing data is even higher. For this reason, we consider only these 6 blood tests and exclude the rest of them. We restrict ourselves to all available measurements in the time frame of length 10 days, starting at the day of surgery. None of the patients have more than one measurement each day.

To ensure that we consider patients where the ratio of missing data is not too high, we create a cohort that consists of those patients that have at least 5 measurements of each of the 6 blood tests after the surgery. Hence patients that leave the hospital within 5 days after the surgery are not considered. This results in a cohort consisting of 139 patients.

We apply the PCK to this dataset and use exactly the same setup as we did for the simulated dataset. We run spectral clustering using the learned PCK similarity and report results for two and three clusters.

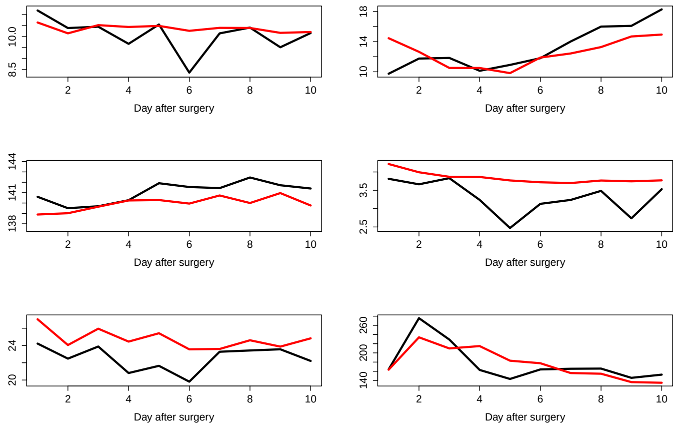

Fig. 3. Plots of the mean as function of days after surgery for the 6 different blood tests obtained running spectral clustering with two clusters using the learned PCK similarity. Black represents cluster 1 and red cluster 2. Top left: Hb, Top right: Leukocytes, middle left: Sodium. Middle right: Potassium. Bottom left: albumin. Bottom right: CRP.

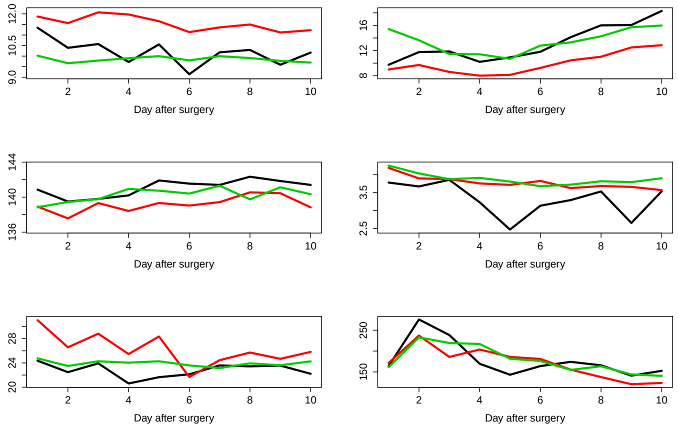

Fig. 4. Plots of the mean for the 6 different blood tests obtained running spectral clustering with three clusters.

TABLE III
CHARACTERISTICS (MEAN AGE, FRACTION MEN, FRACTION ELECTIVE SURGERIES, FRACTION LAPAROSCOPIC SURGERIES, MEAN ASA-SCORE AND FRACTION DEAD AT DISCHARGE) IN THE DIFFERENT CLUSTERS OBTAINED USING PCK ON THE BLOOD DATA.

	age	frac. men	elective	lapa	stoma	ASA	dead
	Two clusters of size 17 and 122						
1	73	0.59	0.31	0.06	0.59	2.75	0.18
2	66	0.56	0.66	0.11	0.28	2.58	0.07
	Three clusters of size 16, 38 and 85						
1	75	0.56	0.31	0.06	0.63	2.73	0.19
2	66	0.74	0.74	0.16	0.32	2.57	0.02
3	66	0.48	0.62	0.09	0.26	2.59	0.08

Figure 3 and Figure 4 show the mean for each of the six blood tests hemoglobin, leukocytes, sodium, potassium, albumin and CRP in the 2 (or 3) clusters obtained using PCK followed by spectral clustering. Table III shows some characteristics of the clusters that are obtained. It seems like cluster 1 contains weaker patients, they are older (73 and 75 years), have a higher fraction of non-planned surgeries (76 and 75 percent) where most of them are open (not laparoscopic) (94 %). Most of them have had a stoma surgery (59 and 63 %). They also have a slightly higher ASA score, which shows that their general health status is weaker than for the general population. This is confirmed by the fact that a larger fraction of the patients in cluster 1 are dead at discharge.

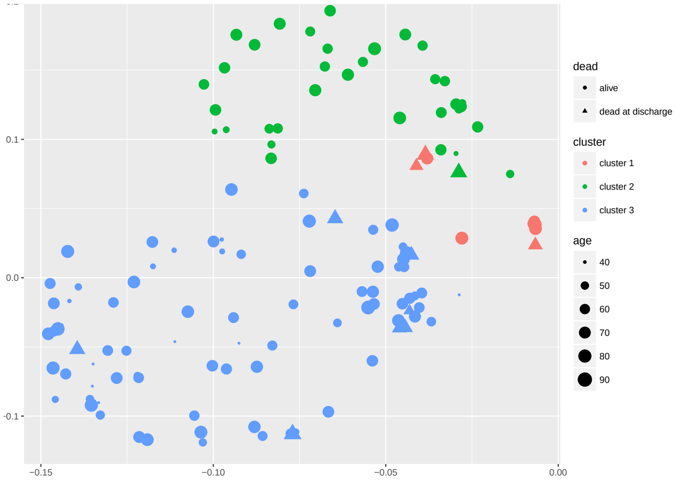

Fig. 5. Plot of first eigenvector of the PCK matrix versus the second eigenvector. The colors indicate the cluster assignment, the size age and the symbol whether the patient is dead or alive at discharge.

By focusing on the two cluster case (Figure3) we can see that in general, even though the differences are not very clear everywhere, cluster 1 has lower hemoglobin levels, a leukocytes level that increases more than for cluster 2, a higher sodium level, a lower potassium level, a lower albumin level and a high CRP level immediately after the surgery. All these things (except the fact that the sodium level in cluster 1 is higher than in cluster 2) could be indicators of complications such as e.g. postoperative delirium.

Figure 5 shows an embedding of the EHR time series data in the PCK space. The different colors represent the three different clusters. Triangles represent patients that died at the hospital after the surgery. We see that most of them are in cluster 1 (red) or in close proximity. The size of the symbols represents the age of the patients. The age distribution is not clear, but it seems like most of the younger patients are placed to the left in cluster 3 (blue).

The results presented above suggest that the proposed method is able to identify sub-cohorts of patients in the data of different characteristics. In particular, it is interesting to see that one of the clusters contains patients that in general are weak and probably more exposed to severe postoperative complications. We are in the process of investigating these results via our close collaboration with the clinicians at UNN. However, due to time limitations a more thorough interpretation of the results is left for further work.

IV. Conclusion

In this work we have introduced the probabilistic cluster kernel for (possibly) irregularly sampled multivariate time series. This kernel learns similarities at different scales and is robust in the sense that it is parameter-free. It can be used as the starting-point for all kinds of machine learning methods, not only clustering.

By applying the PCK to simulated data and comparing to other methods we have shown that the proposed method is robust and provides good results. The method managed to identify clinically interesting sub-cohorts in a blood dataset. Currently we are working on extensions that will include more detailed comparisons with existing methodologies and further applications to real world scenarios.

Acknowledgement

This work (Robert Jenssen and Filippo Bianchi) is partially supported by the Research Council of Norway over FRIPRO grant no. 234498 on developing the *Next Generation Learning Machines*. Cristina Soguero-Ruiz is supported by FPU grant AP2012-4225 from Spanish Government.

References

[1] J. Hu, A. Perer, and F. Wang, *Data Driven Analytics for Personalized Healthcare*, pp. 529–554. Cham: Springer International Publishing, 2016.

[2] S. Theodoridis and K. Koutroumbas, *Pattern Recognition, Fourth Edition*. Academic Press, 4th ed., 2008.

[3] M. Müller, "Dynamic time warping," *Information retrieval for music and motion*, pp. 69–84, 2007.

[4] Z. Liu and M. Hauskrecht, "Learning adaptive forecasting models from irregularly sampled multivariate clinical data," in *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[5] C. Soguero-Ruiz, W. M. Fei, R. Jenssen, K. M. Augestad, J.-L. R. Álvarez, I. M. Jiménez, R.-O. Lindsetmo, and S. O. Skrøvseth, "Data-driven temporal prediction of surgical site infection," in *AMIA Annual Symposium Proceedings*, vol. 2015, pp. 1164 –1173, American Medical Informatics Association, 2015.

[6] M. Shokoohi-Yekta, B. Hu, H. Jin, J. Wang, and E. Keogh, "Generalizing DTW to the multi-dimensional case requires an adaptive approach," *Data Mining and Knowledge Discovery*, pp. 1–31, 2016.

[7] C. A. Ratanamahatana and E. Keogh, "Three myths about dynamic time warping data," in *Mining, in the Proceedings of SIAM International Conference on Data Mining*, pp. 506–510, 2005.

[8] L. P. Cruz, S. M. Vieira, and S. Vinga, "Fuzzy clustering for incomplete short time series data," in *Portuguese Conference on Artificial Intelligence*, pp. 353–359, Springer, 2015.

[9] M. Ghassemi, M. A. F. Pimentel, T. Naumann, T. Brennan, D. A. Clifton, P. Szolovits, and M. Feng, "A multivariate timeseries modeling approach to severity of illness assessment and forecasting in ICU with sparse, heterogeneous clinical data," in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, pp. 446–453, AAAI Press, 2015.

[10] T. A. Lasko, J. C. Denny, and M. A. Levy, "Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data," *PLoS ONE*, vol. 8, pp. 1–13, 06 2013.

[11] B. M. Marlin, D. C. Kale, R. G. Khemani, and R. C. Wetzel, "Unsupervised pattern discovery in electronic health care data using probabilistic clustering models," in *Proceedings of the 2Nd ACM SIGHIT International Health Informatics Symposium*, IHI '12, (New York, NY, USA), pp. 389–398, ACM, 2012.

[12] D. B. Rubin, "Inference and missing data," *Biometrika*, vol. 63, no. 3, pp. 581–592, 1976.

[13] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the royal statistical society. Series B (methodological)*, pp. 1–38, 1977.

[14] E. Izquierdo-Verdiguier, R. Jenssen, L. Gómez-Chova, and G. Camps-Valls, "Spectral clustering with the probabilistic cluster kernel," *Neurocomputing*, vol. 149, Part C, pp. 1299 – 1304, 2015.

[15] A. L. Fred and A. K. Jain, "Combining multiple clusterings using evidence accumulation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 6, pp. 835–850, 2005.

[16] W. W.-S. Wei, *Time series analysis*. Addison-Wesley publ Reading, 1994.

[17] C. Soguero-Ruiz, K. Hindberg, I. Mora-Jiménez, J. L. Rojo-Álvarez, S. O. Skrøvseth, F. Godtliebsen, K. Mortensen, A. Revhaug, R.-O. Lindsetmo, K. M. Augestad, and R. Jenssen, "Predicting colorectal surgical complications using heterogeneous clinical data and kernel methods," *Journal of Biomedical Informatics*, vol. 61, pp. 87 – 96, 2016.