

Received February 18, 2019, accepted March 25, 2019, date of publication April 24, 2019, date of current version May 15, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2912079

Patient-Specific Physiological Monitoring and Prediction Using Structured Gaussian Processes

TINGTING ZHU¹, GLEN WRIGHT COLOPY¹, CLARE MACEWEN¹, KATHERINE NIEHAUS¹,
YANG YANG¹, CHRIS W. PUGH², AND DAVID A. CLIFTON¹

¹Department of Engineering Science, Institute of Biomedical Engineering, University of Oxford, Oxford, U.K.

²Nuffield Department of Medicine, University of Oxford, Oxford, U.K.

Corresponding author: Tingting Zhu (tingting.zhu@eng.ox.ac.uk)

This work was supported in part by the NIHR and in part by the EPSRC. The work of T. Zhu was supported by the St. Hilda's College, Oxford.

ABSTRACT The management of patient well-being can be performed by monitoring continuous time-series vital-sign data via low-cost wearable devices. Automated algorithms may then be used with the resulting data to provide early warning of deterioration of the health of an individual. Such algorithms are typically trained for a large population without considering the time-variability and inter-subject variability of the data being collected. In the case where limited numbers of subjects are available, it is difficult to create a generalized population model from a small sample size. Furthermore, some “normal” patients may exhibit different physiological patterns when compared to other “normal” patients, forming multiple “normal” clusters/subgroups. This also makes inferring a population model difficult. It is, therefore, preferable to develop patient/subgroup-specific time-series models to overcome these challenges. We propose using Bayesian hierarchical Gaussian processes to infer the hidden latent structure of the vital sign's trajectory for each individual patient or group of patients who share similar patterns. We further demonstrate the feasibility of such a model in novelty detection, using the symmetric Kullback–Leibler divergence. This allows us to identify any patterns that correspond to “normal” or “abnormal” physiology, and further classifying “abnormal” patterns from a model of “normal” latent trajectories. We tested our approach using two real datasets for different monitoring scenarios. Our model was compared to the performance of the state-of-the-art unsupervised clustering algorithms, demonstrating at least 10% improvement in accuracy. We further benchmarked against two one-class classifiers and showed at least 5% accuracy improvement when using the proposed metrics in identifying abnormal physiological episodes.

INDEX TERMS Physiology, patient monitoring, pattern analysis, Bayes methods.

I. INTRODUCTION

Time-series data such as vital-sign measurements are commonly used in hospital, as they provide a direct indication of a patient's physiological state and can be easily interpreted by clinicians. Abnormal vital-sign values are thereby often used as indicators of physiological instability. There are many approaches described in the literature to classify the health of a patient according to their vital signs; many such approaches are either heuristic [1] or assume the data are time-invariant, patient-invariant, independent and identically distributed [2], [3]. Modeling the time-series vital-sign data for a patient has been described using linear dynamic systems [4], [5] as well as hidden Markov autoregressive models [6], [7]. How-

ever, such approaches cannot easily cope with the fact that vital-sign recordings commonly contain different numbers of observations, and the times at which observations are made may not be aligned (i.e., the data may be unevenly-sampled and corrupted due to artifacts from sensors). To address these problems, Gaussian processes (GP) are used for modeling physiological time-series data: Stegle *et al.* [8] considered GP regression to model noisy heart-rate data, using the cluster assignments of noise levels as a means of classifying the state of health of the patient. Clifton *et al.* [9] used GP regression to cope with artifactual and missing vital-sign data, which incorporated a novelty score to identify abnormality. The previous works assumed the same kernel function was used for all patients under a population, which maybe not robust in real scenarios. It is unclear how to optimise a patient-specific GP for an individual when there are multiple, co-

The associate editor coordinating the review of this manuscript and approving it for publication was Qingxue Zhang.

existing, similar treatments, and whether there are clusters of patients that exhibit similar time-series patterns while others do not.

In the literature concerning clustering of health-related time-series data to identify patterns of normal and abnormal behavior, the recordings tend to be grouped using traditional approaches such as K-means, mixtures of Gaussians, and hierarchical clustering [10]. Pimentel *et al.* [11] considered a direct clustering of GPs using a likelihood similarity index. However, they only used the mean trajectories (i.e., the GP latent functions) to compute the similarity between data, excluding GP uncertainty in the mean function over the temporal domain. Kim and Lee [12] proposed a clustering algorithm using only the GP-derived variance function to construct cluster boundaries. Dirichlet process (DP) models have been a popular choice for constructing mixtures of Gaussian processes with health-related data: Ross and Dy [13] clustered subjects based on their similarity in response to environmental and disease factors. Xu *et al.* [14] grouped baseline deviation of subjects based on their similarity in the parameters of the GP. In an alternative approach, a hierarchical GP was proposed by Hensman *et al.* [15], [16] to model structured time-series data, and clustering was then performed to describe different behavior of gene expression data. Similar work was also proposed by Park and Choi [17] but using a different objective optimization for the GP regression. It is unclear how such aforementioned approaches translate to time-series vital-sign data that are monitored continuously and change rapidly depending on the health status of the patient. Noting that there are two challenges in vital-sign modeling: (i) the change of physiology can be drastic and sensitive to artifacts regardless of the health status of an individual, and (ii) the same patient may exhibit different patterns of the same time-series vital-sign data under the same health condition. Furthermore, previous works have focused solely on the method development, with no known work utilizing hierarchical GP to perform classification of patient-specific trajectory for being normal or abnormal.

In this work, we propose using a hierarchical structure with Bayesian treatment to infer the latent trajectory from similar sessions/treatments, where each subject (e.g., each individual patient) may have both measurements recorded over multiple time sessions and multiple types of measurements for each session. For example, a patient may have a time-series of one recurring stochastic behavior, and then another time-series with a different stochastic behavior. Thus we introduce the notion of a hierarchy of a set of time-series for a patient, which will be formalized later. We then demonstrate the use of these latent trajectories to build a model of normality for classification purpose. We use the symmetric Kullback-Leibler divergence as a similarity measure between the model and GP-fitted data from an individual patient to classify any patient instability that may exist. We demonstrate the feasibility of the proposed model which is robust to small datasets where a few number

of patients are available with longitudinal time-series that last for minutes to hours, a common scenario in physiological monitoring.

II. METHODS

A. GAUSSIAN PROCESS REGRESSION

Given a dataset \mathbf{D} of M physiological values \mathbf{y} over time intervals \mathbf{x} , $\mathbf{D} = [\mathbf{x}_i, \mathbf{y}_i]_{i=1}^M$, we assume that vital signs \mathbf{y} can be considered as related to an underlying function $f(\mathbf{x})$ through a Gaussian noise model:

$$\mathbf{y} = f(\mathbf{x}) + \epsilon, \quad (1)$$

where $f(\mathbf{x})$ is a latent function of \mathbf{x} , and $\epsilon \sim N(0, \sigma_y^2)$ is defined as the additive Gaussian noise with variance σ_y^2 over the latent function f . We further assume $f(\mathbf{x})$ has a Gaussian Process prior [18]:

$$f(\mathbf{x}) \sim GP(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')), \quad (2)$$

where $m(\mathbf{x})$ is the mean function of the GP, and $k(\mathbf{x}, \mathbf{x}')$ is a covariance function which describes the relationship among the \mathbf{y} values that is determined according to the distance between the \mathbf{x} values. Therefore,

$$\mathbf{y} \sim GP(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}') + \sigma_y^2 \delta(\mathbf{x}, \mathbf{x}')), \quad (3)$$

where δ is the Kronecker delta function with $\delta(\mathbf{x}, \mathbf{x}') = \mathbb{I}(x_i = x'_i)$. Covariance functions may be incorporated in an additive manner in order to model the short- and long-term variability in a time-series trajectory, such as:

$$\begin{aligned} k(\mathbf{x}, \mathbf{x}') + \sigma_y^2 \delta(\mathbf{x}, \mathbf{x}') = & \underbrace{\sigma_1^2 \left(1 + \frac{\sqrt{5}r}{\lambda_1} + \frac{5r^2}{3\lambda_1^2} \right) \exp\left(-\frac{\sqrt{5}r}{\lambda_1}\right)}_{\text{Matern}_{\frac{5}{2}}(\sigma_1, \lambda_1)} \\ & + \underbrace{\sigma_2^2 \left(1 + \frac{\sqrt{3}r}{\lambda_2} \right) \exp\left(-\frac{\sqrt{3}r}{\lambda_2}\right)}_{\text{Matern}_{\frac{3}{2}}(\sigma_2, \lambda_2)} \\ & + \underbrace{\sigma_3^2 \exp\left(-\frac{r^2}{2\lambda_3^2}\right)}_{\text{RBF}(\sigma_3, \lambda_3)} + \underbrace{\sigma_y^2 \mathbb{I}}_{\text{noise}(\sigma_y)}, \end{aligned} \quad (4)$$

where $r = \sqrt{\sum_{i=1}^n (x_i - x'_i)^2}$. Here, we have that λ_1 and σ_1^2 , λ_2 , and σ_2^2 , λ_3 , and σ_3^2 are the length-scale and variance hyperparameters for the Matern $_{\frac{5}{2}}$, Matern $_{\frac{3}{2}}$, and radial basis function (RBF) covariance functions respectively. The likelihood of the observed data can be estimated by marginalizing over the latent function \mathbf{f} :

$$p(\mathbf{y}|\mathbf{x}) = \int_{\mathbf{f}} p(\mathbf{y}|\mathbf{f}, \mathbf{x}) p(\mathbf{f}|\mathbf{x}) d\mathbf{f}. \quad (5)$$

To estimate the values of hyperparameters θ in $k(\mathbf{x}, \mathbf{x}')$, an empirical Bayes approach (i.e., type-II maximum-a-posteriori) may be considered by finding the derivatives of the log marginal likelihood (LML) with respect to θ and maximizing

$$\log [p(\mathbf{y}|\mathbf{x}, \theta)] = -\frac{1}{2}\mathbf{y}^T [\mathbf{K}_{\mathbf{x}, \mathbf{x}} + \sigma_y^2 \mathbb{I}]^{-1} \mathbf{y} + p(\theta) - \frac{1}{2} \log |\mathbf{K}_{\mathbf{x}, \mathbf{x}} + \sigma_y^2 \mathbb{I}| - \frac{M}{2} \log (2\pi). \quad (6)$$

where $\mathbf{K}_{\mathbf{x}, \mathbf{x}}$ denotes the covariance matrix, and its $(i, j)^{th}$ element can be estimated from the covariance function $k(\mathbf{x}[i], \mathbf{x}[j])$, and where $p(\theta)$ describes the probability distribution over θ .

Suppose that we have observed \mathbf{y} (i.e., a noisy version of hidden f) at times \mathbf{x} , and wish to predict the function outputs \mathbf{f}^* at times \mathbf{x}^* , the joint probability of \mathbf{y} and \mathbf{f}^* that is conditional on \mathbf{x} and \mathbf{x}^* can be defined as [18]:

$$p\left(\begin{bmatrix} \mathbf{y} \\ \mathbf{f}^* \end{bmatrix} \middle| \begin{bmatrix} \mathbf{x} \\ \mathbf{x}^* \end{bmatrix}\right) = N\left(\begin{bmatrix} \mathbf{y} \\ \mathbf{f}^* \end{bmatrix} \middle| \begin{bmatrix} \boldsymbol{\mu} \\ \boldsymbol{\mu}^* \end{bmatrix}, \begin{bmatrix} \mathbf{K}_{\mathbf{x}, \mathbf{x}} + \sigma_y^2 \mathbb{I} & \mathbf{K}_{\mathbf{x}, \mathbf{x}^*} \\ \mathbf{K}_{\mathbf{x}, \mathbf{x}^*}^T & \mathbf{K}_{\mathbf{x}^*, \mathbf{x}^*} \end{bmatrix}\right). \quad (7)$$

Conditioned on the observed data \mathbf{D} , the posterior predictive density of \mathbf{f}^* is:

$$p(\mathbf{f}^* | \mathbf{x}^*, \mathbf{x}, \mathbf{y}) = N(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*), \quad (8)$$

$$\boldsymbol{\mu}^* = \mathbf{K}_{\mathbf{x}, \mathbf{x}^*}^T [\mathbf{K}_{\mathbf{x}, \mathbf{x}} + \sigma_y^2 \mathbb{I}]^{-1} \mathbf{y}$$

$$\boldsymbol{\Sigma}^* = \mathbf{K}_{\mathbf{x}^*, \mathbf{x}^*} - \mathbf{K}_{\mathbf{x}, \mathbf{x}^*}^T [\mathbf{K}_{\mathbf{x}, \mathbf{x}} + \sigma_y^2 \mathbb{I}]^{-1} \mathbf{K}_{\mathbf{x}, \mathbf{x}^*}.$$

B. STRUCTURED GAUSSIAN PROCESSES

Here we assume the n^{th} subject is represented by R_n time-series (such as R_n physiological measurements over time), and R_n can be defined as $\mathbf{Y}_n = \{\mathbf{y}_{nr}\}_{r=1}^{R_n}$ taken at times $\mathbf{T}_n = \{\mathbf{x}_{nr}\}_{r=1}^{R_n}$. It is further assumed that there is a latent GP function which governs these R_n time-series for the n^{th} subject, denoted as $g_n(\mathbf{x})$. With the aforementioned assumptions, each $r = 1, \dots, R_n$ time-series of data for this n^{th} subject can be considered as drawn from a GP as [19]:

$$f_{nr}(\mathbf{x}) \sim GP(g_n(\mathbf{x}), k_f(\mathbf{x}, \mathbf{x}')). \quad (9)$$

The structure of hierarchical GPs (denoted thereafter as HGP) can therefore be formulated for the n^{th} subject as [19]:

$$g_n(\mathbf{x}) \sim GP(0, k_g(\mathbf{x}, \mathbf{x}')), \quad [\text{parent structure}]$$

$$f_{nr}(\mathbf{x}) \sim GP(g_n(\mathbf{x}), k_f(\mathbf{x}, \mathbf{x}')). \quad [\text{child structure}] \quad (10)$$

That is, we assume that all $r = 1, \dots, R_n$ GPs for patient n share some common mean function, which is itself modeled by a GP g_n . Thus the difference for GPs may be thought of as modeling the residuals around the (shared) mean function given by g_n . A graphical representation of the HGP model is shown in Figure 1 and its likelihood can be described as:

$$p(\mathbf{Y}_n | \mathbf{T}_n, \theta) = N(\mathbf{Y}_n | 0, \boldsymbol{\Sigma}_n),$$

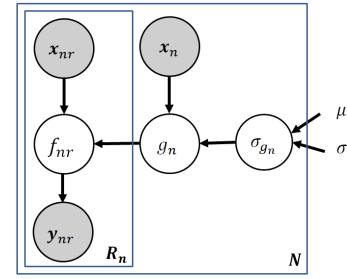


FIGURE 1. Graphical representation of the proposed model [19]: y_{nr} corresponds to the r^{th} time-series for the n^{th} subject, and is modeled by the hidden node f_{nr} given the observed timestamps x_{nr} . The model further assumes there is a latent hidden GP, g_n , which governs the relationship among $\{f_{nr}\}_{r=1}^{R_n}$ over timestamps x_n . A log-normal prior was placed over the variance hyperparameter (i.e., σ_{g_n}) of g_n .

$$\boldsymbol{\Sigma}_n[r, r'] = \begin{cases} \mathbf{K}_g(\mathbf{x}_{nr}, \mathbf{x}_{nr'}) + \mathbf{K}_f(\mathbf{x}_{nr}, \mathbf{x}_{nr'}) + \sigma_y^2 \mathbb{I} & \text{if } r = r' \\ \mathbf{K}_g(\mathbf{x}_{nr}, \mathbf{x}_{nr'}) & \text{otherwise.} \end{cases} \quad (11)$$

where $\boldsymbol{\Sigma}_n[r, r']$ is a block matrix with $\mathbf{K}_g(\mathbf{x}_{nr}, \mathbf{x}_{nr'})$ describing the off-diagonal blocks as the covariance matrices for the latent “parent” structure. Note that the $(i, j)^{th}$ element of $\mathbf{K}_g(\mathbf{x}_{nr}, \mathbf{x}_{nr'})$ is estimated from the covariance function $k_g(\mathbf{x}_{nr}, \mathbf{x}_{nr'})$ from the r^{th} time-series for the n^{th} subject. In a similar manner, $\mathbf{K}_g(\mathbf{x}_{nr}, \mathbf{x}_{nr'}) + \mathbf{K}_f(\mathbf{x}_{nr}, \mathbf{x}_{nr'})$ are the diagonal blocks describing the covariance sub-matrices for the separate GP structure estimated from the covariance function $k_g(\mathbf{x}, \mathbf{x}') + k_f(\mathbf{x}, \mathbf{x}')$. Here we propose using the following covariance functions for our HGPs:

$$k_g(\mathbf{x}_{nr}, \mathbf{x}_{nr'}) = \text{Matern}_{\frac{5}{2}},$$

$$k_f(\mathbf{x}_{nr}, \mathbf{x}_{nr'}) = \text{Matern}_{\frac{3}{2}} + \text{RBF} + \text{noise}. \quad (12)$$

The parent structure uses a Matern kernel to capture short-term variability in a time-series, while the child structure is a variant/mutation of the parent structure. The child structure is therefore composed of RBF and Matern kernels to describe both long- and short-term changes of a time-series (such as the trajectory of heart rates over longer periods of time vs. the inter-beat variability of heart rate, respectively). The noise term can be considered as from unaccounted physiological variability or quantization of measurements made by the sensor/device [20]. Note that the description for each covariance function is detailed in Equation (4). Following the standard method described in Equation (6), we optimize the values of the hyperparameters of the covariance functions. Furthermore, inference about the functions for the r^{th} time-series of subject n can be performed in a similar manner following Equation (8):

$$p(\mathbf{f}_{nr}^* | \mathbf{x}_{nr}^*, \mathbf{x}_{nr}, \mathbf{y}_{nr}) = N(\boldsymbol{\mu}_{nr}^*, \boldsymbol{\Sigma}_{nr}^*), \quad (13)$$

$$\boldsymbol{\mu}_{nr}^* = \mathbf{K}_{gf}^T(\mathbf{x}_{nr}, \mathbf{x}_{nr}^*) [\mathbf{K}_{gf}(\mathbf{x}_{nr}, \mathbf{x}_{nr}) + \sigma_y^2 \mathbb{I}]^{-1} \mathbf{y}_{nr}$$

$$\boldsymbol{\Sigma}_{nr}^* = \mathbf{K}_{gf}^T(\mathbf{x}_{nr}^*, \mathbf{x}_{nr}^*) - \mathbf{K}_{gf}^T(\mathbf{x}_{nr}, \mathbf{x}_{nr}^*)$$

$$\times \left[\mathbf{K}_{gf}(\mathbf{x}_{nr}, \mathbf{x}_{nr}) + \sigma_y^2 \mathbb{I} \right]^{-1} \mathbf{K}_{gf}(\mathbf{x}_{nr}, \mathbf{x}_{nr}^*),$$

where $\mathbf{K}_{gf} = \mathbf{K}_g + \mathbf{K}_f$.

C. SIMILARITY METRIC

The Kullback-Leibler (KL) divergence is commonly used as a measure of the non-symmetric difference between two continuous probability distributions P_1 and P_2 :

$$D_{\widehat{KL}}(P_1 \| P_2) = \int_{-\infty}^{\infty} p_1(x) \log \frac{p_1(x)}{p_2(x)} dx, \quad (14)$$

where p_1 and p_2 are the densities of P_1 and P_2 , respectively. When both P_1 and P_2 are two multivariate normal with dimension d , their KL divergence can be estimated as:

$$\begin{aligned} D_{\widehat{KL}}(P_1 \| P_2) = & \frac{1}{2} \left[\log \frac{|\Sigma_{P_2}|}{|\Sigma_{P_1}|} - d + \text{Tr} \left(\Sigma_{P_2}^{-1} \Sigma_{P_1} \right) \right] \\ & + \frac{1}{2} \left[(\mu_{P_2} - \mu_{P_1})^T \Sigma_{P_2}^{-1} (\mu_{P_2} - \mu_{P_1}) \right], \end{aligned} \quad (15)$$

where μ_{P_2} and Σ_{P_2} , μ_{P_1} and Σ_{P_1} are the mean and covariance of P_1 and P_2 , respectively. To address the asymmetric property of KL divergence, we propose the following:

$$D_{KL} = D_{\widehat{KL}}(P_1 \| P_2) + D_{\widehat{KL}}(P_2 \| P_1). \quad (16)$$

The above symmetric KL divergence will be used later in the paper as a metric for classifying abnormal GPs from a model describing “normality”.

III. DATA DESCRIPTION

In this section, we describe two real datasets which were collected in outpatient and inpatient hospital scenarios, namely the dialysis wards and the step-down unit after surgery. Vital-sign time-series data such as systolic blood pressure and heart rate are used as exemplars in this work: the former was collected non-invasively from each patient undergoing repeated dialysis treatments – the HGPs model can be used to collapse the repeated treatments of the same patient to infer the latent trajectory representing that individual. The latter was collected from non-invasive sensors for multiple patients, where similar patients are pre-clustered into subgroups of individuals who shared similar heart rate time-series – the HGPs model can then be used to infer the latent trajectory of each sub-cluster or subgroup of patients.

A. SYSTOLIC BLOOD PRESSURE IN THE DIALYSIS SETTING

1) DATA DESCRIPTION

35 subjects (26 males and 9 females) undergoing routine haemodialysis (HD) treatment session with age of 60 (51-76.5) years for median and inter-quartile range, were recruited for this study [21]. A Finometer[®] device was used to collect minutely continuous blood pressure from each subject during treatment. Each subject has up to three sessions per week and each visit lasts up to 240 minutes. There were a total of between three and 27 HD sessions for a subject

depending on the number of weeks that they were monitored. During a HD treatment, some subjects may experience one or more intradialytic hypotension (IDH) events. IDH is defined as a sudden drop of blood pressure which leads to increased rates of morbidity and mortality both during and following the dialysis session. Hence there is a urgent need to provide identification of patients exhibiting patterns of IDH either during or between consecutive dialysis sessions, and allow for intervention (such as changing the filtration rate of a device) as early as possible. Such an algorithm could be implemented in real time as vital sign data is acquired by comparing the current vital sign trajectories to reference values. In this case, the reference values are derived from GPs fit to previous time series. An abnormal session is defined as having one or more intradialytic hypotension (IDH) events where the mean arterial blood pressure (MAP) was less than 60 mmHg and SBP less than 80% of the baseline SBP (i.e., the SBP measurement taken prior to treatment); Alternatively, when there is no baseline SBP, a MAP of less than 60 mmHg during treatment also constituted abnormality. A total of 190 abnormal and 156 normal sessions were obtained in this study: the normal cases were distributed between 26 subjects; the abnormal cases were distributed between a different set of 26 subjects (noting that not all subjects have both normal (i.e., no IDH) and abnormal (i.e., with IDH) sessions).

2) PRE-PROCESSING

The minutely-sampled SBP data for each session of an individual subject was initially fitted with a univariate GP for the process of artifact removal. To accommodate both long-term and transient changes, the following covariance function was used:

$$k(\mathbf{x}, \mathbf{x}') + \sigma_y^2 \delta(\mathbf{x}, \mathbf{x}') = \text{Matern}_{\frac{5}{2}} + \text{RBF}_s + \text{RBF}_l + \text{noise}, \quad (17)$$

where RBF_s and RBF_l refer to short and long length-scale variability in the RBF covariance function, respectively, and their length-scales were fixed to be $\lambda = 5$ and $\lambda = 200$ respectively. The length-scale of the $\text{Matern}_{\frac{5}{2}}$ was fixed to be $\lambda = 15$. Furthermore, a log-uniform prior was applied over each variance hyperparameter in the covariance functions. We note that the lengthscales are fixed to define the general trend of a given session's time-series, but the variance as a scaling factor is allowed to change, as this determines variation of function values from their mean. Each session's data were log-transformed and the mean was subtracted prior to fitting a GP. A moving window centered at each data point (± 5 minutes) was used to compute the normalized mean LML for that data point with respect to the GP. The normalization of the mean was performed using the standardized zero mean and unit variance approach. A threshold τ on LML was used to determine outliers, and any data points with LML below τ were deemed to be artifacts and were removed. The process was repeated in an iterative man-

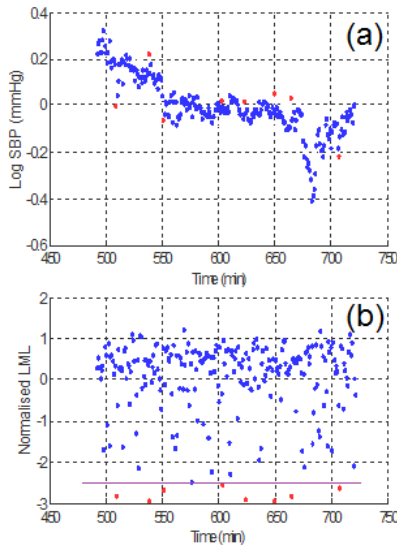


FIGURE 2. Artifact removal using a univariate GP: the GP was fitted to a log-SBP time-series of a dialysis session with 95% confidence interval (horizontal grey line) as shown in (a); the normalized LML was computed for each data point as shown in (b). A threshold τ of normalized LML (i.e., the purple horizontal bar) is also shown in (b). The time was converted from hour-of-day to minutes. The blue-colored points in both plots indicate the data that remained after artifact removal, while the artifactual dots are highlighted in red.

ner until 95% percent of data remained. Figure 2 demonstrates a fitted GP for a session of SBP data in (a), and the mean LML of each data point was computed in (b). We allowed the threshold values to vary in the range of $\tau = [-2.5, -1]$ for all subjects. After artifact removal, we fitted a univariate GP for each session using log-uniform priors over all hyperparameters.

B. HEART RATE MONITORING IN THE HOSPITAL SETTING

1) DATA DESCRIPTION

Heart rate (HR) vital-sign measurements were extracted from 336 patients in a step-down unit at the University of Pittsburgh Medical Centre (UPMC) [22]. 112 clinical emergency events were identified by clinicians for 59 patients. These emergency events were defined to be any single period, at least several minutes in length, in which measurements from any vital-sign channel (HR, RR, SaO₂, BP) were “abnormally” high or low (using clinical definitions of abnormality [22]). For the purpose of this study, only the first emergency event for each of the 59 patients was considered, to avoid the confounding effect that may arise due to clinical intervention subsequent to an emergency event. We note that it could be the case that some of these events arose due to abnormality in a vital sign other than HR. However, we hypothesize that some of these non-HR events contain abnormal HR dynamics and thus our (HR-based) modeling will consider all 59 “first events” independent of which vital sign was the primary cause of the event.

We wish to compare the trajectories of data from these patients with abnormal physiology to the trajectories of data

from normal patients. Ostensibly, patient vital-sign dynamics are at their most “normal” in the period immediately before discharge. Data from normal patients were used if there were at least four hours of data before discharge; data from abnormal patients were used if there were at least four hours of data, centered around the annotated clinical event. These requirements were met by 170 normal patients and 47 abnormal patients.

2) PRE-PROCESSING

Each HR time-series was cleaned of transient artifacts using an artifact-detection algorithm that has been previously validated on the UPMC data set [23]. Unlike the GPR-based artifact detection method described earlier, the artifact detection algorithm in [23] assumes data were independent-and-identically-distributed (iid) with respect to a short time window, $\tau = 5$ minutes, hence a unique 5-minute window was placed at each HR data point. A Gamma distribution was fitted to all HR measurements within the window, and the log-likelihood of each HR measurements was evaluated with respect to the fitted Gamma distribution. A measurement’s artifact score was calculated as the average log-likelihood, across each 5-minute window containing that measurement. Artifacts with extremely low values are considered to be far away from other measurements close in time. These measurements were removed as they were likely to be artifacts.

3) CLUSTERING SUB-POPULATIONS

In order to test the use of the HGPs framework for inferring latent structure from a population of subjects, we identified clusters of sub-populations within the set of normal patients. As the number of clusters is unknown, we applied the mixture-of-Gaussians framework proposed by Hensman *et al.* [15] to estimate the number of clusters using a truncated Dirichlet process prior. This procedure offers the advantage of finding the likely number of clusters in the data via the posterior Dirichlet distribution over the number of clusters. A total of 37 clusters were created from 170 normal subjects, and we then further removed clusters that combined a single HR time-series; this resulted in 32 groups from 165 subjects with between two and 14 subjects in each cluster. The resulting clusters are then used for subsequent analysis as described below.

IV. DATA ANALYSIS AND METHOD OF COMPARISON

A. NOVELTY DETECTION VIA ONE-CLASS LEARNING

Prior to creating a one-class model of normality, HGPs was used to infer the representative latent trajectories for describing the “normal physiology”. For the SBP dataset, this process was applied to each subject with multiple sessions, resulting 26 normal latent subject trajectories from a total of 156 sessions. For the HR dataset, 32 clusters were created among normal subjects using the Mixture of Gaussians with variational inference [15] and HGPs was used to derive the latent trajectory to represent each cluster. A KL-divergence

model of normality can then be created to identify abnormal sessions for the SBP data set and abnormal subjects for the HR dataset, respectively.

1) MODEL TRAINING

Due to limited number of data available, We considered the leave-one-out method to train our model using N latent trajectories [19]: At each fold, pairwise KL values were estimated (i.e., the latent trajectory of a subject or a cluster was compared to other latent trajectories), and resulted a N by N matrix of KL values. A threshold was then used to identify a trajectory being abnormal. To find the optimal threshold, we estimated the upper bound of the normal KL values by calculating the maximum normal KL value from each subject in the training set (denoted as D_{KL_m} hereafter), as it indicated the most-dissimilar behavior within the normal population. From these D_{KL_m} values, the mean and standard deviation D_{KL_m} were derived to define the distribution of D_{KL_m} . Various KL thresholds of “normality” can then be defined (such as a range of ± 3 standard deviation from the mean). Regarding the test set, we considered the normal sessions or subjects from the left-out subject or cluster, and all abnormal sessions or subjects, depending on the datasets. For a fair comparison between normal and abnormal classes in the test set, we drew the same number of sessions or subjects from the abnormal population (i.e., 190 sessions for SBP or 33 subjects for HR) as those from the left-out subject of cluster. To ensure the abnormal population was sampled adequately, we repeated our draw randomly 100 times using the sampling-without-replacement method.

2) STATISTICAL MEASURES

The classification performance of each fold was assessed according to its specificity, sensitivity and accuracy. These statistical measures were estimated from the ratios between the number of true positive (denoted as TP - the number of sessions correctly identified as abnormal), true negative (denoted as TN - the number of trajectories correctly identified as normal), false positive (denoted as FP - the number of trajectories incorrectly identified as abnormal) and false negative (denoted as FN - the number of trajectories incorrectly identified as normal). Out of 100 draws, the median sensitivity ($\frac{TP}{TP+FN}$) was calculated for each fold, and the mean value across N folds was then estimated for sensitivity, specificity ($\frac{TN}{TN+FP}$), and accuracy ($\frac{TP+TN}{TP+TN+FP}$). The Receiver-Operating-Characteristic (ROC) curve was also computed using the mean true positive rate (denoted as TPR as equivalent to sensitivity), and mean false positive rate (denoted as FPR as 1 - specificity) for different D_{KL_m} thresholds of normality.

B. METHOD OF COMPARISON

We began with exploration of five unsupervised clustering methods, with the goal to identify the structural difference between normal and abnormal sessions. These chosen methods are: (1) Spectral Clustering (SClust) [24];

(2) k-means clustering (KMean); (3) Hierarchical clustering with Euclidean distance (HClust); (4) Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) [25]; (5) Mixture of Gaussians (MOG). To further compare the robustness of our proposed one-class classification algorithm, we had compared its performance with the following novelty-detection methods for high-dimensional dataset: (1) One-class SVM with RBF kernel [26]; (2) Isolation Forest (iForest) [27]. Both methods were trained using five-fold cross-validation by permuting the normal sessions. We assigned 20% of data for each fold of the cross-validation as the hold-out test set. To obtain a test set with balanced normal and abnormal classes, we drew the same number of sessions from the abnormal population as those from the hold-out set, following the same method as described previously in our proposed algorithm.

V. RESULTS AND DISCUSSION

In this section, we describe the results of our proposed methods for inferring latent SBP time-series of an individual in dialysis, as well as inferring latent HR time-series of clusters of patients in the step-down ward. We further discuss the potential of our approach in novelty detection, where a model of normality is formulated using the latent trajectories to identify the abnormal trajectories. We also demonstrate the robustness of our approach against the benchmarking methods in clustering and one-class classification.

A. MODEL PARAMETERS

A log-normal prior (i.e., $LN(1, 0.25)$) was used to regularize the variance components, σ_{gn} , in the covariance function of the latent structure. Furthermore, the noise variance in the individual structure as well as that of the HGPs model were constrained with bound of $[1e-3, 1]$ to prevent over-fitting. The parameters of HGPs were optimized using the Python GPy package [28] with ten random restarts.

B. SBP DATASET

1) INFERRING LATENT TRAJECTORIES USING HGPs

The HGPs model was fitting to each subject individually, and the subject-specific latent trajectory was learnt from normal and abnormal sessions separately, resulting 26 normal and 26 abnormal latent trajectories. An example of the normal and abnormal latent trajectories of a subject are shown in Figure 3. In each row, the inferred latent structure of the SBP time-series is plotted in the leftmost panel with 95% confidence interval, and each subsequent panel is the individual session of SBP time-series with 95% confidence interval. It is important to note that HGPs can be applied to timeseries where each session might have different length and contain information at different time-stamps. Furthermore, in comparison to the univariate GP model, HGPs not only can infer a latent structure from multiple correlated sessions, but also provide a better fitting to each individual session, via learning from both the parent and child structures. Examining the latent

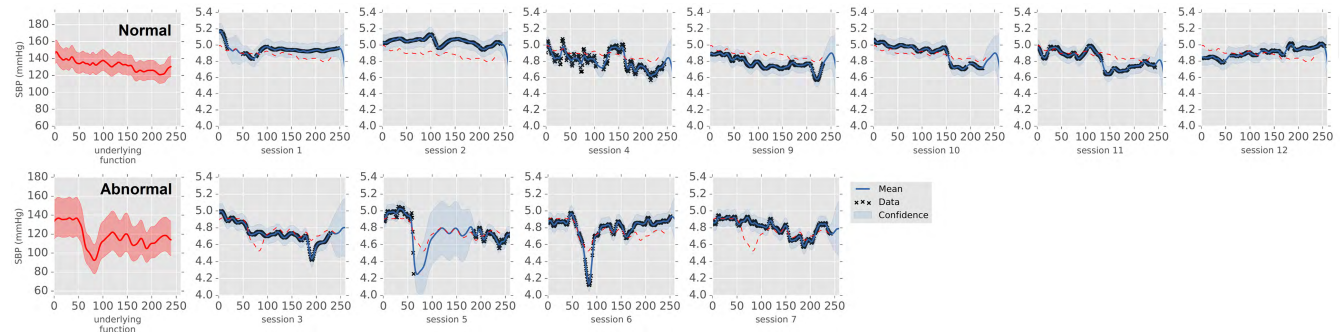


FIGURE 3. The dialysis data HGPs inferred SBP structure of a subject derived from sessions. For each row, the leftmost panel (in red) indicates the inferred latent structure of the SBP time-series $g_n(x)$ for the n th subject. Each subsequent blue panel is the log-scaled SBP time-series $f_{nr}(x)$ of the r th session, in comparison to the log-scaled latent structure (in dotted red line). The session indices are indicated in the bottom of each subplot. The times (measured in minutes) considered for each subplot are the amount of time elapsed since the start of a HD session. The normal and abnormal behaviors are learnt separately from top and bottom rows, respectively. The solid blue line in each subplot indicates the mean of the predicted function with 95% confidence interval (as shown in shaded blue area).

normal and abnormal trajectories in Figure 3), HGPs was able to express their difference via the level of uncertainty due to the disagreement among sessions: it was more confident in the normal case as the sessions were more closely related to each other, due to their similar magnitudes. In the case for abnormal sessions, only sessions 5 and 6 were similar at 50 to 100 minutes period (see Figure 3), while sessions 3 and 7 had a small decrease in the period of 150 to 200 minutes. These disagreements are expected as the human physiology is dynamic throughout the HD treatment, and we anticipate that a patient may deteriorate at different time-points for different HD treatment sessions. Greater uncertainty across the full 4-hour session indicates that abnormalities may occur at anytime within the session.

2) CLASSIFICATION OF HGPS SBP SESSIONS VIA CLUSTERING

Prior to one-class classification, one may argue that clustering methods can be applied here to separate “normal” and “abnormal” sessions directly, especially the case where we have an adequate number of normal vs. abnormal sessions. However, clustering methods cannot deal with missing values directly. For a fair comparison of our proposed model with other bench-marking clustering methods, we computed the missing values using HGPs for all normal and abnormal sessions. Unsupervised clustering was then performed on all data with number of clusters fixed to be two a priori. We had tried both clustering with and without normalization of the concatenated sessions, where the former was considered to cluster different shapes of the SBP trajectories, and the latter was to form clusters of SBP time-series with different magnitudes. The results of the best performance for each method are reported in Table 1. It was observed that all clustering methods had an accuracy around 50%, as similar to a random guess. This demonstrates that transitional time-invariant clustering methods fail to identify normal from abnormal sessions. Furthermore, these methods only take the mean trajectories into consideration, ignoring the variation of each data-point in a timeseries, reducing their abilities

TABLE 1. Results of different clustering approaches for the SBP dataset.

Method	Statistical Measures		
	Sensitivity	Specificity	Accuracy
KMean	0.70	0.31	0.52
SCluster	0.30	0.72	0.49
BIRCH	0.66	0.40	0.54
HClust	0.63	0.39	0.52
MOG	0.58	0.37	0.48

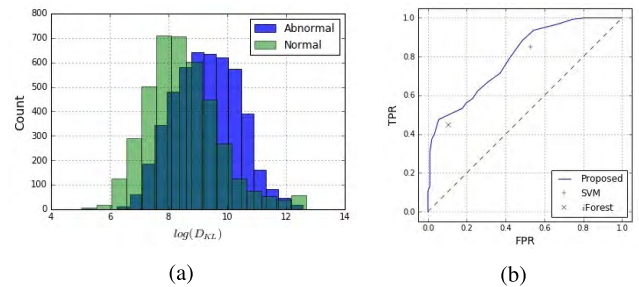


FIGURE 4. (a) The distributions of $\log(D_{KL})$ values of normal latent trajectories versus normal and abnormal sessions in the SBP dataset. (b) Classification result of the ROC curve of our proposed algorithm on the SBP dataset for various $\log(D_{KL})$ values, against one-class SVM and iForest using a classification probability threshold of 0.5. The diagonal dotted line indicate the reference.

to separate different populations. Therefore, formulation of a model of normality from the latent trajectories of normal subjects is the next logical step to improve identification of abnormal sessions as a form of “outlier” or “Novelty”.

Figure 4 (a) shows the distributions of $\log(D_{KL})$ values of normal latent trajectories versus normal and abnormal sessions. Smaller $\log(D_{KL})$ values indicate a higher similarity between two time-series, and vice versa. It is observed that the latent trajectories were similar to the normal sessions, with smaller $\log(D_{KL})$ values that are centered around 8. In comparison, the abnormal sessions tend to have larger $\log(D_{KL})$ as its distribution is shifted to the right of the plot with a flat top and centred around 9. As there is an overlap between the two distributions, making it harder to define a suitable $\log(D_{KL})$ value to set normal and abnormal sessions apart. Nevertheless, the $\log(D_{KL})$ figure shows that a certain

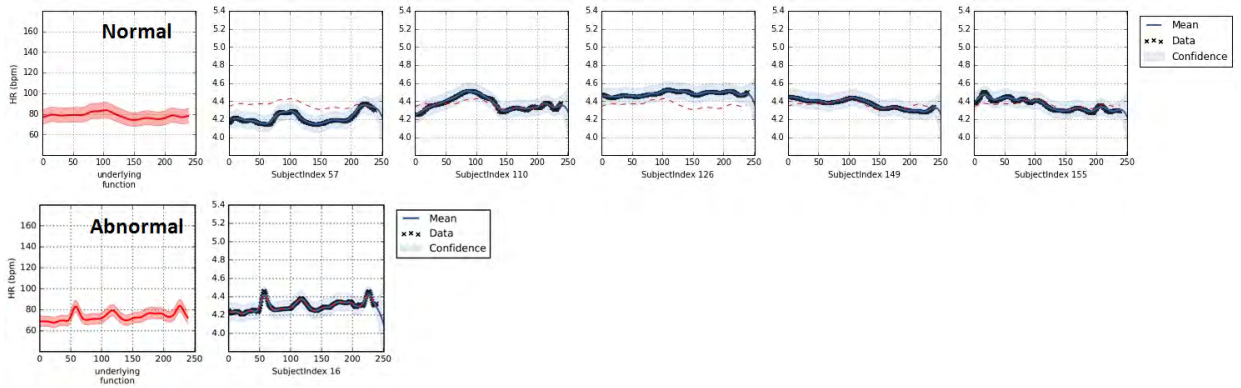


FIGURE 5. The SDU data HGPs inferred HR structure of a cluster derived from subjects. For each row, the leftmost panel (in red) indicates the inferred latent structure of the HR time-series $g_n(x)$ for the n th cluster. Each subsequent blue panel is the log-scaled HR time-series $f_{nr}(x)$ of the r th subject, in comparison to the log-scaled latent structure (in dotted red line). The subject indices are indicated in the bottom of each subplot. The times (measured in minutes) considered for each subplot are the last four hours of discharge. The normal and abnormal behaviors are learnt separately from top and bottom rows, respectively. The solid blue line in each subplot indicates the mean of the predicted function with 95% confidence interval (as shown in shaded blue area).

high quantile of abnormal can be effectively picked out with minimal inclusion of normals. Some of those “normals” may be highly abnormal but just didn’t meet the labeling criteria. The modal difference is useful as a patient’s session is repeated over time. We therefore tested our algorithm for a range of threshold values that were selected from the $\log(D_{KL_m})$ distribution.

3) ONE-CLASS CLASSIFICATION RESULTS

The mean and standard derivation of the $\log(D_{KL_m})$ were estimated to be 8.30 ± 0.11 and 1.01 ± 0.03 using a 26-fold LOO cross-validation. The results show that the KL values were consistently similar among the latent trajectories as they have small variance values. We then proceed with the classification by varying the threshold within the range of $[5.2, 11.4]$ (i.e., ± 3 standard deviation from the mean in the $\log(D_{KL_m})$ distribution), as shown in Figure 4 (b). Furthermore, we compared our results to the bench-marking algorithms such as one-class SVM and *iForest*. It was shown that our approach outperformed other methods with higher mean TPR when comparing with the same mean FPR. Furthermore, we have obtained an Area-Under-Curve of 0.80, with an accuracy of 0.71 ± 0.07 at threshold value of $\log(D_{KL_m}) = 9.22$, outperforming other bench-marking algorithms. The accuracy of one-class SVM and *iForest* is 0.66 ± 0.06 and 0.67 ± 0.01 , respectively. As a whole, our proposed algorithm allows the threshold of $\log(D_{KL_m})$ to vary, and providing the flexibility of choosing a user-defined sensitivity and specificity.

C. HR DATASET

1) INFERRING LATENT HR TRAJECTORIES USING THE HGPS MODEL

A total of 32 latent HR trajectories were inferred using the HGPs model, from 32 clusters of normal subjects. The top row of Figure 5 demonstrates the inferred normal trajectories using HGPs. The latent structure of the HR time-series $g_n(x)$ for the n th cluster is plotted in the leftmost panel with

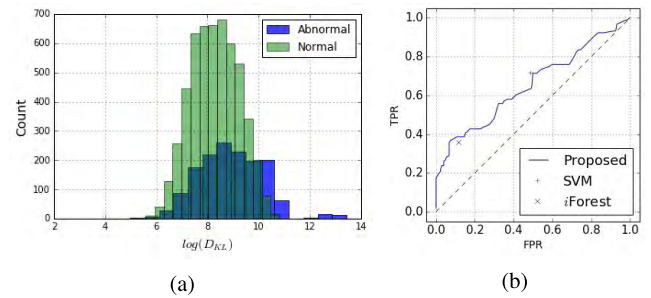


FIGURE 6. (a) The distributions of $\log(D_{KL})$ values of normal latent trajectories versus normal and abnormal subjects in the HR dataset. (b) Classification result of the ROC curve of our proposed algorithm on the HR dataset for various $\log(D_{KL_m})$ values, against one-class SVM and *iForest* using a classification probability threshold of 0.5. The diagonal dotted line indicate the reference.

95% confidence interval, and each subsequent panel shows the log-scaled HR time-series $f_{nr}(x)$ of the r th subject with 95% confidence interval. Similarly, the latent structure of an abnormal subject learnt using HGPs is shown in the bottom row of Figure 5. It was observed that some abnormal subjects had shared similar HR magnitudes and patterns as those of the normal subjects, which might be challenging to perform classification.

Figure 6 (a) shows the distributions of $\log(D_{KL})$ values of normal latent trajectories versus normal and abnormal subjects. Smaller $\log(D_{KL})$ values indicate a higher similarity between two time-series, and vice versa. It was expected that the normal and abnormal latent trajectories were heavily overlapped as the HR time-series were similar among the subjects. They both had $\log(D_{KL})$ values that are centered around 8.5, but the abnormal subjects tend to have a larger $\log(D_{KL})$ as its distribution is slightly-shifted to the right of the plot. As there were only 47 abnormal subjects, it was difficult to inspect their distribution. We then proceed to test our algorithm for various threshold values to attempt to classify normal and abnormal subjects.

TABLE 2. Results of different clustering approaches for the HR dataset.

Method	Statistical Measures		
	Sensitivity	Specificity	Accuracy
KMean	0.49	0.43	0.44
SClust	0.49	0.62	0.59
BIRCH	0.54	0.47	0.52
HClust	0.54	0.47	0.52
MOG	0.72	0.19	0.31

2) CLASSIFICATION OF HGPS HR SUBJECTS VIA CLUSTERING

We inferred the subject-wise mean function using HGP and considered such function as the data representing a subject. Unsupervised clustering approaches were then applied to the population data to identify normal and abnormal subjects, assuming the number of clusters was two. The results of the performance for five clustering methods are reported in Table 2. It was interesting to observe that non-linear methods achieved approximately 50% accuracy, as similar to a random guess, demonstrating the difficulty for separating normal from abnormal subjects. Furthermore, these methods cannot deal with missing data-points in each time-series, and they assumed the data-points were independent, therefore limiting their use for patient-specific physiological monitoring and time-series clustering.

3) ONE-CLASS CLASSIFICATION RESULTS

The mean and standard deviation of the $\log(D_{KL_m})$ (i.e., 5.25 ± 0.03 and 0.47 ± 0.01) were estimated from the 32-fold LOO cross-validation. These KL values show that the latent trajectories were very similar as the variation values were very small. The statistical measures were then performed for various threshold values (i.e., $[0, 15]$) and the results are shown in Figure 6 (b). Furthermore, we compared our results again with the bench-marking algorithms, one-class SVM and *iForest*. It was shown that our method outperformed the *iForest* but was similar when compared with the one-class SVM. Our proposed algorithm has an Area-Under-Curve of 0.65, with an accuracy of 0.65 ± 0.06 at threshold value of $\log(D_{KL_m}) = 9.36$. This indicates our algorithm is better than the bench-marking algorithms, where the accuracy of one-class SVM and *iForest* is 0.62 ± 0.07 and 0.62 ± 0.02 , respectively. The results were expected as normal and abnormal subjects exhibited similar HR trajectories in both magnitude and trend, making it very challenging to distinguish the difference. Furthermore, although the results in the HR dataset does not demonstrate an improvement over other algorithms, there were only 47 abnormal time-series available. More data will be required to perform a better analysis. Since the variability of the abnormal HR time-series (see Figure 5 bottom row) might occur in a smaller window, comparing the trajectories in a long time window might “smooth out” the differences. Future work can analyze the KL divergence of a moving window across the time-series.

VI. CONCLUSION AND FUTURE WORKS

Monitoring and predicting patient well-being over a long period of time is a current challenge in predictive health informatics. Within a time-series, patients with similar

outcomes may exhibit divergent physiological patterns, forming a multiplicity of clusters and sub-groups. These are uninformative when attempting to infer a generalized population model. Our proposed approach addresses these challenges by using patient/subgroup-specific time-series modeling. This is a Bayesian Gaussian Processes framework, from artifact removal of vital-sign time-series to classification of “normal” and “abnormal” patterns of physiological trends. We have introduced a hierarchical structure to infer the latent trajectory from similar sessions/treatments or clusters, and demonstrated the possibility of using these latent trajectories to build a model of normality to classify the patient’s instability.

One limitation in our approach is that we assume that the number of clusters and the cluster memberships are known a priori. One extension to this approach would be to incorporate Bayesian non-parametric clustering to further improve our model of normality. Furthermore, our current approach considers only one vital sign within each of the data sets. Future work will focus on combining multiple vital signs to form a more robust monitoring of patient deterioration. More specifically, combining extreme value statistics and Poisson point processes to model the symmetric KL-divergences from the multiple vital signs would provide a more reliable estimation of the state of health of a patient in a continuous monitoring setting.

REFERENCES

- [1] H. Gao et al., “Systematic review and evaluation of physiological track and trigger warning systems for identifying at-risk patients on the ward,” *Intensive Care Med.*, vol. 33, no. 4, pp. 667–679, 2007.
- [2] L. Tarassenko, A. Hann, and D. Young, “Integrated monitoring and analysis for early warning of patient deterioration,” *Brit. J. Anaesth.*, vol. 97, no. 1, pp. 64–68, Jul. 2006.
- [3] M. A. Pimentel, D. A. Clifton, L. Clifton, P. J. Watkinson, and L. Tarassenko, “Modelling physiological deterioration in post-operative patient vital-sign data,” *Med. Biol. Eng. Comput.*, vol. 51, no. 8, pp. 869–877, 2013.
- [4] J. A. Quinn, C. K. I. Williams, and N. McIntosh, “Factorial switching linear dynamical systems applied to physiological condition monitoring,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 9, pp. 1537–1551, Sep. 2009.
- [5] I. Stanculescu, C. K. Williams, and Y. Freer, “A hierarchical switching linear dynamical system applied to the detection of sepsis in neonatal condition monitoring,” in *Proc. UAI*, 2014, pp. 752–761.
- [6] E. Fox, M. I. Jordan, E. B. Sudderth, and A. S. Willsky, “Sharing features among dynamical systems with beta processes,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 549–557.
- [7] H. L. Li-Wei, S. Nemati, R. P. Adams, and R. G. Mark, “Discovering shared dynamics in physiological signals: Application to patient monitoring in ICU,” in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Aug. 2012, pp. 5939–5942.
- [8] O. Stegle, S. V. Fallert, D. J. C. MacKay, and S. Brage, “Gaussian process robust regression for noisy heart rate data,” *IEEE Trans. Biomed. Eng.*, vol. 55, no. 9, pp. 2143–2151, Sep. 2008.
- [9] L. Clifton, D. A. Clifton, M. A. F. Pimentel, P. J. Watkinson, and L. Tarassenko, “Gaussian processes for personalized e-health monitoring with wearable sensors,” *IEEE Trans. Biomed. Eng.*, vol. 60, no. 1, pp. 193–197, Jan. 2013.
- [10] S. Aghabozorgi, A. S. Shirkhorshidi, and T. Y. Wah, “Time-series clustering—A decade review,” *Inf. Syst.*, vol. 53, pp. 16–38, Oct. 2015.
- [11] M. A. Pimentel, D. A. Clifton, and L. Tarassenko, “Gaussian process clustering for the functional characterisation of vital-sign trajectories,” in *Proc. IEEE Int. Workshop Mach. Learn. Signal Process. (MLSP)*, Sep. 2013, pp. 1–6.
- [12] H.-C. Kim and J. Lee, “Clustering based on Gaussian processes,” *Neural Comput.*, vol. 19, no. 11, pp. 3088–3107, 2007.

- [13] J. C. Ross and J. G. Dy, "Nonparametric mixture of Gaussian processes with constraints," in *Proc. ICML*, 2013, pp. 1346–1354.
- [14] Y. Xu, Y. Xu, and S. Saria. (2016). "A Bayesian nonparametric approach for estimating individualized treatment-response curves." [Online]. Available: <https://arxiv.org/abs/1608.05182>
- [15] J. Hensman, N. D. Lawrence, and M. Rattray, "Hierarchical Bayesian modelling of gene expression time series across irregularly sampled replicates and clusters," *BMC Bioinf.*, vol. 14, no. 1, pp. 1–12, 2013.
- [16] J. Hensman, M. Rattray, and N. D. Lawrence, "Fast nonparametric clustering of structured time-series," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 2, pp. 383–393, Feb. 2015.
- [17] S. Park and S. Choi, "Hierarchical Gaussian process regression," in *Proc. 2nd Asian Conf. Mach. Learn.*, vol. 13, Nov 2010, pp. 95–110.
- [18] C. K. Williams and C. E. Rasmussen, *Gaussian Processes for Machine Learning*. Cambridge, MA, USA: MIT Press, vol. 2, no. 3, 2006, p. 4.
- [19] T. Zhu, G. W. Colopy, C. W. Pugh, and D. A. Clifton, "Identifying patient-specific trajectories in haemodialysis using bayesian hierarchical Gaussian processes," in *Proc. IEEE EMBS Int. Conf. Biomed. Health Inform. (BHI)*, Mar. 2018, pp. 186–189.
- [20] G. W. Colopy, M. A. Pimentel, S. J. Roberts, and D. A. Clifton, "Bayesian optimisation of Gaussian processes for identifying the deteriorating patient," in *Proc. IEEE EMBS Int. Conf. Biomed. Health Inform. (BHI)*, Feb. 2017, pp. 85–88.
- [21] C. MacEwen, "Can data fusion techniques predict adverse physiological events during haemodialysis?" Ph.D. dissertation, Univ. Oxford, Oxford, U.K., 2016.
- [22] A. Hann, "Multi-parameter monitoring for early warning of patient deterioration," Ph.D. dissertation, Univ. Oxford, Oxford, U.K., 2008.
- [23] G. W. Colopy, T. Zhu, L. Clifton, S. J. Roberts, and D. A. Clifton, "Likelihood-based artefact detection in continuously-acquired patient vital signs," in *Proc. 39th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jeju, South Korea, Jul. 2017, pp. 2146–2149.
- [24] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, Aug. 2000.
- [25] T. Zhang, R. Ramakrishnan, and M. Livny, "BIRCH: An efficient data clustering method for very large databases," *ACM SIGMOD Rec.*, vol. 25, pp. 103–114, Jun. 1996.
- [26] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 27:1–27:27, 2011.
- [27] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation-based anomaly detection," *ACM Trans. Knowl. Discovery Data*, vol. 6, no. 1, pp. 3:1–3:39, Mar. 2012.
- [28] GPy. (2012). *GPy: A Gaussian Process Framework in Python*. [Online]. Available: <http://github.com/SheffieldML/GPy>



of Oxford. Her research interests include investigating the development of machine learning for understanding complex patient data, with a special emphasis on Bayesian inference, deep learning, and applications involving the developing world.



metrics for personalized medical modeling, the robust automatization of statistical inference, and the experimental design of machine learning-based clinical trials.

TINGTING ZHU received the B.Sc. degree in electrical engineering from the University of Malta, the M.Sc. degree in biomedical engineering from the University College London, and the D.Phil. degree in information engineering and biomedical engineering from the Centre for Doctoral Training in Healthcare Innovation, Department of Engineering Science, University of Oxford, in 2016. She is currently an Associate Member with the Department of Engineering Science, University

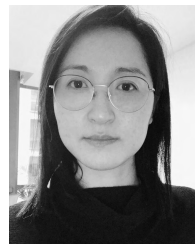
GLEN WRIGHT COLOPY received the B.S. degrees in mathematics and economics from the College of William and Mary, the M.Sc. degree in applied statistics from the University of Oxford, the M.Sc. degree in operations research from the North Carolina State University, and the D.Phil. degree from the Department of Engineering Science, University of Oxford, in 2018. He is currently with Current Health, Edinburgh, Scotland. His research interests include Bayesian nonpara-



CLARE MACEWEN received the medical and B.A. degrees from the University of Oxford, and the D.Phil. degree from the Nuffield Department of Medicine, University of Oxford, in 2016. She is currently a Critical Care Nephrologist with the Oxford University Hospitals NHS Trust. Her research interests include intra-dialytic physiology, and novel methods to predict deterioration during hemodialysis.



KATHERINE NIEHAUS received the B.S. degree in biomechanical engineering and the M.S. degree in bioengineering from Stanford University, and the D.Phil. degree from the Department of Engineering Science, University of Oxford, in 2017. She is currently a Scientist with Apple Inc., where she focuses on health projects. Her research involves machine learning for medical and biological applications (EHR, genomic, and sensor data), with an interest in interpretability.



the Department of Mechanical Engineering, Shanghai Jiao Tong University, China. She is currently leading the bidirectional translation of research between the Oxford and Chinese sites of CHI Lab, with a personal interest in deep learning and healthcare applications.

YANG YANG received the B.Sc. and D.Phil. degrees in mechanical engineering from the Department of Mechanical Engineering, Shanghai Jiao Tong University, China, in 2006 and 2013, respectively. She is currently a Senior Research Associate with the Department of Engineering Science, University of Oxford, which follows two years with the Computational Health Informatics (CHI) Lab as the Oxford University's Second K. C. Wong Fellow, after her previous work at



CHRIS W. PUGH received the D.Phil. degree in immunology from the University of Oxford, in 1981. He subsequently qualified in medicine, in 1985, and has gone on to become a Professor of renal medicine with the University of Oxford. Since 1990, he has been publishing extensively on the biological effects of low oxygen levels. Over the last eight or nine years, he has developed an interest in monitoring the vital signs of patients undergoing hemodialysis.



DAVID A. CLIFTON received the degree in information engineering from the Department of Engineering Science, University of Oxford. He is currently an Associate Professor with the Department of Engineering Science University of Oxford. He spent four years as a Postdoctoral Researcher in biomedical engineering with the University of Oxford, before his appointment to the faculty, at which point he started the CHI Lab. In 2017, CHI Lab opened its second site in Suzhou,

China, with support from the Chinese government. His research focuses on the development of machine learning for tracking the health of complex systems. His previous research resulted in patented systems for jet-engine health monitoring, used with the engines of the Airbus A380, the Boeing 787 "Dreamliner," and the Eurofighter Typhoon. Since 2008, he has been focusing mostly on healthcare applications. During his postdoctoral research, he worked on the early-warning systems that are now implemented within the SEND system, which is now used to monitor 20 000 patients each month in the NHS. He is a Research Fellow of the Royal Academy of Engineering, the Visiting Chair in AI for Healthcare at The University of Manchester, and a Fellow of Fudan University, China.

...