A k-means approach to clustering disease progressions

Duc Thanh Anh Luong
Department of Computer Science and Engineering
University at Buffalo
Buffalo, New York 14260
Email: ducthanh@buffalo.edu

Varun Chandola
Department of Computer Science and Engineering
University at Buffalo
Buffalo, New York 14260
Email: chandola@buffalo.edu

***Abstract*—K-means algorithm has been a workhorse of unsupervised machine learning for many decades, primarily owing to its simplicity and efficiency. The algorithm requires availability of two key operations on the data, first, a distance metric to compare a pair of data objects, and second, a way to compute a representative (centroid) for a given set of data objects. These two requirements mean that k-means cannot be readily applied to time series data, in particular, to disease progression profiles often encountered in healthcare analysis. We present a k-means inspired approach to clustering disease progression data. The proposed method represents a cluster as a set of weights corresponding to a set of splines fitted to the time series data and uses the "goodness-of-fit" as a way to assign time series to clusters. We use the algorithm to group patients suffering from Chronic Kidney Disease (CKD) based on their disease progression profiles. A qualitative analysis of the representative profiles for the learnt clusters reveals that this simple approach can be used to identify groups of patients with interesting clinical characteristics. Additionally, we show how the representative profiles can be combined with patient's observations to obtain an accurate patient specific profile that can be used for extrapolating into the future.**

***Index Terms*—K-means, clustering, Chronic Kidney Disease**

I. Introduction

With the growing adoption of electronic health records (EHRs) by health practices, more and more clinical data is available for research and analysis. This data is particularly useful to understand the course of disease progression for patients. One data analysis task that has gained prominence in recent years, owing to the emphasis on precision medicine [1], is the task of identifying patient subpopulations that exhibit similar progression of the target disease [2]. Under the hypothesis that patients with similar disease progression are likely to share a common disease mechanism (or a phenotype), such discoveries can allow for identifying disease subtypes, a cornerstone of precision medicine [3].

Clearly, clustering methods [4] can be employed to identify the target subpopulations by clustering the disease progression data, which is often available in the form of longitudinal clinical observations (or time series). However, applying standard clustering algorithms, such as k-means, to this data is not straightforward. The primary reason is that most of these algorithms require a *similarity metric* to compare pairs of time series. In the past, researchers have employed time series proximity measures such as cross-correlation, Dynamic Time Warping (DTW), etc., to handle this issue for time series data [5]. However, clinical time series are typically sparse and not necessarily aligned across patients, which makes measures such as DTW ill-suited in this context. Alternatively, researchers have come up with model-based clustering solutions that either involve computationally complex statistical inference [6], or make strict assumptions regarding the nature of the time series data [7].

In this paper, we present an adaptation of the k-means algorithm to admit sparse and irregularly sampled time series data. The core assumption is that the time series within a cluster can be approximated using a weighted sum over a collection of splines or polynomial functions. The weight coefficients are used to represent each cluster (centroid). Each iteration involves estimating the coefficients using the current set of time series in a cluster and then reassigning individual time series to the cluster that gives the best "fit".

The proposed adaptation retains the attractive properties of k-means, viz., efficiency and simplicity, while handling complex time series data. We apply the proposed method to identify patient groups based on their disease progression profiles for a patient cohort suffering from Chronic Kidney Disease (CKD). The identified clusters, obtained by clustering the longitudinal observations of *estimated glomerular filtration rate* (eGFR), reveal interesting patient subtypes for CKD, which are then further analyzed using other information present in the corresponding EHRs.

Additionally, we show how the cluster representative information can be combined with the sparse observations to obtain a patient-specific disease progression profile that can be used for missing data imputation or predicting CKD progression. We show that the predictions obtained using our proposed method that uses only eGFR data closely align

978-1-5090-4881-6/17 $31.00 © 2017 IEEE
DOI 10.1109/ICHI.2017.18

with the predictions using a model tailored for CKD that utilizes multiple patient related inputs [8].

II. Modified k-means algorithm

In this section, we describe the problem of clustering patient disease profiles and present an adaptation of k-means algorithm to solve it. The complexity analysis presented in Section II-C shows that this algorithm grows linearly with the number of measurements in the dataset and can be scalable for large EHR datasets.

A. Problem statement

Let N be the number of patients in our dataset. For each patient i, we denote n_i as the number of lab measurements that a patient has in his/her health record. The vector of lab measurements for patient i is denoted as $\mathbf{x}_i = [x_{i,1}, x_{i,2}, \cdots, x_{i,n_i}]^T$ while $\mathbf{t}_i = [t_{i,1}, t_{i,2}, \cdots, t_{i,n_i}]^T$ is the vector of corresponding timestamps.

Assuming that there are K clusters of patients where K is known as a priori. The cluster assignment of patient i belonging to cluster k is represented by a vector $\mathbf{z}_i \in \{0,1\}^K$ where $z_{i,k} = 1$ and $z_{i,j} = 0$ for $j \neq k$.

For each cluster, the joint progression of all patients belonging to this cluster is represented by a fitted curve using all observations of patients assigned to this cluster. We denote the fitted curve of the cluster k as $f_k(t) = \sum_{l=1}^{L} \boldsymbol{\beta}_l^{(k)} \Phi_l(t)$ where $\Phi(\cdot) = \{\Phi_1(\cdot), \Phi_2(\cdot), \cdots, \Phi_L(\cdot)\}$ is the a set of L basis functions and $\boldsymbol{\beta}^{(k)} = \{\beta_1^{(k)}, \beta_2^{(k)}, \cdots, \beta_L^{(k)}\}$ is the set of corresponding coefficients. For an input timestamp vector, $\mathbf{t}_i \in \Re^{n_i}$, we denote $\Phi(\mathbf{t}_i) = [\Phi_1(\mathbf{t}_i), \Phi_2(\mathbf{t}_i), \cdots, \Phi_L(\mathbf{t}_i)] \in \Re^{n_i \times L}$ as a matrix in which the l^{th} column is the vector of basis function Φ_l applied to each element in $\mathbf{t}_i$.

The dissimilarity between a patient i and a cluster k can be measured using the sum of square difference between his/her actual lab measures and its representative progression of cluster k, and is computed as $\left\|\mathbf{x}_i - \Phi(\mathbf{t}_i)\boldsymbol{\beta}^{(k)}\right\|_2^2$.

With the aim of clustering the set of N patients into K clusters, we can write the error function as follow:

$$E(\{\mathbf{z}_i\}_{i=1}^N, \{\boldsymbol{\beta}^{(k)}\}_{k=1}^K) = \sum_{k=1}^{K} \sum_{i=1}^{N} z_{i,k} \left\|\mathbf{x}_i - \Phi(\boldsymbol{t}_i)\boldsymbol{\beta}^{(k)}\right\|_2^2 \tag{1}$$

In particular, we want to determine the cluster assignment $\mathbf{z}_i$ for the i^{th} patient, and the set of coefficients $\boldsymbol{\beta}^{(k)}$ for the basis functions in each cluster k, such that the error $E()$ is minimized. One can observe that this error function is similar to distortion measure which is used as the objective function of k-means algorithm [9, p. 424-425] except that the square distance between the data point and cluster center is replaced by our own measure of dissimilarity between patient's trajectory and cluster's trajectory. In the next section, we describe the modified k-means algorithm which optimizes the objective function in (1).

B. Algorithm

Similar to k-means algorithm, the error function in (1) can be minimized through an iterative process in which we first randomly assign the patients to K clusters and then perform following two steps: (1) optimize $E()$ with respect to $\{\boldsymbol{\beta}^{(k)}\}_{k=1}^K$, while keeping all $\{\mathbf{z}_i\}_{i=1}^N$ fixed, and (2) optimize $E()$ with respect to $\{\mathbf{z}_i\}_{i=1}^N$ while keeping all $\{\boldsymbol{\beta}^{(k)}\}_{k=1}^K$ fixed. This iterative process continues until convergence occurs. Step (1) can be achieved by using the method of least squares for each cluster k. Step (2) can be easily done by setting $z_{i,k} = 1$ when $k = \arg\min_j \left\|\boldsymbol{x}_i - \Phi(\boldsymbol{t}_i)\boldsymbol{\beta}^{(j)}\right\|_2^2$ and $z_{i,j} = 0$ for $j \neq k$. Similar to k-means algorithm, this iterative procedure can get stuck in local optima and may require multiple restarts with random initializations to achieve the best result.

The pseudocode for this modified k-means algorithm can be written as follow:

Initialization:
for $i \in 1, \cdots, N$ **do**
  randomly assign patient i to a cluster
end for
repeat
  Update step:
  for $k \in 1, \cdots, K$ **do**
    compute $\boldsymbol{\beta}^k$ which minimizes $\sum_{i=1}^{N} \sum_{k=1}^{K} z_{i,k} \left\|\mathbf{x}_i - \Phi(\mathbf{t}_i)\boldsymbol{\beta}^{(k)}\right\|_2^2$ using the method of least squares
  end for
  Assignment step:
  for $i \in 1, \cdots, N$ **do**
    set $z_{i,k} = 1$ where $k = \arg\min_j \left\|\mathbf{x}_i - \Phi(\mathbf{t}_i)\boldsymbol{\beta}^{(j)}\right\|_2^2$ and set $z_{i,j} = 0$ for $j \neq k$
  end for
until Convergence ($\{\mathbf{z}_i\}_{i=1}^N$ does not change)

C. Complexity analysis

We denote $T(H, L, K, M)$ as the complexity of the modified k-means algorithm where H, L, K and M are maximum number of iterations, number of basis functions, number of clusters and total number of measurements respectively. It is easy to observe from the pseudo-code that initialization step only has a negligible computational cost ($O(KN)$ as we need to initialize the values of all $\mathbf{z}_i$) in comparison with the rest of the algorithm. Therefore, we only focus on analyzing the complexity of the iterative process of update and assignment steps. This complexity can be bounded by maximum number of iterations times total complexity of update and assignment steps. We denote T_{update} and $T_{assignment}$ as complexity of update step and complexity of assignment step respectively. With these

notations, the complexity of the algorithm can be bounded by $H(T_{update} + T_{assignment})$.

In the update step, the cost of performing least square for cluster k is $O(L^3+L^2M_k)$ where M_k is the total number of measurements of all patients currently assigned to cluster k. Since the update step requires performing the method of least squares for all clusters, the total complexity of update step is $O(KL^3 + L^2M)$.

In the assignment step, we need to compute the difference $\left\|\mathbf{x}_i - \Phi(\mathbf{t}_i)\boldsymbol{\beta}^{(k)}\right\|_2^2$ for each patient i and each cluster k which costs $O(n_iL)$ where n_i is the number of measurements of patient i. Collectively, using the fact that $M = \sum_{i=1}^{N} n_i$, the total complexity of assignment step is $O(LKM)$.

Overall, the complexity of the modified k-means algorithm is $O(H(KL^3 + L^2M + LKM))$. In actual applications when we usually choose a fixed value for H, L and K, the complexity of the modified k-mean algorithm grows linearly with the number of measurements M in the dataset.

III. Experiments

In this section, we demonstrate the use of the modified k-means algorithm presented in Section II for understanding Chronic Kidney Disease (CKD), a wide-spread disease in both the US and worldwide [10]. In Section III-A, we briefly describe the elements in the dataset as well as the preprocessing steps used for obtaining the target cohort. In Section III-B, the clustering results are presented. A qualitative analysis of clustering result using demographic information and related clinical markers is discussed in Section III-C. Finally, we show how to use the clustering output to obtain patient-specific disease progressions in Section III-D.

A. Data

The data used in the experiments is a subset of a dataset collected by DARTNet Institute [11] which contains eletronic health records of 69,817 patients having various degree of kidney damage. The progression of CKD is typically monitored using a clinical indicator called *estimated Glomerular Filtration Rate* (eGFR) which measures the patient's kidney function. This eGFR value can be estimated using the CKD-EPI equation which takes into account serum creatinine measure as well as patient's age, sex and race [12]. We further refine the CKD cohort by only retaining the patients with eGFR values less than 60 for more than three months. This criteria is used for determining CKD according to clinical guidelines [13]. Moreover, in order to ensure data quality for analysis, we only retain patients with at least one year data record of eGFR values and have more than ten measurements of serum creatinine. The exact preprocessing steps are outlined in Fig. 1. This preprocessing procedure is similar to one in a study of Luong et al. [14].

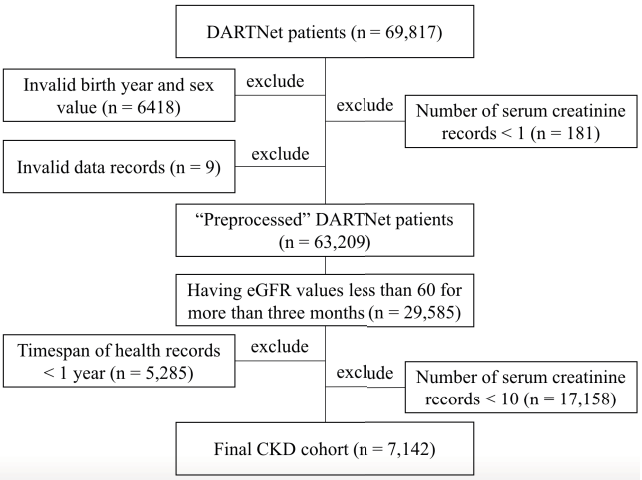


Fig. 1: Preprocessing procedure to obtain CKD cohort

B. Clustering result

We apply the modified k-means algorithm to the eGFR longitudinal data of CKD cohort. In our experiment, we use a linear combination of ten cubic b-spline basis functions with an addition of intercept term to represent cluster trajectory. The knots are chosen based on the quantile of the data. The number of clusters K can be determined by running the algorithm for different values of K and evaluate its corresponding BIC or AIC values. In this experiment, we evaluate the clustering result for $K = 10$.

Fig. 2 shows the output clusters. For each cluster we show the representative profile along with disease profiles of fifty patients who are best fitted for each cluster. In this figure, the representative profile of each cluster is a fitted curve obtained by using all observations of all patients assigned to the cluster.

From Fig. 2, one can observe that CKD patients have various courses of disease progression which can be approximately identified using the modified k-means algorithm. The clusters are presented in this figure from best prognosis to worst. Note that an eGFR value of 60 is typically considered as a cutoff, below which a patient is considered to have CKD [13]. The general disease progressions from clusters 1 to 9 slightly change with more declining trajectories in later clusters. Clusters 1-3 with slightly improving CKD progressions contain 34.49% of total patients in CKD cohorts. This particular set of clusters is interesting because the patients in these clusters have improving CKD progressions, which may be interesting candidates for further investigation of their biological basis and clinical treatments they received. Clusters 4-9 mostly

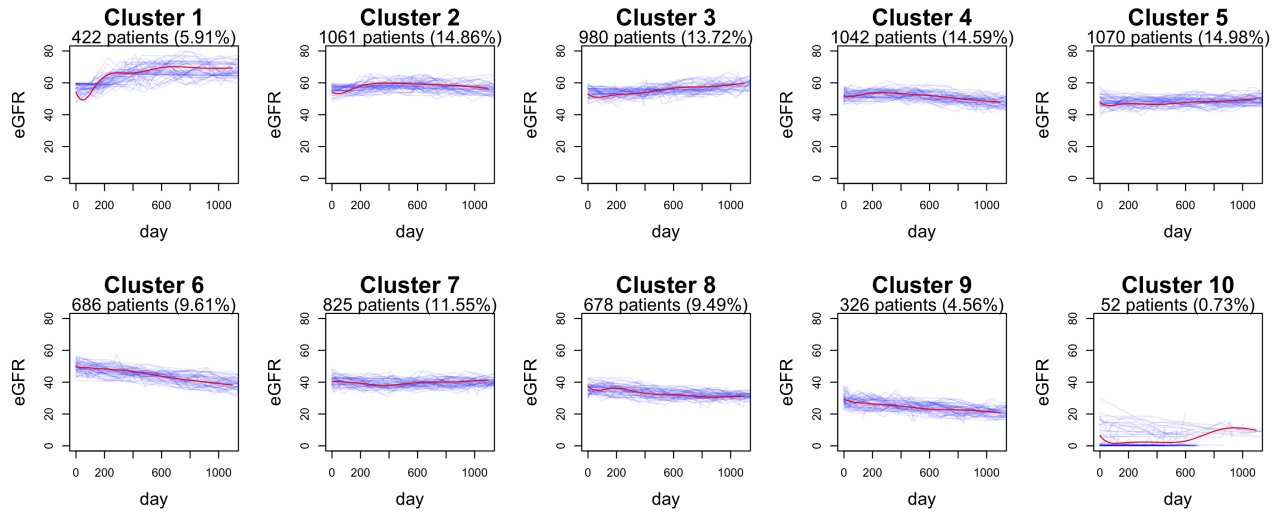


Fig. 2: Clustering output of k-means algorithm applied to the CKD dataset. In blue are the raw patient observations and in red are the representative disease profiles. Number of patients and its proportion in CKD cohort are given for each cluster.

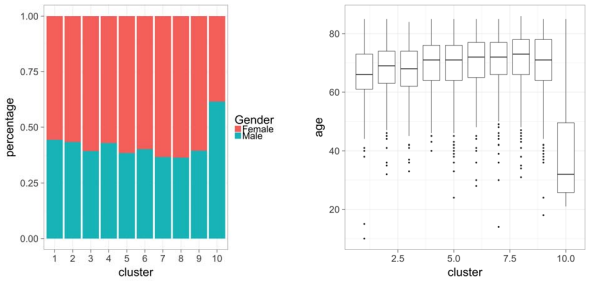
contain patients in stage 3 and stage 4 CKD with different rates of decline and baseline eGFR values. One can note that the differences between clusters 4-9 are not much and perhaps with longer records of eGFR values, these clusters may be merged. Cluster 10 stands out from other clusters because of its very low eGFR trajectories which indicatate most patients in this cluster have very serious kidney damage.

C. Clinical relevance of output clusters

In this section, we investigate the clustering output in terms of demographic information as well as other relevant clinical markers.

Fig. 3 shows the distribution of patients within each cluster along various demographic parameters. The distribution of gender for each cluster in Fig. 3a aligns with the general proportion of men and women in the target age group [15]. It is interesting that for most clusters, gender does not play any additional role. This could be attributed to the fact that the eGFR calculation (CKD-EPI equation) already accounts for gender when estimating eGFR. On the other hand, the distribution of age for each cluster in Fig. 3b shows that the more severe CKD subgroup tends to have older patients with the exception of cluster 10. We also note that cluster 10 is an anomaly in terms of gender and age distribution, as it consists of mostly younger patients and marginally higher proportion of male patients in comparison with other clusters.

For a closer look at the clinical conditions of patients with respect to their clusters, we analyze some related clinical markers associated with those patients. Fig. 4 shows the



(a) The distribution of gender for each cluster

(b) The distribution of baseline age for each cluster

Fig. 3: Demographic distribution for each cluster

distributions of 11 related clinical markers for each cluster. These clinical markers show the risk of patients associated with kidney disease and other interrelated comorbidities. The first column in Fig. 4 contains three clinical markers including albumin-to-creatinine ratio (ACr), phosphorous (Phos) and parathyroid hormone (PTH) that indicate the condition of kidney function. The higher the values of these three markers indicate more deterioration of kidney function [16–18]. As can be seen from the figure, although the trends of those clinical markers are varied among clusters 1-9, cluster 10 distinguishes itself from others as the most severe cluster and contains patients who have substantial kidney damage. The second column in Fig. 4 contains three clinical markers related to diabetes - a common comorbidity among CKD patients. The higher values of these three markers show higher risk of having diabetes [19,20]. From

the distribution of these three markers with respect to clusters, there is no indication of high association between the clusters of CKD patients and risk of diabetes. Three clinical markers in the third column of Fig. 4 are measures of cholesterone and indicators of risk for heart disease. The lower value of high density lipoprotein (HDL) as well as higher value of low density lipoprotein (LDL) and triglyceride level (Trig) indicate higher risk of heart disease [21]. Although there is no distinguishable trends in HDL and Trig among clusters, LDL values in cluster 10 are slightly higher than remaining clusters showing higher risk of heart disease associated with cluster 10. Finally, the two clinical markers - alanine aminotransferase (ALT) and aspartate aminotransferase (AST) in the last column of Fig. 4 are the measurements of enzymes that are used to see if liver is damaged. In particular, lower values of ALT and AST represent more deterioration of liver function. In this last column, one can observe the relationship between clusters and levels of these two enzyms in which the more severe CKD clusters have lower values of ALT and AST. This observation also agrees with other study that analyzes the relationship between CKD and hepatic diseases [22].

D. Generating patient-specific disease profiles

In this section, we show how the representative disease profile for a cluster can be used to obtain patient-specific disease profiles for patients belonging to that cluster. We employ Gaussian Process Regression (GPR) [23], a widely used non-parametric method to predict the eGFR value for patient i, at any time instance t, denoted as $x_i(t)$. Following the previously defined notation from Section II, let $f_k(t)$ be the value of the representative profile for cluster k at time t and let patient i belong to cluster k. The GPR formulation specifies that the observation $x_i(t)$ is connected to the t through a latent function $g_i(t)$, such that $x_i(t) \sim \mathcal{N}(g_i(t), \sigma_i^2)$. The latent function obeys a GP prior, i.e., $g_i(t) \sim GP(f_k(t), \kappa_i)$, where κ_i is a covariance function defined over a pair of time instances. The GP prior essentially specifies that the collection of values taken by function g_i for any vector of time instances, $[t_1, t_2, \ldots, t_m]^T$ is a multivariate Gaussian distribution with mean specified by the vector $[f_k(t_1), f_k(t_2), \ldots, f_k(t_m)]^T$ and covariance matrix, $\mathcal{K}_i$ obtained by applying the covariance function κ_i to every pair of time instances, i.e., $\mathcal{K}_i[m,n] = \kappa_i(t_m, t_n)$. In this paper we use the squared exponential covariance function, i.e., $\kappa_i(t_m, t_n) = a_i \exp\left(\frac{(t_m - t_n)^2}{2l_i^2}\right)$. Assuming that we have observed the eGFR values $\mathbf{y}_i$ with corresponding timestamps $\mathbf{x}_i$ for patient i, the GP prior and the link between $g_i(t)$ and $x_i(t)$ allows us to obtain a prediction at a new time instance t_* as a normal distribution with following mean and variance:

$$\begin{aligned} mean(x_i(t_*)) &= f_k(t_*) + {\boldsymbol{\kappa}_*}^\top (\mathcal{K}_i + \sigma_i^2 I)^{-1} \mathbf{y}_i \\ var(x_i(t_*)) &= \kappa_i(t_*, t_*) - {\boldsymbol{\kappa}_*}^\top (\mathcal{K}_i + \sigma_i^2 I)^{-1} \boldsymbol{\kappa}_* \end{aligned}$$

where $\mathcal{K}_i$ is the covariance matrix of observed timestamps $\mathbf{x}_i$ and $\boldsymbol{\kappa}_*$ is a vector such that $\boldsymbol{\kappa}_*[j] = \kappa(t_*, \mathbf{t}_{ij})$. The hyper-parameters, σ_i^2, a_i, and l_i are learnt by maximizing the log-likelihood of the normal distribution corresponding to the observed data using the R package GPFDA [24].

We use the mean and variance values obtained at regularly sampled time instances to generate the predicted patient profiles and the associated error intervals. Some sample predicted profiles are shown in Fig. 5. The patient-specific profiles generally follow the representative profile for the corresponding cluster, but are "adjusted" according to the sparse observations.

In order to evaluate the quality of prediction, we compare it with a prediction model, tailored for longitudinal eGFR monitoring [8] (we denote this model as *Monitor*). We use the *Salford Royal Hospital Foundation Trust* (SRFT) dataset [8] with same prepocessing steps as shown in Fig. 1. We randomly divide the dataset into five parts and use 5-fold cross-validation with root-mean-square error (or RMSE) as the measure of prediction error. In particular, the RMSE is calculated for the validation set of each fold and the overall RMSE $\sqrt{\frac{RMSE_1^2 + \cdots + RMSE_5^2}{5}}$ is reported. Table I shows the overall RMSE of 5-fold cross-validation for both our model and *Monitor* model. It is important to note that although our model does not give as good prediction as *Monitor* model, our model is not built for prediction while *Monitor* is specialized for monitoring and predicting eGFR. In addition, *Monitor* model uses more information such as baseline age and gender while our model only uses the collection of eGFR trajectories as input. Even then, the performance of our model is not significantly poorer than *Monitor* model.

Model	Overall RMSE
Modified k-means	8.98
Monitor	6.66

TABLE I: Comparison of prediction error in SRFT dataset

IV. Discussion and Related Work

Perhaps the most similar model with our approach is Probabilistic Subtyping Model (PSM) [25] which can explain individual disease progression using different components such as subtype effect, covariate effect, long-term individual effect and short-term individual effect. Luong et al. [14] applied PSM for Chronic Kidney Disease data and empirically identified five subtypes which were evaluated for their clinical relevance. Our k-means approach for clustering disease progression can also be interpreted as

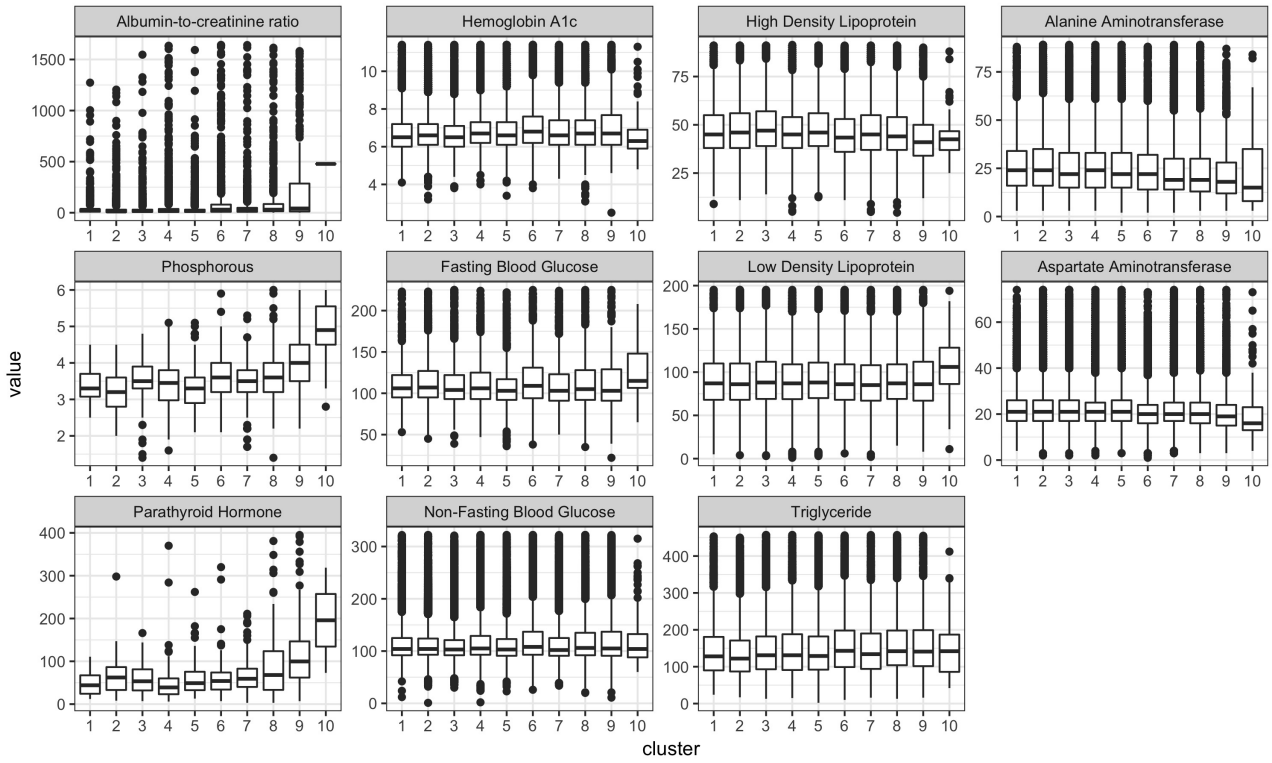


Fig. 4: The distribution of various clinical markers for each cluster

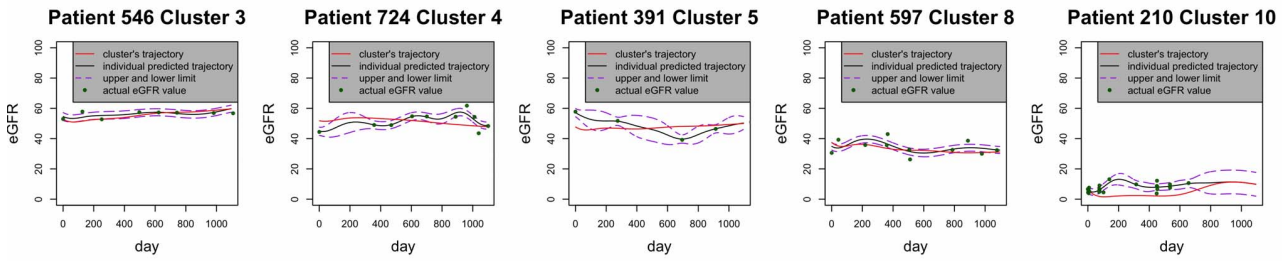


Fig. 5: Patient specific disease profiles obtained using Gaussian Process Regression for five patients.

a hard-clustering version of PSM when only consider the subtype effect and individual short-term effect. Although PSM is an elegant model with probabilistic interpretation, choosing a correct set of hyper-parameters is difficult to achieve and depends on the context of the problem.

In the context of finding disease subtypes using clinical markers, there are few other models that have been proposed in previous studies. Schulam et al. [26] proposed a method to map individual disease trajectory into a low dimensional space and subsequently cluster those trajectories in low dimensional representation using hierarchical clustering. Rusanov et al. [27] presented a method to cluster lab measurements in EHRs data by binning lab values in each 4 months period and imputing missing values before extracting features with discrete wavelet transforms and measuring the pairwise distance between time-series.

From the perspective of using k-means for longitudinal data, there is an R package "KmL" [28] that works specifically for this situation. However, in order to use this package, user has either to choose to have equal length time-series with missing values imputed or use costly distance metric such as dynamic time warping. In the context of clustering disease progression, the choice for using dynamic time warping as a distance metric for unequal length time-series is questionable as the change of eGFR over a long period is very different from the change

over short period.

V. Conclusion

This paper propose a scalable method for stratifying patients with similar disease progressions. The approach is subsequently applied to a CKD dataset to empirically identify ten clusters of patients having similar disease progressions. Our preliminary experimental results show that this approch can find interesting clusters of CKD patient progressions which can be candidates for further analysis on their clinical relevance. The approach can also be generalized to apply to other diseases. Moreover, the proposed approach is also potential for further expansion to cope with multiple clinical markers.

Acknowledgment

We would like to thank Professor. Chester H. Fox for providing us the DARTNet dataset when conducting this research. We also thank two anonymous reviewers for their comments on earlier version of the paper. This material is based in part upon work supported by the National Science Foundation under award numbers CNS - 1409551 and IIS - 1641475.

References

[1] V. Gligorijević, N. Malod-Dognin, and N. Pržulj, "Integrative methods for analyzing big data in precision medicine," *Proteomics*, vol. 16, no. 5, pp. 741–758, 2016.

[2] S. Saria and A. Goldenberg, "Subtyping: What it is and its role in precision medicine," *IEEE Intelligent Systems*, vol. 30, no. 4, pp. 70–75, 2015.

[3] P. N. Robinson, "Deep phenotyping for precision medicine," *Human mutation*, vol. 33, no. 5, pp. 777–780, 2012.

[4] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: A review," *ACM Comput. Surv.*, vol. 31, no. 3, pp. 264–323, Sep. 1999.

[5] T. W. Liao, "Clustering of time series dataa survey," *Pattern Recognition*, vol. 38, no. 11, pp. 1857 – 1874, 2005.

[6] Y. Xiong and D.-Y. Yeung, "Mixtures of ARMA models for model-based time series clustering," in *2002 IEEE International Conference on Data Mining, 2002. Proceedings.*, 2002, pp. 717–720.

[7] C. Li and G. Biswas, "Temporal pattern generation using hidden markov model based unsupervised classification," in *International Symposium on Intelligent Data Analysis*. Springer, 1999, pp. 245–256.

[8] P. J. Diggle, I. Sousa, and . Asar, "Real-time monitoring of progression towards renal failure in primary care patients," *Biostatistics*, vol. 16, no. 3, p. 522, 2015.

[9] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., 2006.

[10] I. Abubakar, T. Tillmann, and A. Banerjee, "Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013." *Lancet*, vol. 385, no. 9963, pp. 117–171, 2015.

[11] W. D. Pace, C. Fox, T. White, D. Graham, L. M. Schilling, and R. David, "The DARTNet institute: seeking a sustainable support mechanism for electronic data enabled research networks," *eGEMs (Generating Evidence & Methods to improve patient outcomes)*, vol. 2, no. 2, p. 6, 2014.

[12] A. S. Levey, L. A. Stevens, C. H. Schmid, Y. L. Zhang, A. F. Castro, H. I. Feldman, J. W. Kusek, P. Eggers, F. Van Lente, T. Greene *et al.*, "A new equation to estimate glomerular filtration rate," *Annals of internal medicine*, vol. 150, no. 9, pp. 604–612, 2009.

[13] National Kidney Foundation, "K/DOQI clinical practice guidelines for chronic kidney disease: evaluation, classification, and stratification." *American journal of kidney diseases: the official journal of the National Kidney Foundation*, vol. 39, no. 2 Suppl 1, p. S1, 2002.

[14] D. T. A. Luong, D. Tran, W. D. Pace, M. Dickinson, J. Vassalotti, J. Carroll, M. Withiam-Leitch, M. Yang, N. Satchidanand, E. Staton, L. S. Kahn, V. Chandola, and C. H. Fox, "Extracting deep phenotypes for chronic kidney disease using electronic health records," *eGEMs (Generating Evidence & Methods to improve patient outcomes)*, vol. 5, no. 1, 2017.

[15] T. Eskes and C. Haanen, "Why do women live longer than men?" *European Journal of Obstetrics & Gynecology and Reproductive Biology*, vol. 133, no. 2, pp. 126–133, 2007.

[16] S. Tomasello, "Secondary hyperparathyroidism and chronic kidney disease," *Diabetes Spectrum*, vol. 21, no. 1, pp. 19–25, 2008.

[17] L. A. Inker, B. C. Astor, C. H. Fox, T. Isakova, J. P. Lash, C. A. Peralta, M. K. Tamura, and H. I. Feldman, "KDOQI us commentary on the 2012 KDIGO clinical practice guideline for the evaluation and management of CKD," *American Journal of Kidney Diseases*, vol. 63, no. 5, pp. 713 – 735, 2014.

[18] J. Da, X. Xie, M. Wolf, S. Disthabanchong, J. Wang, Y. Zha, J. Lv, L. Zhang, and H. Wang, "Serum phosphorus and progression of CKD and mortality: A meta-analysis of cohort studies," *American Journal of Kidney Diseases*, vol. 66, no. 2, pp. 258 – 265, 2015.

[19] S. Rahbar, O. Blumenfeld, and H. M. Ranney, "Studies of an unusual hemoglobin in patients with diabetes mellitus," *Biochemical and Biophysical Research Communications*, vol. 36, no. 5, pp. 838 – 843, 1969.

[20] "Diagnosis and classification of diabetes mellitus," *Diabetes Care*, vol. 37, no. Supplement 1, pp. S81–S90, 2014.

[21] V. Manninen, L. Tenkanen, P. Koskinen, J. K. Huttunen, M. Mänttäri, O. P. Heinonen, and M. H. Frick, "Joint effects of serum triglyceride and LDL cholesterol and HDL cholesterol concentrations on coronary heart disease risk in the Helsinki Heart Study. Implications for treatment." *Circulation*, vol. 85, no. 1, pp. 37–45, 1992.

[22] L. Ray, S. K. Nanda, A. Chatterjee, R. Sarangi, and S. Ganguly, "A comparative study of serum aminotransferases in chronic kidney disease with and without end-stage renal disease: Need for new reference ranges," *International Journal of Applied and Basic Medical Research*, vol. 5, no. 1, p. 31, 2015.

[23] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005.

[24] J. Q. Shi and Y. Cheng, "Gaussian process function data analysis R package GPFDA," 2014.

[25] P. Schulam, F. Wigley, and S. Saria, "Clustering longitudinal clinical marker trajectories from electronic health data: Applications to phenotyping and endotype discovery," 2015.

[26] P. Schulam and R. Arora, "Disease trajectory maps," in *Advances in Neural Information Processing Systems 29*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, Eds. Curran Associates, Inc., 2016, pp. 4709–4717.

[27] A. Rusanov, P. V. Prado, and C. Weng, "Unsupervised time-series clustering over lab data for automatic identification of uncontrolled diabetes," in *2016 IEEE International Conference on Healthcare Informatics (ICHI)*, Oct 2016, pp. 72–80.

[28] C. Genolini and B. Falissard, "Kml: A package to cluster longitudinal data," *Computer Methods and Programs in Biomedicine*, vol. 104, no. 3, pp. e112 – e121, 2011.