Chapter 6

Patient physiological monitoring with machine learning

Marco A. F. Pimentel and David A. Clifton

The task of discovering novel medical knowledge from complex, large-scale and high-dimensional patient data, collected during care episodes, is central to innovation in medicine. The recognition of complex trajectories in multivariate time-series data requires effective models and representations for the analysis and matching of functional data. In this chapter, we describe a method based on Gaussian processes for exploratory data analysis using the observational physiological time-series data.

The method focuses on a representation of unevenly sampled trajectories that allows for revealing physiological recovery patterns in a database of vital signs acquired from post-operative patients. While our primary motivation comes from clinical data, this approach may be applicable to other time-series domains. We first describe methods that have been proposed in the literature for the same purpose. We then provide a brief summary of Gaussian processes, and describe our proposed approach for performing “clustering” of patients’ trajectories.

6.1 Introduction

The task of knowledge discovery from time-series data is important for “tracking” the health status of post-operative patients. An enormous amount of work has been devoted to the task of modelling time-series data.

The autoregressive model is a basic means of analysing time-series data, which specifies that the output variable depends linearly on its previous values. Other examples include state-space models, which are based on the notion that there is an unobserved state of the system, or latent state, that evolves through time and which may only be observed indirectly. For example, the health status of a patient can only be observed through “noisy” observations of the patient’s physiology and mental status.

The most basic state-space model with a continuous-valued latent state is the linear dynamical system (LDS), which is the discrete-time analogue of a linear differential equation. The hidden Markov model (HMM) [1] is the discrete-state space analogue of an LDS. Quinn et al. [2] applied an extension of an LDS model to the problem of monitoring the condition of premature infants receiving intensive care.
A factorial-switching LDS model (equivalent to a switching Kalman filter) was described and tested with continuous time-series data collected from bedside monitors. This model was developed into a hierarchical factorial-switching LDS [3] by adding a set of higher-level variables to model correlations in the physiological factors in order to detect sepsis in ICU patients. Lehman et al. [4] used a switching vector autoregressive framework to systematically learn and identify continuously acquired arterial blood pressure data dynamics. These can possibly be recurrent within the same patient and shared across an entire cohort of ICU patients.

Work by Willsky et al. [5,6] uses Bayesian nonparametric models for capturing the generation of continuous-valued time-series. This method uses an HMM for segmenting time-series data, where the latter are characterised by autoregressive models. Beta processes, which provide prior distributions in the unit interval, are then used to share observation models across several series. Thus, this BP-AR-HMM model is used to capture variability between series by sampling subsets of low-level features that are specific to individual series. Lehman et al. [7] used this model to discover shared dynamics in ICU patients’ continuously acquired blood pressure time-series data. A different Bayesian nonparametric method for exploratory data analysis and feature construction in continuous time-series has been proposed in Reference 8. This method builds on the framework of latent Dirichlet allocation and its extension to hierarchical Dirichlet processes, which allows the characterisation of each series as switching between latent “modes,” where each mode is characterised as a distribution over features that specify the series dynamics. The model was applied to heart-rate data collected from premature infants admitted to a neonatal ICU. A different probabilistic model, the continuous shape template model, has also been applied for discovering time-series’ segments that can repeat within and across different series of continuous heart-rate data [9].

Although conceptually sound, it is unclear how such approaches cope with irregularly sampled data and missing data. As opposed to equally spaced time-series, on which the methods described above have been applied, irregularly sampled time-series data are characterised by variable intervals between successive measurements; i.e., the spacing of observation times is not constant. Different time-series typically contain different numbers of observations and the times at which observations were recorded may not be aligned. Furthermore, periods of missing data are common in clinical scenarios. The properties of these data mean that most common machine learning algorithms and models for supervised and unsupervised learning cannot be directly applied.

One solution to these problems is offered by Gaussian processes. Gaussian processes are a Bayesian modelling technique that has been widely used for various machine learning tasks, such as dimensionality reduction, non-linear classification, and regression [10,11]. It is a nonparametric method, informally suggesting that the number of parameters in the model can grow with the number of observed data. Compared to other related techniques, Gaussian process models have the advantage that prior knowledge of the functional behaviour (e.g., periodicity or smoothness) may be easily expressed. The Bayesian nature of its formulation also means that inference is performed within a probabilistic framework, allowing us to reason in the presence
of noise, incompleteness, and artefacts, all of which are characteristic of the data recorded in hospital settings.

Gaussian processes have been used for modelling physiological time-series data. Clifton et al. [12,13] used Gaussian process regression to cope with artefactual and missing vital-sign data, and incorporated the Gaussian process posterior in their novelty detection schemes. Stegle et al. [14] proposed a robust regression model for noisy heart-rate data based on Gaussian processes and a preliminary clustering procedure that learns the structure of outliers and noise bursts. In the work described in chapter 6 of Reference 15, trend analysis was performed using dependent Gaussian processes, in which the correlation between two or more physiological variables is used to obtain improved regression results. Clifton et al. [16] extended extreme value theory such that a function-wise approach to novelty detection was taken, as opposed to point-wise approaches that are most commonly described in the literature. The method was illustrated using Gaussian process regression, which offers a probabilistic framework in which distributions over a function space are defined. Gaussian process regression has also been used for the ranking of gene expressions [17].

In this work, we propose a representation of vital-sign trajectories using Gaussian process regression, which may be used for the recognition of “normal” and “abnormal” patterns of physiological trends. Figure 6.1 illustrates the components of our proposed approach. We model the evolution of the unevenly sampled physiological trajectories using Gaussian process regression, and we introduce a kernel similarity measurement for the comparison of the latent functions based on the likelihoods of the data points in each trajectory. This patient-to-patient similarity measurement can be used for the functional characterisation of vital-sign trajectories, which may then be used for recognising known trajectories and identifying unknown trajectories as would be required for identifying “abnormal” vital-sign time-series.

6.2 Methodology

In the following sections we describe our proposed approach and analysis conducted.

6.2.1 Dataset

For this analysis we selected a cohort of post-operative patients who stayed for a minimum of 24 hours on the post-operative ward (after upper-gastrointestinal surgery for removal of cancer), and for a period no longer than 20 days (which corresponds approximately to the 95th quantile for the length of stay on the ward of the entire cohort of patients considered, \( N = 407 \)). The rationale for this was to exclude both very short or very long stayers from our analysis and focus on a more “homogeneous” cohort of patients with regard to length of stay. For the analysis considered in this work, we also excluded patients who died on the ward, had an emergency admission to the Intensive Care Unit (ICU), or cardiac arrest. This resulted in a total of 326 patients that were included in this analysis. For all patients, the first day of their vital-sign trajectories corresponds to the day on which surgery took place.
Although the proposed approach can be applied to multivariate time-series data, given the small size of the dataset we demonstrate the proposed approach using univariate observational data from our cohort of post-operative patients. We can, however, take into account the contribution of all five vital signs (and not focus only on a single vital sign). For this, we consider the output of a model constructed using our proposed approach described in Reference 18, which provides a parsimonious representation of the overall physiological trajectories for each patient. In short, a multivariate model of normality based on pre-discharge vital-sign data from patients (who were discharged alive), $U$, is constructed using kernel density estimates \cite{19,20}; then, for each patient, the likelihood $p(u_i|U,\sigma)$ of each observation set $u_i$ (with $i = 0, ..., N$) with respect to this model is computed, and the correspondent novelty score $z(u_i)$ is finally obtained: 

$$z(u_i) = -\log p(u_i|U,\theta).$$

Thus, for each patient, we obtain a “univariate”, unevenly sampled time-series of novelty score values; i.e., a collection $n$ pairs of $(t, z(u))$, where $t$ corresponds to the time of the observation set $u$, and $n$ is the number of observation sets for that patient. The details of how the model is constructed have previously been described (see Reference 18).

### 6.2.2 Gaussian processes

We provide a brief summary of Gaussian processes in this section. It therefore makes a rather compressed introduction to the topic. A more thorough introduction is available in Reference 11.
When performing a regression task we assume there exists some optimal prediction function $f \in \mathcal{X} \rightarrow \mathcal{Y}$, possibly with a noise distribution. In linear regression, we assume that the outputs $y$ are a linear function of the inputs $X$, with some parameters $\theta$, usually fewer than the number of training examples $N : |\theta| \ll N$. However, for many real-world datasets a simple parametric form, such as a linear form, is an unrealistic assumption. Therefore, we would like to have models that can learn general functions $f$. Since the functions may not be summarised by a small (fixed) number of parameters $\theta$, maximum likelihood estimation of the parameters may cause overfitting. In fact, in a Gaussian process, the effective number of parameters is often infinite. Therefore, in order to perform inference we need to place a prior probability distribution on functions. We make predictions using our posterior on an underlying predictive function $f$ given a set of training examples in the form of input-output pairs: $D = \{(x_i \in \mathbb{R}^D, y_i \in \mathbb{R})\}_{i=1}^N$.

Gaussian processes provide a distribution over real-valued functions which is widely used for non-linear regression and classification tasks [11]. By definition, a function $f : \mathcal{X} \rightarrow \mathbb{R}$ is distributed according to a Gaussian process if and only if $p(f(x_1), ..., f(x_N))$, the density of that function’s values at any $N$ points $x_i \in \mathcal{X}$, is multivariate Gaussian. This allows Gaussian processes to be parameterised tractably by a mean function $m(x)$ and a covariance kernel function $K(x_i, x_j)$ specifying the correlations within any finite point set, such that $y = f(x) \sim \mathcal{GP}(m(x), K(x_i, x_j))$, (6.1)

with possibly some Gaussian observation noise. Note that the covariance matrix $K$, or Gram matrix, whose entries $K_{ij}$ are often thought of as the “similarity” between inputs $x_i$ and $x_j$, encodes our prior knowledge concerning the functional behaviour we wish to model. Without loss of generality, the prior mean function is typically set to zero: $m(x) = 0$. The most commonly used covariance function is the squared-exponential,\footnote{It is also known as the exponentiated-quadratic, or the Gaussian kernel function.}

$$k_{SE}(x_i, x_j) = \sigma_0^2 \exp \left( -\frac{\|x_i - x_j\|^2}{2\ell^2} \right).$$

(6.2)

where $\theta = \{\sigma_0, \ell\}$ are hyperparameters modelling the $y$-scaling and $x$-scaling (or time-scale if the data are time-series), respectively, and where $\|\cdot\|$ denotes the Euclidean norm. The squared-exponential covariance function is said to be stationary because it only depends on the difference between points $x_i - x_j$, rather than on their absolute value. In general, covariance functions have to fulfil Mercer’s theorem, meaning that $K(x_i, x_j)$ has to be symmetric and positive semidefinite, and therefore $k_{SE}(\cdot, \cdot)$ is a valid kernel. Many mathematical operations, such as summation or taking a product, preserve positive definiteness and can therefore be used for combining basic kernels to make more complex kernels. A survey of covariance functions can be found in chapter 4 of Reference 11.
Given a training set \( \mathcal{D} \), using the standard conditioning rules for a Gaussian distribution, we can obtain the predictive distribution on a new observation \( y_* \) at test input \( x_* \):

\[
\begin{bmatrix}
y \\
y_*
\end{bmatrix} = \mathcal{N}
\left( 
\begin{bmatrix}
0 \\
0
\end{bmatrix},
\begin{bmatrix}
K_x & K_{ss} \\
K_{sx} & K_{ss}
\end{bmatrix}
\right)
\]

(6.3)

implying

\[
p(y_*|x_*, X, y) \sim \mathcal{N}(\mu_*, \sigma_*^2), \text{ with}
\]

\[
\mu_* = K_x^T K^{-1} y \in \mathbb{R},
\]

(6.5)

\[
\sigma_*^2 = K_{ss} - K_{sx}^T K^{-1} K_{sx} \in \mathbb{R}^+.
\]

(6.6)

Here, \( K_x = k(X, x_*) \in \mathbb{R}^{N+1} \) is the cross-covariance between the test input \( x_* \) and the training inputs \( X; K_{ss} = k(x_*, x_*) \in \mathbb{R}^+ \) is the prior variance of \( x_* \).

The values of the hyperparameters \( \theta \) may be optimised by, for example, minimising the negative log marginal likelihood (NLML) which is defined as

\[
\text{NLML} = -\log p(y|x, \theta)
\]

(6.7)

\[
= \frac{1}{2} \log |K| + \frac{1}{2} y^T K^{-1} y + \frac{N}{2} \log(2\pi)
\]

(6.8)

This is sometimes called the type-II maximum likelihood (if we remove the negative logarithm). Interpreting the NLML as a cost function reveals that the first term penalises model complexity and the second term penalises low data likelihood (i.e., low data fitness). Bias-variance trade-off is therefore performed by minimising the NLML, which is commonly achieved using gradient descent. In a full Bayesian treatment, we should integrate out the hyperparameters. Unfortunately, this cannot be performed analytically in general, e.g., for the input scale. Sampling methods, or other approximations, are usually used to estimate these integrals [11].

In our experiments, we used a single squared-exponential covariance function and a zero-mean function to capture the overall physiological recovery of post-operative patients. During training, each time-series was centred by removing the mean of the time-series data to achieve a zero-mean function. The hyperparameters \( \{\sigma_0^2, \ell\} \) were selected using a grid-search optimiser for minimising the NLML: \( \sigma_0^2 \in [3, 4, 5, ..., 15] \) (in units of \( z(x) \)) and \( \ell \in [2.0, 2.5, 3.0, ..., 5.0] \) (in units of days). We then evaluated the resulting function over a uniform grid of test points \( x_* \) sampled every hour within the range \( x_*^n \in [t_1, t_f] \), where \( t_1 \) and \( t_f \) correspond to the time of the first and last observations for patient \( n \). Figure 6.2 shows a few examples of the regression results obtained with our dataset using this procedure.

We observe that small (daily) variations of the novelty scores are smoothed by use of this approach. Nevertheless, the model is able to capture the overall trajectory of recovery of the patients. For example, patient 31 exhibits a high initial physiological derangement following major surgery, and a clear return to normality (decrease in the physiological novelty score), as a result of recovery on the ward. Patient 105, on the other hand, appears to be within the normal range of novelty score values throughout their stay on the ward.
6.2.3 Time-series clustering

In this section we describe our proposed approach for performing clustering of the Gaussian process posteriors over the uniform grid of test points (sampled every hour).

To quantify the similarity of time-series we make use of kernels. Kernel-based classifiers, like any other classification scheme, should be robust against invariances and distortions. Dynamic time warping (DTW), a method based on dynamic programming [21], has been previously combined with kernel methods [22,23].

Let $\mathcal{X}^N$ be the set of discrete-time time-series taking values in an arbitrary space $\mathcal{X}$. One can try to align two time-series $u = (u_1, ..., u_n)$ and $v = (v_1, ..., v_m)$ of lengths $n$ and $m$, respectively, in various ways by distorting them. An alignment $\pi$ of length $|\pi| = p$ between two sequences $u$ and $v$ (with $p \leq n + m - 1$ since the two series have $n + m$ points and they are matched at least at one point in time) is a pair of increasing integer vectors $(\pi_1, \pi_2)$ such that $1 \leq \pi_1(1) \leq ... \leq \pi_1(p) = n$ and $1 \leq \pi_2(1) \leq ... \leq \pi_2(p) = m$, with unitary increments and no simultaneous repetitions (we use the notation of Reference 24). We write $\mathcal{A}(u, v)$ for the set of all possible alignments between $u$ and $v$, which can be conveniently represented by paths in an $n \times m$ matrix. Following the well-known DTW metric, the cost of the alignment can be defined by means of a distance $\phi$ that measures the discrepancy between any two points $u_i$ and $v_j$, such that

$$D_{u,v}(\pi) = \sum_{i=1}^{||\pi||} \phi(u_{\pi_1(i)}, v_{\pi_2(i)})$$

Dynamic programming algorithms provide an efficient way to compute the optimal path $\pi^*$ which gives the minimum cost among all possible alignments,

$$\pi^* = \arg \min_{\pi \in \mathcal{A}(u, v)} \frac{1}{||\pi||} D_{u,v}(\pi)$$
Different kernel distances (or scores) $\phi$ have been proposed in the literature to compute the similarity between time-series based on DTW, such as the negative squared Euclidean distance $\phi(u, v) = -||u - v||^2$ [22],

$$k_{DTW_1}(u, v) = \exp \left( -\arg\min_{\pi \in \mathcal{A}(u,v)} \frac{1}{|\pi|} \sum_{i=1}^{||\pi||} ||u_{\pi_1(i)} - v_{\pi_2(i)}||^2 \right), \quad (6.11)$$
or a Gaussian kernel [23],

$$k_{DTW_2}(u, v) = \arg\max_{\pi \in \mathcal{A}(u,v)} \frac{1}{|\pi|} \sum_{i=1}^{||\pi||} \exp \left( -\frac{1}{\sigma^2} ||u_{\pi_1(i)} - v_{\pi_2(i)}||^2 \right). \quad (6.12)$$

The global alignment (GA) kernel, proposed by Cuturi et al. [25], assumes that the alignment that gives the minimum cost may be sensitive to peculiarities of the time-series and intends to take advantage of all possible alignments weighted exponentially. Hence, it is defined as the sum of exponentiated costs of the individual alignments, such that

$$k_{GA}(u, v) = \sum_{\pi \in \mathcal{A}(u,v)} \exp \left( -D_{u,v}(\pi) \right) \quad (6.13)$$

$$= \sum_{\pi \in \mathcal{A}(u,v)} \exp \left( -\sum_{i=1}^{||\pi||} \phi(u_{\pi_1(i)}, v_{\pi_2(i)}) \right) \quad (6.14)$$

$$= \sum_{\pi \in \mathcal{A}(u,v)} \prod_{i=1}^{||\pi||} k(u_{\pi_1(i)}, v_{\pi_2(i)}) \quad (6.15)$$

where $k = \exp -\phi$. It has been argued that $k_{GA}$ runs over the whole spectrum of the costs and leads to a smoother measure than the minimum of the costs, i.e., the DTW distance [25].

In our implementation, we use the kernel suggested in Reference 24,

$$k(u, v) = \exp (-\phi_{\sigma}(u,v)), \quad (6.16)$$

$$\phi_{\sigma}(u,v) = \frac{1}{2\sigma^2} d(u,v) + \log \left( 2 - e^{-\frac{1}{2\sigma^2} d(u,v)} \right) \quad (6.17)$$

where the bandwidth $\sigma$ of the kernel can be set as a multiple of a simple estimate of the median (Euclidean) distance of different points observed in different time-series of the training set, scaled by the square root of the median length of time-series in the training set; as suggested in Reference 24; $d(u,v)$ corresponds to the distance between any two points of the time-series $u$ and $v$. Cuturi et al. [25] used $d(u,v) = ||u - v||^2$.

In our case, as previously described, the time-series or trajectories obtained with the Gaussian process framework are characterised by a mean function and a measure of the uncertainty in the trajectory estimation, which handles the incompleteness, noise and artefacts underlying the observational data considered. That is, because we used

$^2$That is, $\hat{\sigma} = \text{median}(||u - v||)\sqrt{L}$, where $L$ corresponds to the median length of the time-series in $\mathcal{X}$. 


a Gaussian likelihood function, each point $u_i$ in a given trajectory $u$, is defined by

$$u_i \sim \mathcal{N}(m_i, \Sigma_i).$$

In order to take this into account, we use the 2-Wasserstein distance between two Gaussian distributions [26], which is given by

$$d(u, v) = d(\mathcal{N}(m_u, \Sigma_u), \mathcal{N}(m_v, \Sigma_v)) = ||m_u - m_v||^2 + ||\Sigma_u^{1/2} - \Sigma_v^{1/2}||_F^2 \quad (6.18)$$

where $||\cdot||_F$ is the Frobenius (also called Hilbert-Schmidt) norm.

Using the measure of discrepancy (or similarity) described above, classification or clustering of the trajectories may be performed. There are a large number of clustering methods proposed in the literature. In this work, we use an agglomerative hierarchical clustering method. Other partitioning techniques, such as $k$-means or model-based clustering, share the property that objects in a dataset are partitioned into a specific number of clusters at a single step. In contrast, hierarchical clustering methods produce a cluster tree; i.e., a series of nested clusters through a series of partitions.

Hierarchical clustering can be either agglomerative, with fewer clusters at the higher level (by fusing clusters generated at the lower level), or divisive, which separate the $n$ objects into more and finer groups in sequential steps. Agglomerative hierarchical clustering, in particular, starts with $n$ clusters, each of which contains a single object in the dataset. In the second step, the two clusters that have the closest between-cluster distance are fused and are then treated as a single cluster in the next step. As the procedure continues, it results in a single cluster containing all the $n$ objects. Agglomerative methods vary in the ways of defining the distance between two clusters when more than one object is present in either of them. For example, the single linkage method considers the shortest pair-wise distance between objects in two different clusters as the distance between the two clusters. In contrast, with the complete linkage method, the distance between two clusters is defined as the distance between the most distant pair of objects. Here, we use average linkage clustering, in which the average of the pair-wise distances between all pairs of objects coming from each of two clusters is taken as the distance between the two clusters.

The number of clusters was estimated using the gap method (which is described in Reference 27, together with a short review on methods for estimating the optimal number of clusters).

### 6.3 Results

We applied this method to the trajectories of the 326 post-operative patients in order to find different patterns of physiological recovery from major surgery. For this, the Gaussian process posteriors (over the uniform grid of test points sampled every hour) of the physiological trajectories were used. Hierarchical clustering was used to group similar trajectories based on the modified GA kernel distance described earlier.
Figure 6.3  Representation of the clusters obtained during training using our patient-to-patient similarity approach: each node of the graph corresponds to a patient, and the edges connecting any two nodes represent the similarity between them. In each sub-plot, 10 random mean trajectories from each cluster are represented.

The number of clusters obtained, determined using the gap method, was 5 functional clusters. Figure 6.3 illustrates the clusters of patients obtained. Figure 6.4 shows examples of trajectories associated with each cluster of patient trajectories, as an overall representation of the results obtained in this experiment. From the 326 patients included in the normal group, 87 (27%) were part of the cluster represented in the left part of the network represented in Figure 6.3 (first row of Figure 6.4), 79 (24%) were part of the cluster coloured with the dark colour (represented in the right part of the network in Figure 6.3, and last row of Figure 6.4), and the remaining of the patients were part of the other clusters (58 or 17% in the cluster represented in the second row of Figure 6.4, and 51 or 16% in each of the other two clusters).

6.4 Discussion

As expected, different patients may exhibit different physiological trajectories during recovery. Although all patients included in this analysis did recover from surgery and

---

The graph was obtained using the freely available software called Gephi; for that, the similarity matrix (as computed by the proposed approach) was provided to the software, and the ForceAtlas 2 algorithm was used to reorganise the layout of the graph, which takes into account the degree of similarity between the nodes and their neighbourhood.
Figure 6.4 Examples of the Gaussian process posteriors for three patients belonging to each cluster (each row represents patients from one cluster, which correspond to the clusters shown in Figure 6.3). Circles correspond to the raw data. Thick lines correspond to the posterior means, and dashed lines indicate the 95% confidence area for the posterior mean obtained. Shaded areas denote the uncertainty level of the mean (in darker areas, uncertainty is lower).
were discharged home without any major adverse event in the course of their stay in the hospital, these results suggest that the physiological trajectories (based on the novelty scores) for these patients may be different from one another, as expected. While some patients exhibit a recovery trend with a pronounced decrease in the novelty score $z(x)$ in the first couple of days after surgery and a constant $z(x)$ for the remainder of their stay (Figure 6.4, bottom row), other patients present a relatively “stable” trajectory, with only small variations of $z(x)$ throughout their stay (e.g., Figure 6.4, top row). For other patients, a certain variation of $z(x)$ is manifested in their physiological trajectories.

Using the similarity metric and clustering procedure described earlier, the set of entities that are alike appear (visually) to be assigned to the same cluster, and entities from different clusters are also clearly less alike. There are many possible explanations for the different recovery patterns observed: the type of surgery that the patient underwent, the age of the patient, how fit the patient is at the time of operation and other possible underlying conditions. No direct and clear associations between the first two factors (surgery type and age) and the clusters of data were found, which may be due to the small number of patients included in this study and the variety of procedures that patients underwent, although there is room for exploring additional factors.

A few points should be made here regarding the analysis conducted and the results obtained. In the first place, we observe that small (daily) variations of the novelty scores were smoothed out in this analysis. The main goal of this approach was to capture the overall trajectory of recovery of post-operative patients, which motivated the selection of the covariance function (and hyperparameters priors) that was used to model the data with Gaussian processes. In order to also capture short-term variations, one could use a more complex covariance function derived by combining simple covariance functions. For example, the addition of two squared-exponential covariance functions, one to model the short-term variations in the novelty score, and one to model the long-term trends, could be used to provide a better fit to the data. Nevertheless, some additional work would be required to select the set of priors for each hyperparameter. A fully Bayesian approach would be advantageous in this case to better encode the level of uncertainty in the hyperparameters; i.e., rather than using an expensive grid-search optimisation procedure over all possible values for each hyperparameter, prior distributions could be set for each of the hyperparameters, which would be integrated out to obtain the Gaussian process posterior mean and variance.

It is also important to mention that, although we focused on the analysis using trajectories of novelty scores (resulting in “univariate” time-series data streams), the same approach could be used for multivariate time-series; for example, by considering all the vital signs, rather than the novelty score that combines them into a single score. As described earlier, the GA kernel distance is able to cope with multivariate time-series data. Nevertheless, the visualisation of the results for evaluating the performance of the method would be more challenging than that for the univariate case. Moreover, due to the increase of degrees of freedom in the multivariate case, a larger sample of data would be needed to derive a more representative set of trajectories for clustering.
We observe that the proposed approach may be used to recognise “normal” or previously observed physiological patterns and identify abnormal or “novel” physiological trajectories. For example, one may determine the distance (or similarity) between each test Gaussian posterior trajectory and the training Gaussian process posterior trajectories. According to this distance, the test trajectory may be either assigned to one of the five clusters of trajectories or classified as a “novel” trajectory (that is, it is substantially different from the trajectories computed during training). A similar approach has been described in Reference 28.

Finally, we also note that the comparison of our method with other approaches proposed in the literature (such as those proposed in References 9, 8, 6) may be difficult due to the characteristics of our observational dataset. The work described in this chapter may be more advantageous and provide promising results, as the described method includes a direct quantification of the uncertainty in the trajectory estimation (provided by the Gaussian process model), handling incompleteness, noise and artefact in a robust manner.

6.5 Conclusion

We have described a method by which unevenly sampled time-series data may be analysed to better understand the overall recovery trajectories of post-operative patients. Using a similarity metric, which is based on the concepts of DTW and GA kernel, and a hierarchical clustering method, different groups of physiological behaviours of recovery from surgery were revealed. The majority of patients were found to belong to one of two functional clusters: one group of patients who exhibited a recovery trend with a pronounced decrease in the novelty score in the first couple of days after surgery and a constant score for the remainder of their stay on the ward; and a group of patients who presented a relatively “stable” trajectory, with only small variations of the novelty score throughout their stay post-operatively.

The proposed approach may provide a new tool for studying and better understanding the recovery phase of patients post-operatively, which is known to be heterogeneous. As electronic medical records continue to collect data from other interventions (e.g., elective surgery), there will be a growing need for such tools based on machine learning to refine the characterisation of what constitutes a “normal” and an “abnormal” recovery from a major intervention, and quantify the effects of variability in treatment protocols across individuals in these groups.

Acknowledgements

Marco A. F. Pimentel was supported by a Health Innovation Challenge Fund by the Wellcome Trust and the Department of Health. David A. Clifton was supported by a Royal Academy of Engineering Research Fellowship; Balliol College, Oxford; and the Centre of Excellence in Personalised Healthcare funded by the Wellcome Trust and EPSRC [Grant WT 088877/Z/09/Z].
References


