An Extreme Function Theory for Novelty Detection

David A. Clifton, Lei Clifton, Samuel Hugueny, David Wong, and Lionel Tarassenko

***Abstract*—We introduce an *extreme function theory* as a novel method by which probabilistic novelty detection may be performed with functions, where the functions are represented by time-series of (potentially multivariate) discrete observations. We set the method within the framework of Gaussian processes (GP), which offers a convenient means of constructing a distribution over functions. Whereas conventional novelty detection methods aim to identify individually extreme data points, with respect to a model of normality constructed using examples of "normal" data points, the proposed method aims to identify extreme *functions*, with respect to a model of normality constructed using examples of "normal" functions, where those functions are represented by time-series of observations. The method is illustrated using synthetic data, physiological data acquired from a large clinical trial, and a benchmark time-series dataset.**

***Index Terms*—Functional analysis, Gaussian processes, signal processing algorithms.**

I. Introduction

NOVELTY detection [1] is a fundamental task in anomaly detection, outlier detection, and one-class classification, in which we wish to identify if newly-observed data are in some sense "novel" with respect to previously-observed examples. Novelty detection can be viewed as a hypothesis test, in which we wish to determine if a previously-unseen test dataset has the same characteristics as a training set of "normal" data.

Definition I.1: Let a "training" dataset $\mathbf{X} = \{\mathbf{x}_i\}_{i=1\ldots n}$ be a collection of n examples of "normality". For the case of non-timeseries data, let the i^{th} example $\mathbf{x}_i$ be a q-dimensional point $\mathbf{x}_i \in \mathbb{R}^q$.

Definition I.2: Let a test set $\mathbf{X}^* = \{\mathbf{x}_i^*\}_{i=1\ldots n^*}$, where each of the n^* test examples $\mathbf{x}_i^*$ is defined as for the training data in definition I.1.

In novelty detection, we wish to test the null hypothesis H_0 that the test set $\mathbf{X}^*$ has the same characteristics as the training set $\mathbf{X}$, preferably in some probabilistic sense. A model of normality $\mathcal{M}$ is typically constructed from the training set $\mathbf{X}$, and H_0 is tested by comparing the test set $\mathbf{X}^*$ to the model $\mathcal{M}$. The majority of novelty detection work in the literature uses point-wise novelty detection, in which $\mathcal{M}$ defines a decision boundary in the data space $\mathbb{R}^q$. Individual test points $\mathbf{x}_i^*$ are then compared to the "normal" region of data space $\mathbb{F}_{\mathcal{M}} \subseteq \mathbb{R}^q$ defined by the decision boundary, and H_0 is deemed to hold (and $\mathbf{x}_i^*$ is classified as being "normal") if $\mathbf{x}_i^* \in \mathbb{F}_{\mathcal{M}}$. One of the most widely used of such methods is discriminant and based on the support vector machine [2], [3], in which novelty detection takes place in the high-dimensional reproducing kernel Hilbert space (RKHS) corresponding to some kernel function. Here, test points $\mathbf{x}_i^*$ are classified independently from one another, by being compared to a decision boundary in the RKHS. Generalizations of this point-wise approach to novelty detection have been proposed in, for example, the states within a hidden Markov model [4], the output of a Kalman filter [5], or the states of a factorial switched Kalman filter [6]. More recently, one-class classifiers using Gaussian processes (GPs) have been proposed [7]–[10] that take a similarly point-wise approach to novelty detection, constructing a function f to use as the model of normality $\mathcal{M}$ and dividing data space $\mathbb{R}^q$ into regions with high support $f(\mathbf{x}) \to 1$ and low support $f(\mathbf{x}) \to 0$ depending on whether those regions are close to those occupied by "normal" training data $\mathbf{X}$, or not, respectively. Change-point detection has been implemented within a GP framework to identify and rectify sensor failures [11].

This paper aims to tackle the case of timeseries novelty detection, as typically occurs in the analysis of data acquired from critical systems such as jet engines and human patients. In such applications, the test data are a potentially multivariate time-series that we wish to classify as being either normal or abnormal; i.e., the question may be framed "is this timeseries of human vital signs indicative of a normal patient or a deteriorating patient?" Conventional point-wise novelty detection is appropriate when instances of *independent* objects are to be classified; e.g., the classification of different mammograms as being "normal" or "abnormal" [12]. However, when the test data represent a timeseries, the i.i.d. assumption typically does not hold, and adopting a point-wise, sample-by-sample approach to classification can result in large numbers of misclassifications, because we are making large numbers of assumedly independent decisions (perhaps at the sampling rate of the data). Instead, we suggest that a single decision can be taken, testing an entire timeseries, which represents a test function. We will thus adopt a function-wise approach to novelty detection, where the functions are represented by timeseries of discrete observations.

The GP framework offers a convenient, non-parametric method of defining a probability distribution over a Lebesque space of functions $\ell_p \triangleq (\mathcal{X}, \mu)$ where $\mathcal{X}$ is Borel-measurable (often $\mathbb{R}^q$ or discretizations thereof) and where μ is a valid probability measure, such as a multivariate Gaussian density

Manuscript received August 01, 2012; revised November 06, 2012; accepted December 01, 2012. Date of publication December 13, 2012; date of current version January 22, 2013. The work of D. Clifton was supported by the Centre of Excellence in Personalised Healthcare funded by the Wellcome Trust and EPSRC under Grant WT 088877/Z/09/Z. The work of L. Clifton and D. Wong was supported by the NIHR Biomedical Research Centre, Oxford, U.K. The work of S. Hugueny was supported by the EPSRC LSI Doctoral Training Centre, Oxford, U.K. Matlab implementations of the methods described in this paper may be downloaded from http://www.robots.ox.ac.uk/~davidc. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Ery Arias-Castro.

The authors are with the Department of Engineering Science, University of Oxford, Oxford OX3 7DQ, U.K. (e-mail: davidc@robots.ox.ac.uk)

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTSP.2012.2234081

over a (potentially infinite) number of random variables (rvs). This paper describes a principled, probabilistic approach to functional novelty detection by considering *extreme functions*.

II. Background and Notation

Our work is related to that of functional data analysis (FDA), which is an active branch of research concerning inference where "the data are functions" [13]. Reviews of FDA [14]–[16] identify two main approaches: (i) regularization-based methods, in which functions are resampled (interpolated) such that observations occur on a regular sampling grid, and (ii) filtering-based methods, in which functions are expanded onto a finite-dimensional function basis after smoothing has been applied. The latter is the current focus of most research in FDA, where popular tasks include using the basis coefficients for clustering [17], principal component analysis, and linear discriminant analysis [13], [18], often set within a GP regression framework. However, while multiclass classification has been performed using the latter, there is little focus on novelty detection. The closest existing work to tackling the problem considered in this paper is an F-test for functional data [19], which compares two functions, analogous to the standard F-test between two sets of data points.

Similarly, timeseries classification has been identified as a topic within anomaly detection [20], where many of the approaches are shared with those in the field of novelty detection (and where the latter is most often presented as a branch of machine learning).

In common with the majority of FDA work, and in keeping with the GP literature, we will first consider univariate timeseries (generalizing later to multivariate timeseries). Keeping to the notation of the functional and timeseries literature, we must redefine the notation from definitions I.1 and I.2, which were conventional notation for the i.i.d. case, giving:

Definition II.1: Let a training set of $i = 1 \ldots n$ examples of functional data (e.g., timeseries) be $\{\mathbf{x}_i, \mathbf{y}_i\}$. Let the i^{th} example (e.g., timeseries object) be a sequence $\mathbf{y}_{i,j}$ of $j = 1 \ldots n_i$ observations occurring at locations $\mathbf{x}_{i,j}$. For the case of univariate timeseries, the r.v. $\mathbf{y} \in \mathbb{R}$ and the index set of the r.v. $\mathbf{y}$ is time.

Note that, in general, there is no requirement that each of the n examples in a collection of functional data should be of the same length, and so the length of the i^{th} example is denoted by n_i. Furthermore, in the general case, there is no requirement that the n examples should consist of observations occurring at the same locations; for timeseries, this means that each example could be observed at different times.

We first consider the case of univariate GP regression to set notation, and initially define a single GP over one of the $i = 1 \ldots n$ sequences of observations y_i. We follow the standard treatment of [21].

Let a GP prior be defined over a latent variable $f(\mathbf{x}) \sim \mathcal{GP}(\mu_f(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$, using, for example, a squared-exponential (SE) covariance function $k(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\left(-\| \mathbf{x} - \mathbf{x}' \|^2 / 2\sigma_l^2\right)$, where $\| \cdot \|$ is the ℓ_2-norm, where σ_l^2 and σ_f^2 are the length-scale in the x-direction and the variance of f, respectively, and where the mean function $\mu_f(\mathbf{x}) = 0$. The $j = 1 \ldots n_i$ observations for this i^{th} timeseries are related to the latent variables via $y_{i,j} = f(\mathbf{x}_{i,j}) + \varepsilon$, with $\varepsilon \sim \mathcal{N}(0, \sigma_y^2)$, or $\mathbf{y}_i = \mathbf{f} = f(\mathbf{x}_i)$, dropping the j subscripts for the i^{th} timeseries object, as in [21].

Following the novelty detection approach, in which a model of normality $\mathcal{M}$ is constructed from "normal" training examples, we will assume the set of training timeseries $\{\mathbf{x}_i, \mathbf{y}_i\}$ comprises "normal" timeseries. We now wish to construct a model $\mathcal{M}$ from this training set, where $\mathcal{M}$ is a GP that describes the dynamics of the whole collection of "normal" examples. For instance, each example timeseries in the training set could be a timeseries of vital-sign observations taken from a "normal" patient, as will be considered in Section VII. The hyperparameters of the GP used to represent $\mathcal{M}$ could be determined in a number of ways; this paper maximizes the joint likelihood of all timeseries in the training set $p(\{\mathbf{x}_i, \mathbf{y}_i\}) = \prod_{i=1}^{n} p(\mathbf{y}_i | \mathbf{x}_i)$[21], an example of which will be demonstrated in Section VII. More complex examples of model construction considered in FDA include constructing a mixture of GPs [22].

Assuming the presence of a GP model $\mathcal{M}$, we next focus on determining how can we formulate a hypothesis test H_0 to classify a test timeseries $(\mathbf{x}^*, \mathbf{y}^*)$ as being "normal" (i.e., generated from $\mathcal{M}$ to some probability α) or otherwise "abnormal".

III. Assigning Probabilities to Functions

The formulation of our hypothesis test requires a mapping from functions to probabilities, $\mathbb{P} : f \to [0\ 1]$. A common approach used to assign a sequence of test observations $\mathbf{y}$ to one of many clusters in FDA [13] is to consider the marginal likelihood given the set of inputs $\mathbf{x}$ to the function f:

$$p(\mathbf{y}|\mathbf{x}) = \int p(\mathbf{y}|\mathbf{f}, \mathbf{x})\, p(\mathbf{f}|\mathbf{x})\, d\mathbf{f} \tag{1}$$

in which we have marginalized over the function values $\mathbf{f}$, using the GP distribution[1] over functions $p(\mathbf{f}|\mathbf{x}) \sim \mathcal{N}(\mu_f(\mathbf{x}), k(\mathbf{x}, \mathbf{x}))$, and where the likelihood $p(\mathbf{y}|\mathbf{f}, \mathbf{x}) \sim \mathcal{N}(\mathbf{f}, \sigma_y^2 \mathbf{I})$. The log marginal likelihood can be found in closed form as a marginalized Gaussian,

$$\log p(\mathbf{y}|\mathbf{x}) = -\frac{1}{2}\mathbf{y}^\top (\mathbf{K} + \sigma_y^2 \mathbf{I})^{-1} \mathbf{y} - \frac{1}{2} \log |\mathbf{K} + \sigma_y^2 \mathbf{I}| - \frac{n}{2} \log 2\pi \tag{2}$$

where $\mathbf{K} = k(\mathbf{x}, \mathbf{x})$. However, the resulting quantities are probability densities, not probabilities, and therefore scale with the dimensionality n of the input (i.e., the number of data in the timeseries). Fig. 1 shows log marginal likelihoods for 500 sample functions drawn from an example GP model $\mathcal{M}$ with a random positive semi-definite covariance matrix $\mathbf{K}$, where we have evaluated the sample functions at both $n = 50$ and $n = 100$ points in $\mathbf{x}$. It may be seen that doubling the dimensionality of the vector of function values $\mathbf{s}$ causes a scaling of the log marginal likelihoods. Therefore, while likelihoods may be useful for some tasks (e.g., maximum likelihood approaches to parameter fitting for some fixed value of n), they are not

[1] In keeping with the literature, we use the term *distributions* to refer to probability density functions (pdfs), which we will denote with lower-case letters $f, g, p \in \mathbb{R}^+$, and which we will make distinct from *distribution functions* (dfs), which we will denote with upper-case letters $F, G, P \in [0\ 1]$.

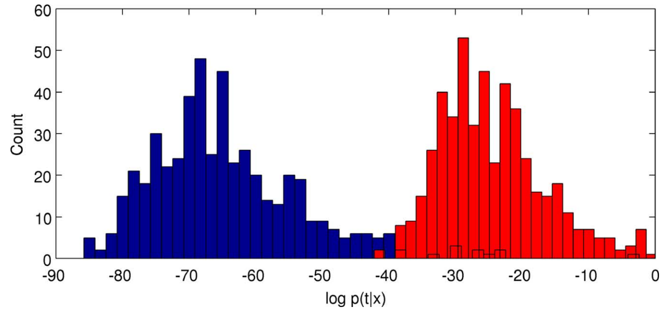


Fig. 1. Histograms of densities (marginal likelihoods) for 500 sample functions generated randomly from a GP model $\mathcal{M}$ with random $\mathbf{K}$, evaluated at $n = 50$ and $n = 100$ points in $\mathbf{x}$, shown in red (rightmost) and blue (leftmost), respectively.

suitable for comparing functions of arbitrary n, nor for direct probabilistic inference, because they are not probabilities.

Instead, we consider that the GP used as a model $\mathcal{M}$ may be used to determine the predicted values of observations $\mathbf{y}^*$ at times $\mathbf{x}^*$ using the GP regression framework [21]:

$$p(\mathbf{y}^*|\mathbf{x},\mathbf{y},\mathbf{x}^*) = \int p(\mathbf{y}^*|\mathbf{f}^*)\,p(\mathbf{f}^*|\mathbf{x},\mathbf{y},\mathbf{x}^*)\,\mathrm{d}\mathbf{f}^* \quad (3)$$

$$p(\mathbf{f}^*|\mathbf{x},\mathbf{y},\mathbf{x}^*) = \int p(\mathbf{f}^*|\mathbf{x},\mathbf{f},\mathbf{x}^*)\,p(\mathbf{f}|\mathbf{x},\mathbf{y})\,\mathrm{d}\mathbf{f} \quad (4)$$

In keeping with our function-wise philosophy, we will consider the joint distribution over all $(\mathbf{x}^*,\mathbf{y}^*)$ and the model, to evaluate the whole test function in a single classification decision, which is the joint posterior distribution over all functions, conditioned on the model,

$$p(\mathbf{f}^*|\mathbf{x},\mathbf{f},\mathbf{x}^*) \sim \mathcal{N}(\boldsymbol{\mu}^*,\mathbf{K}^*) \quad (5)$$

where the mean function and covariance matrix are, respectively,

$$\begin{aligned}\mu^* &= \mathbb{E}\left[p(\mathbf{f}^*|\mathbf{x},\mathbf{f},\mathbf{x}^*)\right] = k(\mathbf{x}^*,\mathbf{x})\left[k(\mathbf{x},\mathbf{x}) + \sigma_y^2\mathbf{I}\right]^{-1}\mathbf{y}\\ \mathbf{K}^* &= k(\mathbf{x}^*,\mathbf{x}^*) - k(\mathbf{x}^*,\mathbf{x})\left[k(\mathbf{x},\mathbf{x}) + \sigma_y^2\mathbf{I}\right]^{-1}k(\mathbf{x},\mathbf{x}^*)\end{aligned}$$

The hypothesis test H_0 could initially be framed by considering the chance of obtaining the test function $(\mathbf{x}^*,\mathbf{y}^*)$ from the GP model $\mathcal{M}$. The GP offers the convenience that the joint distribution over the test data (5) is multivariate Gaussian, and so we could formulate H_0 by determining the probability of observing an n-dimensional vector from the n-dimensional multivariate Gaussian distribution. For low n (such as $n = 1$ or 2), standard results from extreme value theory may be applied [23]. However, this is not straightforward for multivariate Gaussians of increasingly large dimensionality, as occurs in our case with $n \to \infty$ with functional data represented by timeseries, and standard results cannot be used [24]. Instead, we consider forming a distribution over the densities allowed by the GP model $\mathcal{M}$.

IV. A Distribution Over Function Densities

For some dimensionality n, the GP distribution over functions is multivariate Gaussian by definition, and a test function $\mathbf{f}^*$ may be evaluated to give a density z using the GP:

Definition IV.1: Let probability density z be given by a pdf, $z = f_n(\mathbf{f}^*)$, where $f_n = p(\mathbf{f}^*|\mathbf{x},\mathbf{f},\mathbf{x}^*) \sim \mathcal{N}(\mu_f,\mathbf{K})$, given by (5).

We emphasize that the notation $f_n(\mathbf{f}^*)$ refers to the multivariate Gaussian distribution f_n, defined over the latent function $\mathbf{f}^* = f(\mathbf{x}^*)$, drawing careful distinction between f_n (the multivariate Gaussian defined by the GP), $\mathbf{f}^*$ (the vector of output values from the latent function), and f (the latent function).

We note that the largest likelihood obtained from f_n occurs at the mode, and has value $\sup\{f_n\} = C_n^{-1}$, where $C_n = (2\pi)^{n/2}|\mathbf{K}|^{1/2}$. Therefore, densities take the range $z \in [0\,C_n^{-1}]$. Then,

Definition IV.2: Let a df $G_n(z)$ be defined over densities $z \in [0\ \ C_n^{-1}]$, according to

$$G_n(z) = \int_{\mathcal{R}} f_n(\mathbf{f}^*)\,\mathrm{d}\mathbf{f}^* \quad (6)$$

where the region of integration $\mathcal{R} = \{\mathbf{f}^* \mid f_n(\mathbf{f}^*) \in [0\ z]\}$.

This is an integration over all those points in $\mathbb{R}^n$ that result in a density $z' = p(\mathbf{f}^*|\mathbf{x},\mathbf{f},\mathbf{x}^*)$ which is lower than the density z. Thus, $G_n(z)$ is the probability that a sample function generated from the GP will be "more extreme" (i.e., have lower density z w.r.t. the GP) than our test function $\mathbf{f}^*$. It is important to note that $G_n(z)$ is effectively a df over level sets on the output of multivariate Gaussian f_n. That is, it is a distribution function defined over probability densities.

The integration in (6) can be evaluated in closed form by casting $f_n(\mathbf{f}^*)$ into an equivalent function over Mahalanobis radius r with respect to $\mathbf{K}$, such that

$$r^2 = \left(\mathbf{f}^* - \mu_f(\mathbf{x})\right)^\top \mathbf{K}^{-1}\left(\mathbf{f}^* - \mu_f(\mathbf{x})\right) = \mathbf{D} \circ \mathbf{D} \quad (7)$$

where $\mathbf{D} = \left(\mathbf{f}^* - \mu_f(\mathbf{x})\right)\mathbf{L}$, in which $\mathbf{L}$ is the Cholesky decomposition of $\mathbf{K}$, and where $\circ$ denotes the Hadamard (elementwise) matrix product. The integration may be performed in polar form, taking advantage of the radial symmetry of our multivariate Gaussian distribution over functions (5), which simplifies our expression for $G_n(z)$

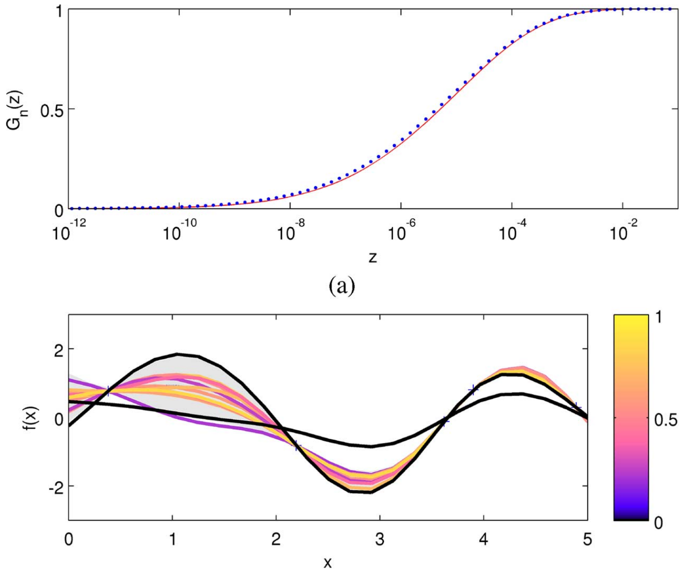
$$G_n(z) = \int \cdots \int_{\theta_1 \ldots \theta_{n-1}} \int_{r=r_0}^{\infty} f_n(r)\,\mathrm{d}\theta_1 \ldots \mathrm{d}\theta_{n-1}\,\mathrm{d}r \quad (8)$$

where r_0 is the Mahalanobis radius on our multivariate Gaussian distribution over functions that gives density z, via

$$z = f_n(r_0) = C_n^{-1}\exp\left(-\frac{1}{2}r_0^2\right), \quad r_0 = \sqrt{-2\log(zC_n)} \quad (9)$$

and where the latter is obtained by rearranging the former. This explicitly shows that $G_n(z)$ is the tail mass associated with the level set defined by z on our multivariate Gaussian $p(\mathbf{f}^*|\mathbf{x},\mathbf{f},\mathbf{x}^*)$, and is therefore the probability of observing a sample function generated from the GP of greater Mahalanobis radius than r_0, and hence the probability of observing a sample function with a lower probability density z' than z. Integration over all of the angles $\{\theta_1 \ldots \theta_{n-1}\}$ in the radial integration (8) yields

$$G_n(z) = \Omega_n|\mathbf{K}|^{1/2}\int_0^z (r_0)^{n-2}\,\mathrm{d}r \quad (10)$$



(a)

(b)

Fig. 2. (a) Df $G_n(z)$ over densities $z = f_n(\mathbf{f}^*) = p(\mathbf{f}^*|\mathbf{x}, \mathbf{f}, \mathbf{x}^*)$ and resulting empirical df for densities of 10^6 sample functions generated randomly from a GP, shown by red line and blue dots, respectively. Densities have been normalized by dividing through by $\sup\{f_n\}$; i.e., using zC_n. (b) Sample functions drawn from a GP, $\mathcal{M}$, colored by $G_n(z)$. "Abnormal" functions are shown by dark lines. A 95% confidence region (2 standard deviations) around the mean function is shown by the shaded gray background.

where $\Omega_n = (2\pi)^{n/2}/\Gamma(n/2)$ is the total solid angle of the hypersphere in n dimensions, and where $\Gamma(\cdot)$ is the Gamma function. As $n \to \infty$, the densities $z \to 0$, and so it will be convenient to express $G_n(w)$, where $w = \log z$. It will also be convenient to use the Cholesky decomposition $|\mathbf{K}|^{1/2} = \prod_{i=1}^{n} \mathbf{L}_{i,i}$ as $n \to \infty$. After iterative integration-by-parts (see the appendix),

$$\log G_{2p}(w) = w + \log H_1(w) \tag{11}$$

$$\begin{aligned}\log G_{2p+1}(w) &= w + \log H_2(w)\\ &\quad + \log\Big[1 + \exp(\log H_3(w) - \log H_2(w) - w)\Big]\end{aligned} \tag{12}$$

where the functions H_1, H_2, and H_3 are given in the appendix. Fig. 2(a) shows that the predicted df $G_n(z)$ over a range of densities using (11) is a close estimate of the empirical df obtained by generating 10^6 sample functions randomly from an example GP, with $n = 100$. It may be seen that as densities $z \to 0$, the df $G_n(z) \to 0$ as required: for a given density z, the probability that a sample function generated randomly from the GP will have a lower density (i.e., be more extreme) decreases as $z \to 0$. Conversely, as the density tends towards its maximum value, at the mode of the Gaussian distribution over function space, $z \to \sup\{f_n\} = C_n^{-1}$, the probability $G_n(z)$ of observing a more extreme function tends to 1.

Finally, we may use the df $G_n(z)$ to assign probabilities to each function that we wish to examine, given some GP model $\mathcal{M}$.

Definition IV.3: Let a test of the null hypothesis H_0 be defined as the comparison of $G_n(z)$ to some threshold probability $\alpha \in [0\ 1]$, where the null hypothesis holds if $G_n(z) \geq \alpha$ (thus classifying the data as being "normal"). The null hypothesis is rejected (and the test function is deemed "abnormal") when the corresponding $G_n(z) < \alpha$.

For example, $\alpha = 0.05$ would result in a hypothesis test in which functions are deemed "abnormal" if they could have been generated from the GP model $\mathcal{M}$ with a probability below significance level $\alpha = 0.05$.

Fig. 2(b) shows an example in which ten sample functions have been drawn from an example GP model $\mathcal{M}$, and which are therefore all "normal", and where two other functions have been shown for comparison (black lines). Each of the sample functions $\mathbf{f}^*$ has been colored according to $G_n(z)$, given $z = f_n(\mathbf{f}^*) = p(\mathbf{f}^*|\mathbf{x}, \mathbf{f}, \mathbf{x}^*)$. Those functions that lie close to the mean function take higher values of $G_n(z)$, while those that stray away from the mean function take lower values. Importantly, we note that the ten "normal" sample functions (for which the null hypothesis H_0 should not be rejected) take values of $G_n(z) > 0.15$, and so a hypothesis test (e.g., with $\alpha = 0.05$) would typically not reject H_0. The two functions shown by black lines, however, have $G_n(z) < 10^{-3}$. Therefore, H_0 for these two functions would be rejected by a hypothesis test with $\alpha = 0.05$, and the functions shown by black lines would be deemed to be "abnormal".

V. Extreme Function Distributions

Extreme value theory is a branch of statistics that considers the distribution of extrema (such as the minimum or maximum value observed in a set of data) in low-dimensional spaces. The majority of the literature [25], [26] is concerned with univariate data; some extensions into $\mathbb{R}^2$ or $\mathbb{R}^3$ have been described using copulae to estimate the dependence between rvs. These are, by definition, point-wise approaches, in which the single most extreme point in a set of data is considered. This point-wise approach has been extended to higher-dimensional work in [24] and [27]. We will now consider the extension of this method to our functional application, and show how it can be used in conjunction with $G_n(z)$ to provide identification of extreme functions, characterized by n discrete observations with $n \to \infty$, using *extreme function distributions*.

Fig. 3(a) shows functions sampled randomly from the GP model $\mathcal{M}$ considered previously, where each is the most extreme of a set of m sample functions, for increasing m, and where "most extreme" is defined as being that function with the lowest density $z = f_n(\mathbf{f}^*) = p(\mathbf{f}^*|\mathbf{x}, \mathbf{f}, \mathbf{x}^*)$, as given by the n-dimensional multivariate Gaussian distribution in (5). With increasing m, as shown by the colors in Fig. 3(a), the most extreme function observed in a set of m randomly-generated functions becomes increasingly more extreme, moving away from the mean function. This follows the intuition that, as we draw more data from the underlying distribution, we would expect the extremum of those data to be "more extreme" as we draw more and more data [24]. However, all of the functions generated from $\mathcal{M}$ are all "normal" functions, in that they have been generated from our GP that represents $\mathcal{M}$, and so we would like the null hypothesis H_0 not to be rejected. That is, they may be extreme, but they are not "abnormal" – they are extreme only because we have observed many realizations from the GP.

If we treat these functions as before, and assign probabilities to them using $G_n(z)$, we obtain the results shown in Fig. 3(b),

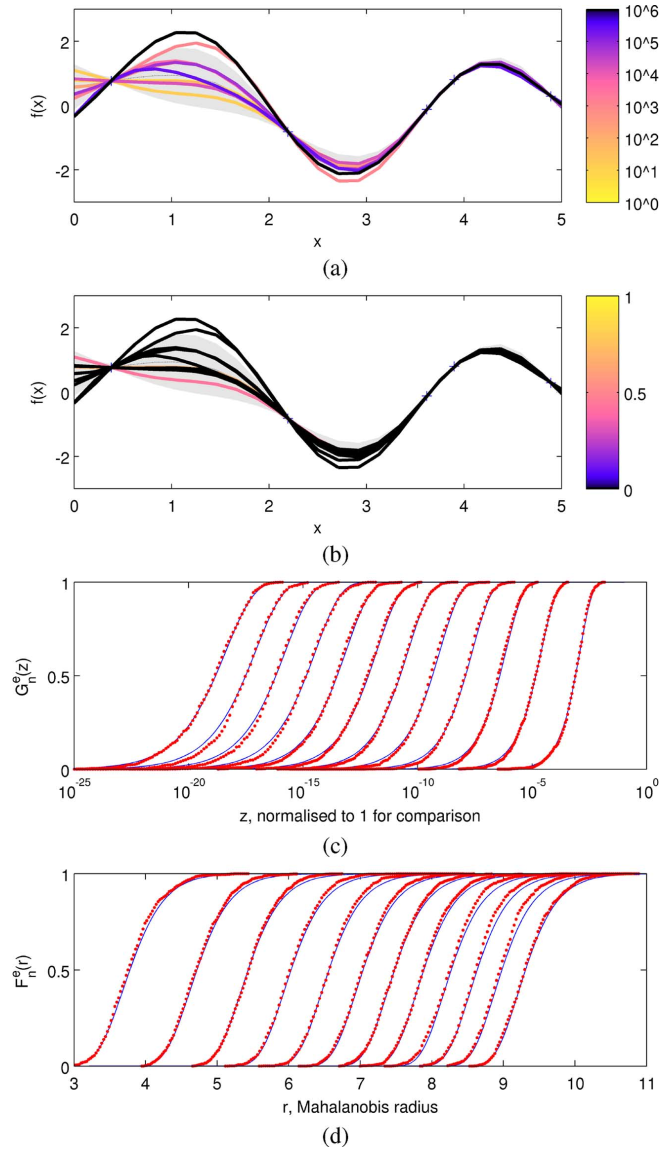


Fig. 3. Extreme functions generated by observing $m = 10^1, 10^2, \ldots, 10^6$ functions from a GP and selecting that with the lowest probability $G_n(z)$; (a) colored according to m (b) colored according to $G_n(z)$. Plots (c) and (d) show extreme function distributions $G_n^e(y)$ and $F_n^e(r)$, respectively defined over densities y and Mahalanobis radii r, giving predictions (blue lines) and results obtained by randomly generating 10^6 extreme functions (red dots). Distributions are shown for $m = 100$, with dimensionality $n = 5, 10, 15, \ldots, 55, 60$ from right to left in (c), and from left to right in (d).

in which most of the functions are assigned very low probabilities, $G_n(z) < 10^{-3}$ and hence H_0 would be incorrectly rejected for the majority of the functions shown. However, we can adapt extreme value theory to our functional case by considering the extremes in probability density z. Given the df $G_n(z)$, which is itself univariate in densities z, we note that the most extreme sample of a set of observations from $G_n(z)$ will have a density the distribution of which asymptotically converges to the Weibull df for minima [26],

$$G_n^e(z) = 1 - \exp\left[-\left(\frac{z}{c_m}\right)^{\alpha_m}\right] \qquad (13)$$

This is because the Fisher-Tippett theorem [28], on which extreme value theory is based, may be used to show that all non-degenerate functions[2] are in the domain of attraction of the generalized extreme value (GEV) distribution in their extrema, and that the limiting form of the GEV for our case, in which probability densities are truncated over the domain $z \in [0\ C_n^{-1}]$, is the Weibull.

The Weibull has scale and shape parameters c_m and α_m, respectively, for some value of m, which may be estimated [24] as being $c_m = G_n^{\leftarrow 1/m}$ and $\alpha_m = m\, c_m\, g_n(c_m)$, in which $G_n^{\leftarrow 1/m}$ is the $1/m$ quantile on $G_n(z)$, which may found to arbitrary accuracy because we can use $\log G_n(w)$ in closed form (11) using $w = \log z$. Also, $g_n(z)$ is the pdf associated with the df $G_n(z)$, and is straightforwardly the integrand in (10),

$$g_n(z) = \Omega_n(r_0)^{(n-2)} \prod_{i=1}^{n} \mathbf{L}_{i,i} \qquad (14)$$

Thus, using this non-standard extension of extreme value theory over a probability density, we have obtained a df $G_n^e(z)$ that allow us to determine the location in density space z for the densities of extreme functions, given some number of observed functions m. This is shown in Fig. 3(c) for $m = 100$, with increasing dimensionality n, where it may be seen that these dfs $G_n^e(z)$ closely match the densities of extreme functions observed from random sampling of the GP.

Definition V.1: Define an *extreme function distribution* as being $F_n^e(r)$ over Mahalanobis radii r on the multivariate Gaussian posterior distribution $f_n(\mathbf{f}^*) = p(\mathbf{f}^*|\mathbf{x}, \mathbf{f}, \mathbf{x}^*)$ using $r = \sqrt{-2\log(zC_n)}$ from (9). $F_n^e(r)$ is the probability that, if a set of m functions is observed from $\mathcal{M}$, then the most extreme function of that set will have Mahalanobis radius (with respect to $\mathcal{M}$) less than r.

Therefore, $F_n^e(r)$ is the probability that a test function (which has Mahalanobis radius r w.r.t. $\mathcal{M}$) is "abnormal", given a model of normality $\mathcal{M}$.

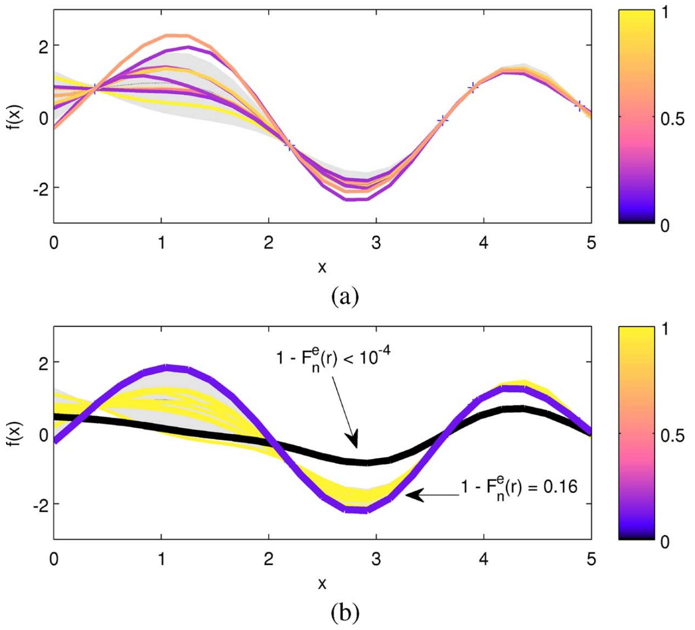
The extreme function distribution for our example $\mathcal{M}$ is shown in Fig. 3(d), again showing close agreement with the results of random sampling of functions from the GP. As required, the probability of abnormality $F_n^e(r)$ increases with r. Increasing the value of m (i.e., generating more functions from $\mathcal{M}$) results in the support of $F_n^e(r)$ shifting further up the r-axis, as expected: if we observe more functions from $\mathcal{M}$, then we are more likely to observe functions with larger Mahalanobis radius r (i.e, be more "extreme"), with respect to $\mathcal{M}$.

VI. Hypothesis Testing With Extreme Functions

Finally, we may return to our extreme functions previously shown in Fig. 3(b) and, instead of assigning them probabilities w.r.t. $G_n(z)$, we assign probabilities for each using the extreme function distributions for the appropriate number of functions observed in each case, m.

Definition VI.1: Let a hypothesis test be defined as the test comparing the extreme value distribution $F_n^e(r)$ to some

[2] Non-degenerate functions are those that do not assign all probability mass to a single point in their domain. The multivariate Gaussian defined by the GP is therefore non-degenerate, as is the df defined over its densities, $G_n(z)$.



(a)

(b)

Fig. 4. (a) Extreme functions previously shown in Fig. 3, with probabilities now assigned using extreme function distribution $1 - F^e_n(r)$ given the appropriate value of m for each, and (b) using $1 - F^e_n(r)$ with $m = 100$, including two other functions shown by dark lines.

threshold probability $\alpha \in [0\ 1]$, rejecting the null hypothesis H_0 (and thus classifying functions as "abnormal") for which $1 - F^e_n(r) < \alpha$.

Fig. 4(a) shows the result of using the hypothesis test from definition VI.1, in which it may be seen that each extreme function now takes probabilities $1 - F^e_n(r) > 0.15$, and that the null hypothesis H_0 will no longer be rejected if $1 - F^e_n(r) < \alpha$ is used (with, for example, $\alpha = 0.05$). Therefore, the sample functions will be classified as being "normal" functions w.r.t to our GP, $\mathcal{M}$, as we would hope, because they were all generated from $\mathcal{M}$ and are thus "normal". We have therefore successfully identified extreme-but-normal functions that have arisen due to the fact that we have observed multiple functions.

Our original example in Fig. 2(b), which included two additional functions shown by dark lines, may now be revisited in light of our extreme function distributions. Setting $m = 100$ and assigning probabilities to each function in the example using the resulting extreme function distribution $1 - F^e_n(r)$ gives the result shown in Fig. 4(b). Here it may be seen that the ten sample functions randomly generated from the posterior GP (shown by yellow lines) are now assigned probabilities $1 - F^e_n(r) \approx 1$, indicating their obvious "normality" with respect to the GP model $\mathcal{M}$. However, one of the two functions shown by dark lines has been assigned $1 - F^e_n(r) = 0.16$, which would cause a hypothesis test based on, for example, $1 - F^e_n(r) < \alpha = 0.05$ to accept H_0 and classify the function as being "normal". This classification is correct in this instance, because the function was generated from the GP model $\mathcal{M}$, and is the most extreme function (i.e., that with the lowest likelihood z) from a set of $m = 100$ sample functions generated from $p(\mathbf{f}^*|\mathbf{x}, \mathbf{f}, \mathbf{x}^*)$.

The other function shown by a dark line is assigned a very low probability, $1 - F^e_n(r) < 10^{-4}$, and is therefore classified "abnormal" by a hypothesis test based on $1 - F^e_n(r) < \alpha = 0.05$. This is correct in this instance, because it was drawn from a GP with $\sigma_y = 0.25$, whereas the other functions were drawn from an otherwise-identical GP with $\sigma_y = 1$.

We have demonstrated that the proposed extension of extreme value theory to the problem considered in this paper can correctly separate functions that are extreme-but-normal (i.e., generated by a model of normality, but with a low probability) from those that are actually abnormal (i.e., those not generated by the model).

VII. Illustration With Patient Vital-Sign Data

We now illustrate the use of the method with data representing patient vital-sign trajectories following upper gastro-intestinal cancer surgery, where the patients are subsequently nursed in a "step-down" recovery ward. This dataset comprises 154 examples of "normal" patient timeseries, in which the patients were discharged home after a variable length-of-stay in the hospital, and 17 examples of "abnormal" patient timeseries, in which the patients either died or were admitted under emergency conditions to the intensive care unit (ICU). Patients in this latter category are associated with highly increased risks of mortality and morbidity, and so the automatic determination of physiological deterioration from timeseries of acquired data is an important task – made more so by the fact that vital-sign observations are taken every four hours at best [29].

Our study [30] was undertaken at the Cancer Hospital, within the Oxford University Hospitals NHS Trust, and was granted ethical approval by the local ethics committee.

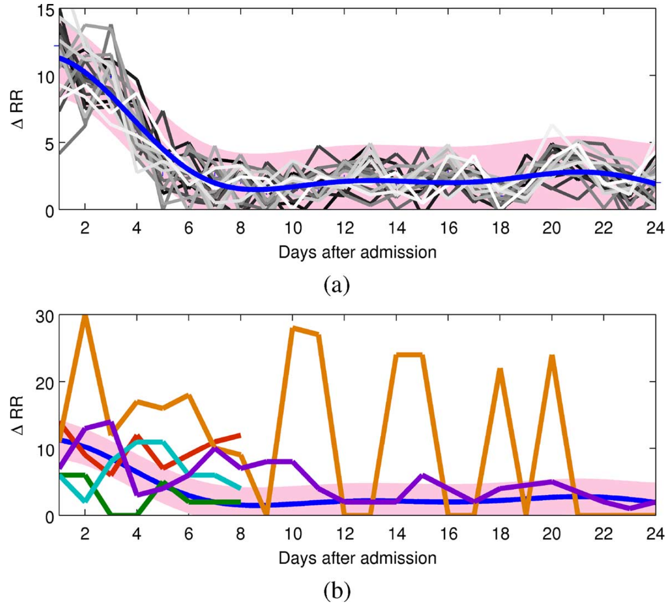
A. Data

Fig. 5(a) shows a subset of the normal data, illustrating 15 time-series of variabilities in respiratory rate (RR) from the "most normal" patients[3], in which the differences, over a 24-hour period, between the maximum and minimum RR (as observed by the nurses) are shown for the first 24 days after admission to the recovery ward. It may be seen that these timeseries typically take high values immediately after discharge from surgery, which then reduce as the patient recovers on the ward.

For comparison, timeseries from a selection of five patients from the abnormal dataset are shown in Fig. 5(b). It may be seen that these example abnormal timeseries can take highly extreme values (such as that shown in orange, which reaches $\Delta RR = 30$); however, many examples of abnormality (such as the four shown in purple, red, light blue, and green) occupy the same range of values on the vertical axis as do the normal patients.

The dataset of 154 normal patients has a median length-of-stay of 9 days (IQR 5 days), while the dataset of 17 abnormal patients has a median length-of-stay of 5 days (IQR 4 days). There is significant overlap between the length-of-stays in each set, and both sets contain timeseries for all lengths-of-stay up to 24 days. The majority of the datasets exhibited incomplete data within their record.

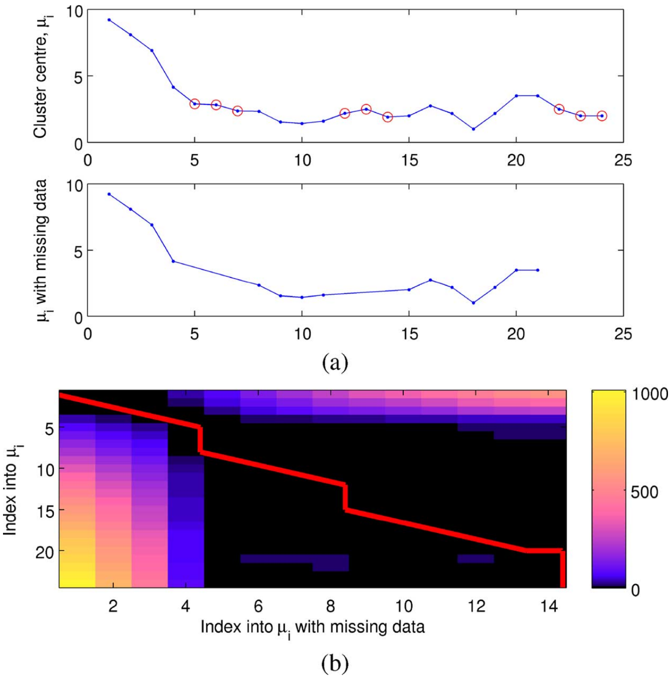
[3] defined to be those closest to the median length-of-stay for normal patients



(a)

(b)

Fig. 5. Normal and abnormal patient data shown in (a) and (b), respectively, with the EFT GP model $\mathcal{M}$ shown by its mean function (in blue) and 95% confidence region (shaded pink). (a) "Normal" vital-sign trajectories. (b) ICU readmission trajectories.



(a)

(b)

Fig. 6. (a) Normal data, corresponding to the centroid of the normal data, in the upper subplot, with the same timeseries having data removed shown in the lower subplot. Missing data are circled in red in the upper subplot. (b) DTW match of the original timeseries to the version containing missing data, where color shows DTW distance.

B. Method

We compared five methods for performing novelty detection: (i) the popular one-class SVM formulation of [3]; (ii) the recently-proposed one-class GP classifier of [10]; (iii) a k-nearest neighbors (k-NN) approach using the Euclidean distance metric; (iv) a k-NN approach using a dynamic time-warping (DTW) distance [31]; and (v) our GP-based extreme function theory. We note that (i) and (ii) are point-wise approaches to novelty detection, and the Gaussian kernel was used as the default choice in each; the former is distance-based, in the RKHS of its kernel, while the latter is probabilistic. Methods (iii) and (iv) are distance-based methods, often used as benchmarks in the anomaly detection and novelty detection literature [20], [32], [33]. We denote the methods (i)-(v) from above as OC-SVM, OC-GP, kNN, kNN-DTW, and EFT, respectively.

TABLE I
MEAN (AND 1 S.D.) CLASSIFICATION ERRORS FOR PATIENT VITAL-SIGN DATA OVER $N = 50$ EXPERIMENTS, SHOWING FALSE-POSITIVE RATE ($\mathrm{FPR} = \mathrm{FP}/38$) AND FALSE-NEGATIVE RATE ($\mathrm{FNR} = \mathrm{FN}/9$).

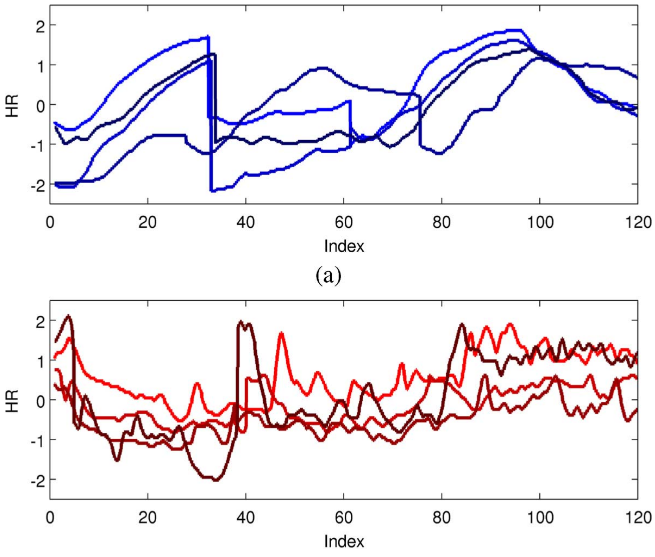
	OC-SVM	OC-GP	kNN	kNN-DTW	EFT
FPR	0.21 (0.08)	0.21 (0.12)	0.28 (0.12)	0.15 (0.07)	0.16 (0.05)
FNR	0.23 (0.07)	0.22 (0.10)	0.31 (0.09)	0.19 (0.08)	0.13 (0.08)

Ten-fold cross-validation was used to set model parameters for each method, using a training set that comprised 75% of the available normal data (116 of 154 examples, selected randomly) and 50% of the available abnormal data (8 of the 17 examples, selected randomly). Model parameters were selected in each method to minimize the classification error[4]. The remaining 38 normal examples and 9 abnormal examples were "held out" and used as test data. The entire experiment (random selection of training data, followed by ten-fold cross-validation to set model parameters) was repeated $N = 50$ times for each classifier. Each method was trained using feature vectors comprising data up to 24 days. For those methods that require fully-specified input vectors (OC-SVM, OC-GP, kNN), the data were padded with zeros where data were missing, after zero-mean, unit-variance normalization was applied (such that the missing data appear "normal", at the mean of the training set). The kNN-DTW and EFT methods allow non-probabilistic and probabilistic handling of missing data, using DTW and marginalization, respectively.

C. Results

Results from each of the $N = 50 \mathrm{experiments}$, reported on the "held out" test data in each case, are shown in Table I. It may be seen that the best-performing techniques are those that can adequately cope with the missing data, the kNN-DTW and the proposed EFT-based method. While both of the latter exhibit lower FPR than the other methods, the proposed method has greater success at classifying abnormal data, with a lower FNR across all experiments. An example of DTW is shown in Fig. 6(a), which shows one of the normal examples with data removed for the purposes of illustration. Fig. 6(b) demonstrates that the DTW method accurately aligns the original copy of the normal timeseries with that containing missing data. However, the results shown in Table I indicate that the principled treatment of missing data from the proposed method (achieved via straightforward marginalization of the Gaussian process [21]) results in better sensitivity to "abnormal" data, with a lower FNR.

[4] Classification error is defined to be false negatives (FN) + false positive (FP). The former is an erroneous classification of an abnormal example as being "normal", while the latter is erroneous classification of a normal example as being "abnormal".



(a)

(b)

Fig. 7. Randomly selected examples from the "normal" and "abnormal" classes in (a) and (b), respectively.

VIII. Illustration With Benchmark Dataset

The PAMAP2 benchmark dataset [34], [35] was used from the UCI Machine Learning Repository [36], which comprises activity data from heart-rate and inertial measurement units, connected to subjects who are asked to perform a protocol of 12 activities: lying, sitting, standing, ironing clothes, using a vacuum cleaner, walking normally, "Nordic" walking, running, cycling, ascending stairs, descending stairs, and using a skipping rope. Data were acquired for a total of 27,000 seconds, resulting in over 3.8×10^6 labeled data.

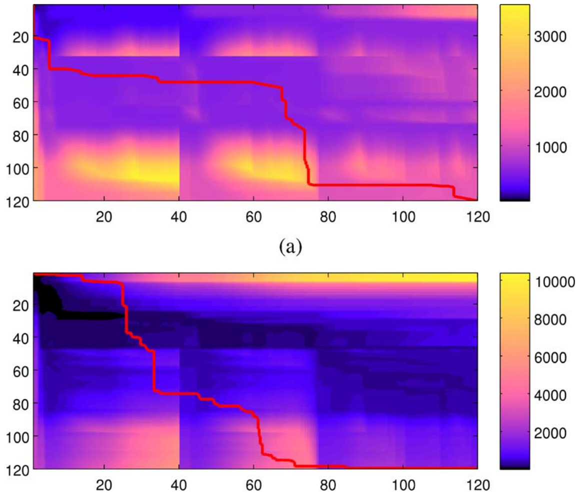
This dataset is multivariate, where we have investigated timeseries of the following features: heart-rate (at approximately 9 Hz) and average spectral power (per 5 s window of data) for each of the three axes of a patient-worn accelerometer (at approximately 100 Hz), following [35].

The methodology from Section VII was repeated, where the "normal" class was taken to comprise all timeseries examples of ascending and descending stairs, and where all other activities were taken as being "abnormal" for the purposes of this investigation. Fig. 7 shows randomly-selected examples of both the "normal" and "abnormal" classes, where it may be seen that the former (ascending and descending stairs) exhibits particular dynamics that a novelty detector could learn, whereas the latter seems more inconsistent (as expected, given that these are random examples of other types of activity).

While this benchmark dataset does not suffer from the incompleteness of data evident in the dataset considered in Section VII, the length of each timeseries can vary. The multivariate nature of the dataset is straightforwardly used by methods (i)-(iv). For the proposed EFT-based method, we adopt the usual GP approach [21] of providing the kernel function $k(\mathbf{x}, \mathbf{x})$ with multivariate inputs.

TABLE II
MEAN (AND 1 S.D.) CLASSIFICATION ERRORS FOR UCI BENCHMARK DATA OVER $N = 50$ EXPERIMENTS.

	OC-SVM	OC-GP	kNN	kNN-DTW	EFT
FPR	0.18 (0.06)	0.22 (0.07)	0.31 (0.10)	0.18 (0.08)	0.18 (0.07)
FNR	0.27 (0.09)	0.27 (0.10)	0.33 (0.10)	0.21 (0.11)	0.19 (0.06)



(a)

(b)

Fig. 8. Examples of kNN-DTW resulting in misclassifications in which (a) two "normal" examples are matched only for the latter half of the timeseries; (b) "abnormal" and "normal" timeseries are closely matched due to the warping effect introduced by DTW.

Table II presents the results of $N = 50$ experiments, in which it may be seen that most methods have performed less well using this benchmark dataset compared with the patient-based analysis in Section VII. The OC-GP and kNN methods are the least able to separate "normal" from "abnormal" classes in this instance, while the OC-SVM has achieved a low FPR. The kNN-DTW method performs similarly to the proposed EFT-based method, although the latter achieves a lower FNR (i.e., it is more sensitive to "abnormal" data), and has generally less variable performance (standard deviation of FNR 0.06 compared with 0.11 for the kNN-DTW).

Again, the DTW procedure can occasionally perform undesirable warping, such that "normal" data are sometimes associated with larger DTW distances to "normality" than some "abnormal" timeseries. This effect is shown in Fig. 8, in which an "abnormal" timeseries in (b) matches the centroid of "normal" data more closely than the example "normal" timeseries shown in (a), with a DTW distance for the "abnormal" case in (b) being $0.67 \times$ the DTW distance for the "normal" case in (a). Again, the GP-based method straightforwardly marginalises over shorter timeseries to perform a classification using only the relevant input domain of the model $\mathcal{M}$.

IX. Discussion

We have extended extreme value theory such that we may take a function-wise approach to novelty detection, following the FDA method in which functions are initially represented by timeseries of discrete observations. While much of previous work in novelty detection has concentrated on pointwise approaches (which are by far the most commonly described in the literature), these reach the limit of their usefulness in the assessment of timeseries.

GPs offer a natural probabilistic framework in which to define distributions over a function space, and we have used the GP regression case to illustrate our method due to the convenience of working with the Gaussian distribution, which has allowed us to find extreme function distributions in closed-form. We note that while the df of densities $G_n(z)$ tends asymptotically towards a parametric distribution (the Weibull, in this case), the whole approach is non-parametric, due to the GP distribution over function space [21].

Results reported in this paper suggest that the proposed method provides slightly better performance than distance-based methods employing dynamic time-warping, while offering the advantages of a principled probabilistic inference framework. While DTW copes well with small quantities of missing data, or with timeseries that differ in length by a small amount, it is less robust when coping with noisy, incomplete timeseries that are often encountered in practice, such as in physiological patient monitoring. By contrast, the proposed method copes well with missing data, using straightforward marginalization of the GP, afforded by its *consistency* property, in which any subset of the rvs over which the GP is defined also has a joint distribution which is multivariate Gaussian.

A natural extension for this work would be to non-Gaussian processes, in which the distribution over functions is not constrained to be Gaussian; e.g., a Beta process or a generalized Pareto process. However, the GP framework is sufficiently flexible to allow a wide-range of functional forms, as illustrated throughout this paper.

There is further scope for estimation of the Weibull parameters within the Bayesian GP framework, which would require an approach using approximate inference, such as via sampling or deterministic (variational) methods.

Appendix
$\log G_n(w)$ in Closed Form

From (10), we apply iterative integration-by-parts and use $w = \log z$ to obtain cases for n even and odd, respectively,

$$G_{2p}(w) = \exp(w) \sum_{k=1}^{p-1} A_{2p}^{k} \Big[-2\log[C_n \exp(w)]\Big]^{p-k-1} \tag{15}$$

$$\begin{aligned} G_{2p+1}(w) = \exp(w) \sum_{k=1}^{p-1} A_{2p+1}^{k} \Big[-2\log[C_n \exp(w)\Big]^{p-k-1/2} \\ + \operatorname{erfc}\left(\sqrt{-\log[C_n \exp(w)]}\right) \end{aligned} \tag{16}$$

where $\operatorname{erfc}(\cdot)$ is the complementary Gaussian error function. These reduce to (11), in which

$$B = -2\log C_n = -n\log(2\pi) - 2\sum_{i-1}^{n} \log \mathbf{L}_{i,i}$$

$$H_1(w) = \sum_{k=0}^{p-1} A_{2p}^{k} (B - 2w)^{p-k-1}, \tag{17}$$

$$H_2(w) = \sum_{k=1}^{p-1} A_{2p+1}^{k} (B - 2w)^{p-k-1/2} \tag{18}$$

$$H_3(w) = \operatorname{erfc}\left(\sqrt{\frac{B}{2} - w}\right) \tag{19}$$

$$A_{2p}^{k} = \Omega_{2p} \frac{2^k (p-1)!}{(p-1-k)!} \prod_{i=1}^{n} \mathbf{L}_{i,i}, \tag{20}$$

$$A_{2p+1}^{k} = \Omega_{2p+1} \frac{(2p-1)!(p-k)!}{2^{k-1}(p-1)!(2p-2k)!} \prod_{i=1}^{n} \mathbf{L}_{i,i} \tag{21}$$

References

[1] L. Tarassenko, D. Clifton, P. Bannister, S. King, and D. King, "Novelty detection," *Encyclopaedia of Structural Health Monitoring*, pp. 653–675, 2009.

[2] D. Tax and R. Duin, "Support vector domain description," *Pattern Recognit. Lett.*, vol. 20, pp. 1191–1199, 1999.

[3] B. Schölkopf, J. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Comput.*, vol. 13, no. 7, pp. 1443–1471, 2001.

[4] N. Hughes, L. Tarassenko, and S. Roberts, "Markov models for automated ECG interval analysis," in *Advances in Neural Information Processing Systems*, S. Thrun, L. Saul, and B. Schoelkopf, Eds. Cambridge, MA: MIT Press, 2004, vol. 16, pp. 611–618.

[5] H. Lee and S. Roberts, "On-line novelty detection using the Kalman filter and extreme value theory," in *Proc. Int. Conf. Pattern Recognit.*, 2008, pp. 1–4.

[6] J. Quinn, C. 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.

[7] H. Kim and J. Lee, J. Wang, Z. Yi, J. Zurada, B.-L. Lu, and H. Yin, Eds., "Pseudo-density estimation for clustering with Gaussian processes," in *Proc. Advances in Neural Networks – ISNN 2006*, Berlin, Germany, 2006, vol. 3971, Lecture Notes in Computer Science, pp. 1238–1243.

[8] R. Adams, I. Murray, and D. Mackay, "The Gaussian process density sampler," in *Advances in Neural Information Processing Systems*, D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, Eds. Cambridge, MA: MIT Press, 2009, vol. 21, pp. 9–16.

[9] Y. Gao and Y. Li, R. Kimmel, R. Klette, and A. Sugimoto, Eds., "Improving Gaussian process classification with outlier detection, with applications in image classification," in *Proc. Comput. Vis.—ACCV '10*, Berlin, Germany, 2011, vol. 6495, Lecture Notes in Computer Science, pp. 153–164.

[10] M. Kemmler, E. Rodner, and J. Denzler, R. Kimmel, R. Klette, and A. Sugimoto, Eds., "One-class classification with Gaussian processes," in *Proc. Comput. Vis.—ACCV '10*, Berlin, 2011, vol. 6493, Lecture Notes in Computer Science, pp. 489–500.

[11] R. Garnett, M. A. Osborne, S. Reece, A. Rogers, and S. J. Roberts, "Sequential Bayesian prediction in the presence of changepoints and faults," *Comput. J.*, vol. 53, no. 9, pp. 1430–1446, 2010.

[12] L. Tarassenko, P. Hayton, N. Cerneaz, and M. Brady, "Novelty detection for the identification of masses in mammograms," in *Proc. 4th IEE Int. Conf. Artif. Neural Netw.*, Perth, Australia, 1995, vol. 4, pp. 442–447.

[13] J. Ramsay and B. Silverman, *Functional Data Analysis*. New York: Wiley, 2005.

[14] G. M. James and C. A. Sugar, "Clustering for sparsely sampled functional data," *J. Amer. Statist. Assoc.*, vol. 98, pp. 397–408, 2003.

[15] J. Rice, "Functional and longitudinal data analysis: Perspectives on smoothing," *Statist. Sinica*, vol. 14, pp. 631–647, 2004.

[16] H.-G. Muller, "Functional modelling and classification of longitudinal data," *Scandinavian J. Statist.*, vol. 32, no. 2, pp. 223–240, 2005.

[17] F. Yao, H.-G. Muller, and J. Wang, "Functional linear regression analysis for longitudinal data," *Ann. Statist.*, vol. 33, no. 6, pp. 2873–2903, 2004.

[18] J. Shi and T. Choi, *Gaussian Process Regression Analysis for Functional Data*. London, U.K.: Chapman & Hall, 2011.

[19] P. Hall and I. van Keilegom, "Two-sample tests in functional data analysis starting from discrete data," *Statist. Sinica*, vol. 17, pp. 1511–1531, 2007.

[20] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surveys*, vol. 41, no. 3, 2009.

[21] C. Rasmussen and C. Williams, *Gaussian Processes for Machine Learning*. Cambridge, MA: MIT Press, 2006.

[22] J. Shi and B. Wang, "Curve prediction and clustering with mixtures of Gaussian process functional regression models," *Statist. Comput.*, vol. 18, pp. 267–283, 2008.

[23] S. J. Roberts, "Extreme value statistics for novelty detection in biomedical signal processing," *IEE Proc. Sci., Technol. Measur.*, vol. 47, no. 6, pp. 363–367, 2000.

[24] D. Clifton, S. Hugueny, and L. Tarassenko, "Novelty detection with multivariate extreme value statistics," *J. Signal Process. Syst.*, vol. 65, pp. 371–389, 2011.

[25] P. Embrechts, C. Kluppelberg, and T. Mikosch*, Modelling Extremal Events*, 4th ed. Berlin, Germany: Springer-Verlag, 2003.

[26] L. de Haan and A. Ferreira*, Extreme Value Theory*. Berlin, Germany: Springer-Verlag, 2006.

[27] S. Hugueny, D. Clifton, and L. Tarassenko, "Probabilistic patient monitoring with multivariate, multimodal extreme value theory," *Communicat. Comput. Sci.*, vol. 127, pp. 199–211, 2011.

[28] R. A. Fisher and L. H. C. Tippett, "Limiting forms of the frequency distributions of the largest or smallest members of a sample," in *Proc. Cambridge Philosoph. Soc.*, 1928, vol. 24.

[29] L. Tarassenko, D. Clifton, M. Pinsky, M. Hravnak, J. Woods, and P. Watkinson, "Centile-based early warning scores derived from statistical distributions of vital signs," *Resuscitat.*, vol. 82, no. 8, pp. 1013–1018, 2011.

[30] L. Clifton, D. Clifton, M. Pimentel, P. Watkinson, and L. Tarassenko, "Gaussian processes for personalised e-health monitoring with wearable sensors," *IEEE Trans. Biomed. Eng.*, vol. 60, no. 1, pp. 193–197, Jan. 2012.

[31] K. Buza, "Fusion methods for time-series classification," Ph.D. dissertation, Univ. of Hildesheim, Hildesheim, Germany, 2011.

[32] M. Markou and S. Singh, "Novelty detection: A review – Part 1: Statistical approaches," *Signal Process.*, vol. 83, no. 12, pp. 2481–2497, 2003.

[33] M. Markou and S. Singh, "Novelty detection: A review – Part 2: Neural network based approaches," *Signal Process.*, vol. 83, no. 12, pp. 2499–2521, 2003.

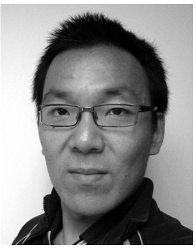
[34] A. Reiss and D. Stricker, "Introducing a new benchmarked dataset for activity monitoring," in *Proc. 16th IEEE Int. Symp. Wearable Comput.*, Newcastle, U.K., 2012.

[35] A. Reiss and D. Stricker, "Creating and benchmarking a new dataset for physical activity monitoring," in *Proc. 5th Workshop Affect and Behavior Related Assistance*, Crete, Greece, 2012.

[36] A. Frank and A. Asuncion, UCI Machine Learning Repository 2010 [Online]. Available: http://archive.ics.uci.edu/ml

David A. Clifton is a Research Fellow at Mansfield College, Oxford and a College Lecturer at Balliol College, Oxford. He received an MEng degree in Engineering Mathematics from the University of Bristol and a DPhil degree in Information Engineering from the University of Oxford. His research interests are in statistical signal processing, particularly in biomedical informatics and other biomedical applications. His doctoral research led to patented methods for monitoring the jet engines of the Eurofighter Typhoon, the Airbus A380, and the Boeing 787 Dreamliner, and he has won the nPower Science, Engineering and Technology (SET) award at the UK Houses of Parliament and the J.A. Lodge award for biomedical engineering by the IET.

Lei Clifton is a post-doctoral research assistant at the Institute of Biomedical Engineering, in the Department of Engineering Science at the University of Oxford. She received BSc and MSc degrees in Electrical Engineering from the Beijing Institute of Technology, China, and a PhD degree in Information Engineering from UMIST. Her research interests include the use of statistical machine learning for health informatics and physiological monitoring.



Samuel Hugueny received an MSc degree in Computer Engineering from the Ecole National Superieure de Techniques Avancees, Paris (ENSTA-Paritech), and is currently studying a DPhil degree at the Institute of Biomedical Engineering in the Department of Engineering Science, University of Oxford. His research interests include medical imaging, signal processing, novelty detection, and patient monitoring.

David Wong is a BRC post-doctoral research assistant at the Institute of Biomedical Engineering in Oxford. He graduated from the University of Oxford with MEng and DPhil degrees, for the latter of which his research focussed on developing systems to detect deterioration in hospital patient vital-signs using machine learning signal processing techniques. His recent interests include determining vital-sign trends using time-series analysis. David is currently involved in a number of clinical studies to assess the effect of computer-assisted systems in increasing patient safety, and reducing nursing workload during routine hospital observations.

Lionel Tarassenko received the B.A. degree in engineering science in 1978 and the Ph.D. degree in medical engineering in 1985, both from the University of Oxford. He has held the Chair in Electrical Engineering at Oxford University since October 1997. He was elected to a Fellowship of the IEE in 1996, when he was also awarded the IEE Mather Premium for his work on neural networks, and to a Fellowship of the Royal Academy of Engineering (RAE) in 2000. Prof. Tarassenko was awarded the Silver Medal of the Royal Academy of Engineering for his contribution to British engineering in 2006, and was awarded the CBE for services to engineering in 2011.