Knowl Inf Syst (2014) 38:717–741
DOI 10.1007/s10115-013-0685-z

REGULAR PAPER

Maritime abnormality detection using Gaussian processes

Mark Smith · Steven Reece · Stephen Roberts · Ioannis Psorakis · Iead Rezek

Received: 1 February 2013 / Revised: 16 August 2013 / Accepted: 17 August 2013 /
Published online: 30 August 2013

Abstract Novelty, or abnormality, detection aims to identify patterns within data streams that do not conform to expected behaviour. This paper introduces novelty detection techniques using a combination of Gaussian processes, extreme value theory and divergence measurement to identify anomalous behaviour in both streaming and batch data. The approach is tested on both synthetic and real data, showing itself to be effective in our primary application of maritime vessel track analysis.

Keywords Gaussian processes · Extreme value theory · Novelty detection · Hellinger distance · Nonnegative matrix factorisation · Maritime traffic · Outlier detection

1 Introduction

The global picture of maritime traffic is large and complex, consisting of dense volumes of (mostly legal) ship traffic. Techniques that identify illegal traffic could help to reduce the impact from smuggling, terrorism, illegal fishing, etc. In the past, surveillance of such traffic has suffered due to a lack of data. However, since the advent of electronic tracking, the amount of available data has grown beyond an analyst's ability to process without some form of automation. One part of the analyst's workload lies in the detection of anomalous behaviour in otherwise normal appearing tracks. Our goal is to detect anomalous vessels using an automated approach. In this paper, we exploit techniques from the field of anomaly detection, particularly *extreme value theory*, to identify potential deviations from normal behaviour.

M. Smith (✉)
ISSG, Babcock Marine and Technology Division, Devonport Royal Dockyard, Plymouth, PL1 4SG, UK
e-mail: marks@robots.ox.ac.uk; mark.x.smith@babcockinternational.com

S. Reece · S. Roberts · I. Psorakis
Department of Engineering Science, University of Oxford, Oxford, UK

I. Rezek
Schlumberger Research, Cambridge, UK

The latter is modelled using a nonparametric Bayesian approach, namely sequential *Gaussian process* regression. This is also extended by applying techniques from the field of divergence measurement namely the *Hellinger distance* to construct an adjacency matrix from which clusters or communities of vessel types can be formed and anomalous tracks identified.

An anomaly has many different interpretations depending on the context in which it is used, i.e. it may refer to a data point arising from a different distribution, measurement error, population variability or execution error. However, fundamentally an anomaly is a data point that stands out in contrast to the other data points around it [5]. It is the task of anomaly detection to infer whether this data point deviates significantly given the intrinsic variability of the population of normal data. An effective technique should therefore be capable of recognising and modelling data points that occur due to such anomalous events and distinguish these from outliers associated with the tails of the reference distribution of nonanomalous data.

This paper applies *extreme value statistics* to identify likely anomalous samples that are extreme values. The probability distribution governing these extreme values is sequentially updated to enable context-sensitive decisions. This is achieved by means of linking this distribution to a sequential *Gaussian process* model, which regresses the vessel's track and forecasts a distribution over future data. Whilst application in this manner can identify individual anomalous points, it is unable to determine whether the entire track can be considered an anomaly. We address this issue by undertaking unsupervised clustering of vessel tracks to identify clusters of a given vessel class and detecting abnormalities from the track cluster. This is achieved through bivariate *Gaussian process* regression modelling of vessel track latitudinal and longitudinal data, and applying the inverse *Hellinger distance* in order to discover an adjacency matrix. An unsupervised approach to clustering based on community detection using *nonnegative matrix factorisation* is then used to identify communities of similar tracks within the adjacency matrix. Anomalies can then be identified by noting a discrepancy between the class of vessel and vessels assignment to a labelled community of vessels.

This paper begins by providing a brief overview of existing approaches to maritime situational awareness, followed by a description of our methods for identifying anomalous tracks. The methods section is broken down into two subsections. Firstly, the identification of anomalous points is considered; secondly, we consider the detection of anomalous tracks. In each instance, the techniques are applied to both synthetic and real vessel tracks extracted from their GPS coordinates. This is to provide both insights into the workings of the approach and provide an empirical validation of the technique.

2 Current techniques

The field of marine anomaly detection has employed a variety of methods including neural networks in [24], Bayesian networks in [14], support vector machines in [12], GPs in [27] and Kalman filters in [8]. Common between all methods are two main tasks: creating a model of normality (free from the presence of anomalies) and using a metric from this model to (allowing for some quantifiable variability) identify anomalous points. These tasks are inherent within the two main uses for the detection of anomalies: *accommodation* and *discordancy*. Accommodation is the task in which the goal is to create a model of normality that does not include anomalous observations. Discordancy tests provide a metric indicator of a point being an anomaly. Both have different aims but within each is some model of normality and a measure of deviation.

Assessing the performance of these different methods is a difficult task as there exists no established benchmarks of what are considered to be marine anomalies, therefore hindering comparison [9]. This implies that a data set can be considered under a variety of contexts, leading to different types of anomalies being identified on the same input data. For example, a sequential time series analysis of a vessel track, where the previous location and the dynamics of the vessel are considered, could highlight sudden changes in vessel dynamics, possibly indicative of evasive manoeuvring. Such a sequential time series model has the advantage that it can be used in online analysis, but it may miss patterns when the entire track is considered as a whole [14]. Other indications of anomalous behaviour within the data could be deviations from a standard route, unexpected port arrival, close approach and zone entry [7]. Even when a particular marine anomaly has been selected for identification, the data need to be considered in the context of external factors, for example the class of vessel, time of day and tidal status. Since these may have to be taken into account when analysing, the data as the form of the anomaly may vary.

Critical to marine anomaly detection is an interpretation of the data that allow the salient features of the desired anomaly to be identified [10]. Further, models for different kinds of anomalies may need to be combined or considered to increase the certainty of an anomaly being detected. For example, a model identifying anomalous vessel speeds could be combined with a model of anomalous zone entries (anomalous spatial locations). Vessels identified as demonstrating anomalous speeds may be pleasure craft in a known unrestricted speed location, giving false positives if simply considered only on the basis of the speed model. Conversely, a vessel entering a port at high speed may be highly anomalous behaviour.

3 Identifying anomalous points

The approach we detail in this paper develops methods for the two main underlying tasks within anomaly detection: modelling normality and subsequently using a cost function to identify points as anomalous. In this section, we construct a model of normality using Gaussian processes (GPs), which allows us to capture the dynamics of vessels in a nonanomalous data set without prescribing a particular parametric form (such as is required for Markov state models, for example). The GP provides a sequentially updated posterior distribution over unseen data, which we link to an extreme value distribution to provide a robust and adaptive metric for anomaly detection.

3.1 Gaussian processes

In order to model the vessel track, we use a Gaussian process, providing a mechanism to continuously predict vessel locations at any future time point, *including* a measure of uncertainty about the vessel location. The GP is a stochastic process [21] that expresses the dependent variable, $\boldsymbol{y}$, in terms of an independent variable (called the input variable) $\boldsymbol{x}$, via a function $f(\boldsymbol{x})$. This function we can see as a draw from a probability distribution over functions,

$$\boldsymbol{y} = f(\boldsymbol{x}) \sim \mathrm{GP}(m(\boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x})),$$

where $m(\boldsymbol{x})$ describes the mean function of the distribution at $\boldsymbol{x}$ and k is a covariance function that describes the information coupling between two values of the independent variable as a function of the distance of their respective inputs. The matrix resulting from the application of the covariance function to a vector of input variables can then be expressed as $\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x})$.

The covariance function thus encodes our prior beliefs and assumptions about the function that we wish to model [21]. Valid covariance functions can take a variety of forms, which we quantify empirically in Sect. 4. Denoting $r = |x_p - x_q|$ as the (Euclidean) distance between two independent values, x_p and x_q, we consider three covariance functions: the squared exponential

$$k_{\mathrm{SE}}(r) = \sigma_0^2 \exp\left(-\frac{r^2}{2\lambda^2}\right), \tag{1}$$

the Matérn $\frac{3}{2}$

$$k_{\frac{3}{2}}(r) = \sigma_0^2 \left(1 + \frac{\sqrt{3r}}{\lambda}\right) \exp\left(-\frac{\sqrt{3r}}{\lambda}\right), \tag{2}$$

and the Matérn $\frac{1}{2}$ covariance function

$$k_{\frac{1}{2}}(r) = \sigma_0^2 \exp\left(-\frac{r}{\lambda}\right). \tag{3}$$

The above selection was driven by prior knowledge about typical vessel trajectories, for example periodic kernels [21] were excluded from the set of covariance functions as the data in our feature space are not recurrent. Hence, we opted for covariance functions capable of reflecting the physical properties of shipping vessels, such as smoothness and differentiability.

We also assume that observations of vessel tracks are corrupted by additive i.i.d Gaussian noise with variance component ε^2. Thus, the full covariance function for the observations is given as

$$v(x_p, x_q) = k(x_p, x_q) + \varepsilon^2 \delta(|x_p - x_q|),$$

where δ is the Kronecker delta, which is one if $p = q$ and zero otherwise. The matrix resulting from the application of the covariance function to a vector of input variables can then be expressed as $\boldsymbol{V}(\boldsymbol{x}, \boldsymbol{x})$.

The hyperparameters σ_0, λ and ε are, respectively, the amplitude, output and noise scale. They encode the characteristics of the track and so depend on the dynamics of the vessel. A vessel undertaking manoeuvring will not exhibit the same smooth track characteristics as one exhibiting regular motion. Thus, the hyperparameters need to be learnt from an anomaly-free training data set that consists of n observations, $D = \{(x_i, y_i) | i = 1, \ldots, n\}$. The x_i and y_i points represent the independent and dependent variable values, respectively.

The nature of the Gaussian process is such that, conditional on the observed data, predictions can be made about the function values, $f(x_*)$ at any location x_*. The distribution of these values at point x_* is Gaussian

$$f_* | x_*, \boldsymbol{x}, \boldsymbol{y} \sim \mathcal{N}(\bar{f}_*, \mathrm{Var}[f_*]).$$

with the mean and variance given as

$$\begin{aligned} \bar{f}_* &= m(x_*) + \boldsymbol{K}(\boldsymbol{x}, x_*)^\top \boldsymbol{V}(\boldsymbol{x}, \boldsymbol{x})^{-1} (\boldsymbol{y} - m(\boldsymbol{x})), \\ \mathrm{Var}[f_*] &= \boldsymbol{K}(x_*, x_*) - \boldsymbol{K}(\boldsymbol{x}, x_*)^\top \boldsymbol{V}(\boldsymbol{x}, \boldsymbol{x})^{-1} \boldsymbol{K}(\boldsymbol{x}, x_*). \end{aligned} \tag{4}$$

3.2 Sequential Gaussian process updates

In many real-world problems, we receive data sequentially and the data set can grow to an arbitrarily large size. If we were to continue to update our beliefs in the light of new

observations, we could naively repeat the matrix inversion in Eq. 4 with every observation. This inversion is expensive as its computational complexity grows as $O(n^3)$ in the number of samples, i.e. the dimension of $\boldsymbol{V}$ above. Closer inspection however reveals that matrix $\boldsymbol{V}$ is changed only in the addition of a new row and column with increasing n. Hence, it is possible to reformulate the matrix inversion as a sequential Cholesky decomposition [16], and the entire $n \times n$ matrix inversion does not have to be recalculated each time n increases.

We decompose a matrix into the product of a lower triangular matrix, $\boldsymbol{R}$, and its conjugate transpose

$$\boldsymbol{V}(\boldsymbol{x}, \boldsymbol{x}) \triangleq \boldsymbol{R}(\boldsymbol{x}, \boldsymbol{x})^\top \boldsymbol{R}(\boldsymbol{x}, \boldsymbol{x}).$$

Based on this decomposition, the parameters of the predictive distribution are given as

$$\begin{aligned} \bar{f}_* &= m(x_*) + \boldsymbol{b}_{\boldsymbol{x},x_*}^\top \boldsymbol{a}_{\boldsymbol{x}} \boldsymbol{V}(x_*, x_*), \\ \mathrm{Var}[f_*] &= \boldsymbol{V}(x_*, x_*) - \boldsymbol{b}_{\boldsymbol{x},x_*}^\top \boldsymbol{b}_{\boldsymbol{x},x_*}, \end{aligned} \tag{5}$$

and where $\boldsymbol{a}$ and $\boldsymbol{b}$ are given as

$$\begin{aligned} \boldsymbol{a}_{\boldsymbol{x}} &\triangleq \boldsymbol{R}(\boldsymbol{x}, \boldsymbol{x})^\top \backslash (\boldsymbol{y} - m(\boldsymbol{x})), \\ \boldsymbol{b}_{\boldsymbol{x},x_*} &\triangleq \boldsymbol{R}(\boldsymbol{x}, \boldsymbol{x})^\top \backslash \boldsymbol{V}(\boldsymbol{x}, x_*). \end{aligned} \tag{6}$$

When we receive new data, the $\boldsymbol{V}$ matrix is changed only in the addition of some new rows and columns, i.e.

$$\boldsymbol{V}(\boldsymbol{x}, \boldsymbol{x}) = \begin{pmatrix} \boldsymbol{V}(\boldsymbol{x}_{1:n-1}, \boldsymbol{x}_{1:n-1}) & \boldsymbol{V}(\boldsymbol{x}_{1:n-1}, x_n) \\ \boldsymbol{V}(x_n, \boldsymbol{x}_{1:n-1}) & \boldsymbol{V}(x_n, x_n) \end{pmatrix}.$$

Consequently, the Cholesky decomposition can also be computed iteratively [16], via

$$\boldsymbol{R}(\boldsymbol{x}_{1:n}, \boldsymbol{x}_{1:n}) = \begin{pmatrix} \boldsymbol{R}(\boldsymbol{x}_{1:n-1}, \boldsymbol{x}_{1:n-1}) & \boldsymbol{S} \\ \boldsymbol{0} & \boldsymbol{U} \end{pmatrix},$$

where

$$\begin{aligned} \boldsymbol{S} &= \boldsymbol{R}(\boldsymbol{x}_{1:n-1}, \boldsymbol{x}_{1:n-1})^\top \backslash \boldsymbol{V}(\boldsymbol{x}_{1:n-1}, \boldsymbol{x}_{1:n}), \\ \boldsymbol{U} &= \mathrm{chol}(\boldsymbol{V}(x_n, x_n) - \boldsymbol{S}^\top \boldsymbol{S}). \end{aligned}$$

Using the iterative Cholesky update, the predictive distribution, Eq. 5, can also be expressed iteratively by computing the vector $\boldsymbol{a}$, in Eq. 6, via the following rule

$$\boldsymbol{a}_{1:n} = \begin{pmatrix} \boldsymbol{a}_{1:n-1} \\ \boldsymbol{U}^\top \backslash (y_n - m(x_n) - \boldsymbol{S}^\top \boldsymbol{a}_{1:n-1}) \end{pmatrix}.$$

The recursion avoids the computationally expensive matrix inversion, in Eq. 4, and allows the Cholesky factor to be expressed as an efficient update rule. Thus, the GP can be adapted to efficiently update its belief in the light of new evidence, as necessitated for online operation. We next consider how to further adapt the GP model in order to capture the correlation between time varying longitudinal and latitudinal data.

3.3 Multiple-output Gaussian processes

The GP model can be adapted to consider the correlation between multiple outputs through the parametrisation of the covariance matrix [16]. Thus, it is possible to discover the function that describes the relationship between latitude and longitude.

Through unconstrained optimisation, where the upper-triangular elements in the variance–covariance matrix are re-parameterised in such way that the resulting estimate must be positive semi-definite, we can discover the correlation between the selected output for the given inputs.

Whilst many techniques for the parametrisation of the covariance matrix exist [17], the spherical parametrisation benefits from the computational efficiency of the Cholesky decomposition allows the parametrisation to be expressed in terms of the variance and covariance between outputs. In this form, it is the Cholesky factor of the output $n \times n$ covariance, which is parameterised, $\boldsymbol{\Sigma} = \boldsymbol{L}^{\top}(\boldsymbol{\theta})\boldsymbol{L}(\boldsymbol{\theta})$. Thus, it can be ensured that the final covariance matrix remains positive semi-definite. By letting $\boldsymbol{L}_i$ denote the ith column of the Cholesky factorisation of $\boldsymbol{\Sigma}$ and $\boldsymbol{l}_i$ denotes the spherical coordinates of the first i elements of $\boldsymbol{L}_i$, where $i = 2, \ldots, n$. We can then express the parameterisations as

$$\begin{aligned}
[\boldsymbol{L}_1]_1 &= [\boldsymbol{l}_1]_1 \\
[\boldsymbol{L}_i]_1 &= [\boldsymbol{l}_i]_1 \cos([\boldsymbol{l}_i]_2) \\
[\boldsymbol{L}_i]_2 &= [\boldsymbol{l}_i]_1 \sin([\boldsymbol{l}_i]_2) \cos([\boldsymbol{l}_i]_3) \\
&\ldots \\
[\boldsymbol{L}_i]_{i-1} &= [\boldsymbol{l}_i]_1 \sin([\boldsymbol{l}_i]_2) \ldots \cos([\boldsymbol{l}_i]_i) \\
[\boldsymbol{L}_i]_i &= [\boldsymbol{l}_i]_1 \sin([\boldsymbol{l}_i]_2) \ldots \sin([\boldsymbol{l}_i]_i).
\end{aligned}$$

The product of the Cholesky factor parametrisation of the output spherical covariance between two outputs (latitude and longitude) can therefore be expressed as

$$\boldsymbol{\Sigma} = \begin{bmatrix} [\boldsymbol{l}_1]_1^2 & [\boldsymbol{l}_1]_1[\boldsymbol{l}_2]_1 \cos([\boldsymbol{l}_2]_2) \\ [\boldsymbol{l}_1]_1[\boldsymbol{l}_2]_1 \cos([\boldsymbol{l}_2]_2) & [\boldsymbol{l}_2]_2^2 \end{bmatrix}.$$

We now consider a principled means of determining whether a new data point should be updated into the model of normal system behaviour.

3.4 Extreme value theory

Extreme value theory has previously been used in novelty detection [11,25]. Extreme value theory provides a probability that a data point is anomalous. A threshold probability is chosen beyond which we can quantify a value as having not arisen from the underlying distribution. The theory itself focuses on the statistical behaviour of $M_n = \max\{X_1, \ldots, X_n\}$ where $X_1, \ldots, X_n$ is a sequence of independent random variables with a distribution function F. In theory, the distribution of M_n can be derived exactly for all values of n, i.e.

$$\begin{aligned}
Pr\{M_n \leq z\} &= Pr\{X_1 \leq z, \ldots, X_n \leq z\} \\
&= Pr\{X_1 \leq z\} \times \cdots \times Pr\{X_n \leq z\} \\
&= \{F(z)\}^n. \qquad (7)
\end{aligned}$$

In practice, the distribution function F is unknown and extreme value theory allows us to approximate this distribution. It states that the entire range of possible limit distributions for M_n is given by one of three types of cumulative distribution function, I, II and III, known as the Gumbel, Fréchet and Weibull, respectively, and given as

$$I : G(z) = \exp\left\{-\exp\left[-\left(\frac{z-\beta}{\alpha}\right)\right]\right\} \quad -\infty < z < \infty \qquad (8)$$

$$II : G(z) = \begin{cases} 0, & z \leq \beta, \\ \exp\left\{-\left(\frac{z-\beta}{\alpha}\right)^{-\xi}\right\}, & z > \beta \end{cases}$$

$$III : G(z) = \begin{cases} \exp\left\{-\left[-\left(\frac{z-\beta}{\alpha}\right)^{\xi}\right]\right\}, & z < \beta \\ 1, & z \geq \beta \end{cases}$$

Each family has a scale and location parameter, α and β, respectively. Additionally, the Fréchet and Weibull families have a shape parameter ξ [3]. Although we have three models to choose from, the underlying target distribution, F, in our case is assumed to be Gaussian, due to the modelling constraints imposed by the GP. The extreme value probability is then restricted to the analytical form of the Gumbel distribution.

Assuming that some "normal" data are identically and independently Gaussian-distributed, one can obtain the extreme quantiles by inverting equation 8

$$z_p = \beta - \alpha \log\left(-\log\left(p\right)\right).$$

The value of p acts as a novelty threshold, below which a test point is classified "abnormal." The parameters α and β require estimation and typically depend on the sample size n of the data set. As proposed in [25], we make use of decoupled estimators for α and β given, respectively, as

$$\alpha = \left(2\log\left(n\right)\right)^{-\frac{1}{2}} \tag{9}$$

$$\beta = \left(2\log\left(n\right)\right)^{\frac{1}{2}} - \frac{\log\left(\log\left(n\right) + \log\left(2\pi\right)\right)}{2\left(2\log\left(n\right)\right)^{\frac{1}{2}}}. \tag{10}$$

In the next subsection, we show how extreme value theory can be applied to anomaly detection in Gaussian processes (GPs).

3.5 Gaussian process–extreme value theory (GP–EVT)

In much of the existing work on novelty detection using extreme value methods, the work has focused on nonsequential conditions, or more precisely, on a fixed training data set. Whilst the extreme value will adequately account for the changes in our belief about the location of extreme events for a fixed sample size, the framework is rarely extended to account for dynamic changes in the underlying generating distribution *and* changes in the sample size.

In this work, we model the typical system dynamics using GP regression. At some arbitrary point in the future, say x_*, we can interrogate the GP and compute the predictive (Gaussian) distribution at that point, conditional on the trajectory's past samples. This predictive distribution, which now features a context (time)-dependent mean, $\bar{f}_*$, and variance, $\mathrm{Var}[f_*]$, allows rescaling of the extreme event quantile e,

$$e = \bar{f}_* + \sqrt{\mathrm{Var}[f_*]} z_p \tag{11}$$

and so reflects temporal changes in the statistics of the base distribution.

In order to estimate the number density of data points $n(x_*)$ at each x_* in Eq. 5, a Gaussian kernel smoother is applied using

$$n(x_*) = \sum_{i=1}^{m} \phi_h\left(x_*; x_i\right), \tag{12}$$

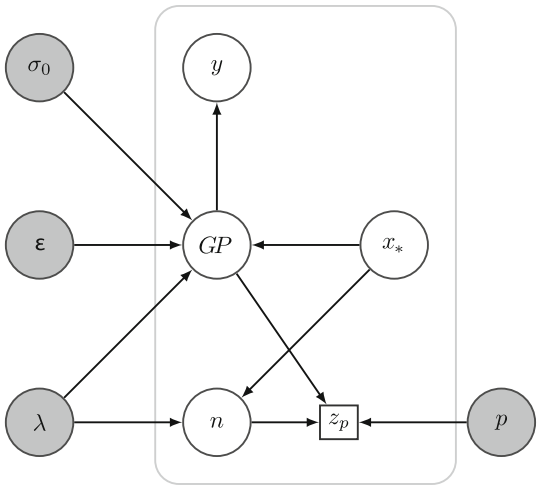

Fig. 1 A graphical model representation of the GP–EVT model. At the centre is the Gaussian process (GP) that models the track's dynamics. It has fixed (pre-inferred) hyperparameters shown as *grey nodes* to the *left*. Also shown is the estimate of the sample size, n. The extreme value percentile is a deterministic node, shown as a *square box*, which depends upon p, the novelty level, and sample size n

where m is the total number of previously observed nonanomalous observations and $\phi_h(x_*; x_i)$ is a (nonnormalised Gaussian) radial basis function

$$\phi_h(x_*; x_i) = \exp\left\{-\frac{|x_* - x_i|^2}{2h^2}\right\},$$

in which x_i is the most recent observation and $|\cdot|$ denotes the Euclidean distance. The kernel width h is set to be equal to twice the length scale λ in Eqs. 1, 2 or 3, depending on the choice of kernel used to model the vessel tracks. This coupling of λ to the GP regression model ensures that tracks with long correlation lengths and smaller sampling rates will feature the same sensitivity to outliers as tracks with short correlation lengths and high sampling rates. Also, the coupling ensures that the smoothing of the sampling processes does not come at a cost of an additional parameter that would require additional estimation.

With the expected number of observations obtained by Eq. 12, the extreme value distribution parameters can be updated in a timely fashion to reflect also the dynamics of the sampling process. Thus, the scaling (Eq. 9), and location (Eq. 10), parameters can be estimated, using the predicted number of data points, $n(x_*)$, contributing information at the location of interest x_* [15], by

$$\alpha(n_*) = (2\log(n(x_*)))^{-\frac{1}{2}} \tag{13}$$

and

$$\beta(n_*) = (2\log(n(x_*)))^{\frac{1}{2}} - \frac{\log(\log n(x_*)) + \log(2\pi)}{2(2\log(n(x_*)))^{\frac{1}{2}}} \tag{14}$$

for a fixed novelty detection threshold, p, which in this work is set to 0.95.

To reiterate, the GP provides a mechanism to predict the distribution of future mean values and to adjust the scaling of the extreme value quantile. Also, the kernel smoothing approach to the sampling process provides an estimate of the future sample size. Their combination is used for novelty detection. If the new data point value falls within a novelty measure of the predicted value, then the new data point is included in the model update. The key advantage

of using our approach is thus the incorporation of future uncertainty in both sampling and observation processes to provide the means for a more accurate novelty detection algorithm. The graphical model representation of the complete model is shown in Fig. 1.

4 Application

We demonstrate the efficacy of the approach presented in the previous section by application to both synthetic data and real data. We use synthetic data to illustrate some of the features of our method and provide a real-world example of its application to vessel tracks.

4.1 Synthetic data illustration

Synthetic data were generated from the Matérn $\frac{3}{2}$ kernel, Eq. 2, with parameters set to $\sigma_0 = 1$, $\lambda = 2$ and $\sigma = 0.01$. Anomalies were generated by offsetting randomly selected samples that were previously drawn from the GP by a fixed offset value, making the point anomalous with respect to surrounding points. The GP predictive distribution was calculated for 1,000 samples within the windowed region of track, the window ending at the time period for the new observed sample. A fixed kernel width was used in order to estimate n at each x_*.

GP extreme value theory was then applied by considering each new data point with respect to the previously learnt underlying function. If the new point falls within the predictive uncertainty of the next data point, it is included in the sequential update, otherwise it will be excluded. An example of such an update step is shown in Fig. 3. The new data point falls outside the EVT bound, Eq. 11, and so has been excluded. Notice that if a data point has not been observed for a period, the predictive uncertainty grows, allowing for the possibility of a dynamic change in the underlying base function and the new data point to be included in the update. In this manner, anomalous points within the data can be clearly identified whilst accommodating for the dynamics of the underlying function and irregular observations. An intuitive illustration of how the irregularity of observed data points affects the scaling of the extreme value distribution, and hence our novelty bounds, is given in Fig. 2.

4.2 Vessel track anomaly detection

The GP–EVT methodology was also applied to real-world vessel track data, consisting of a set of GPS coordinates. This is presented in an appropriate one-dimensional feature space parameterisation, providing a means of identifying changes in vessel dynamics.

Feature extraction In order to convert data to a sufficient feature space representation, we consider the first received data point as the beginning of the vessel track. We relate all subsequent data points to it by computing both the distance and time taken from this originating sample point. In order to take into account the approximated spherical geometry of the earth's surface, we calculate this distance by the application of the Haversine formula,

$$A = \cos\phi_s \cos\phi_f$$
$$\Delta\hat{\sigma} = \arctan\left(\sqrt{\sin^2\left(\frac{\Delta\phi}{2}\right) + A\sin^2\left(\frac{\Delta\lambda}{2}\right)}\right)$$

where ϕ_s and ϕ_f are the latitude of two points, and $\Delta\lambda$ and $\Delta\phi$ are their differences in longitude and latitude, respectively. This choice of feature space has the advantage of con-

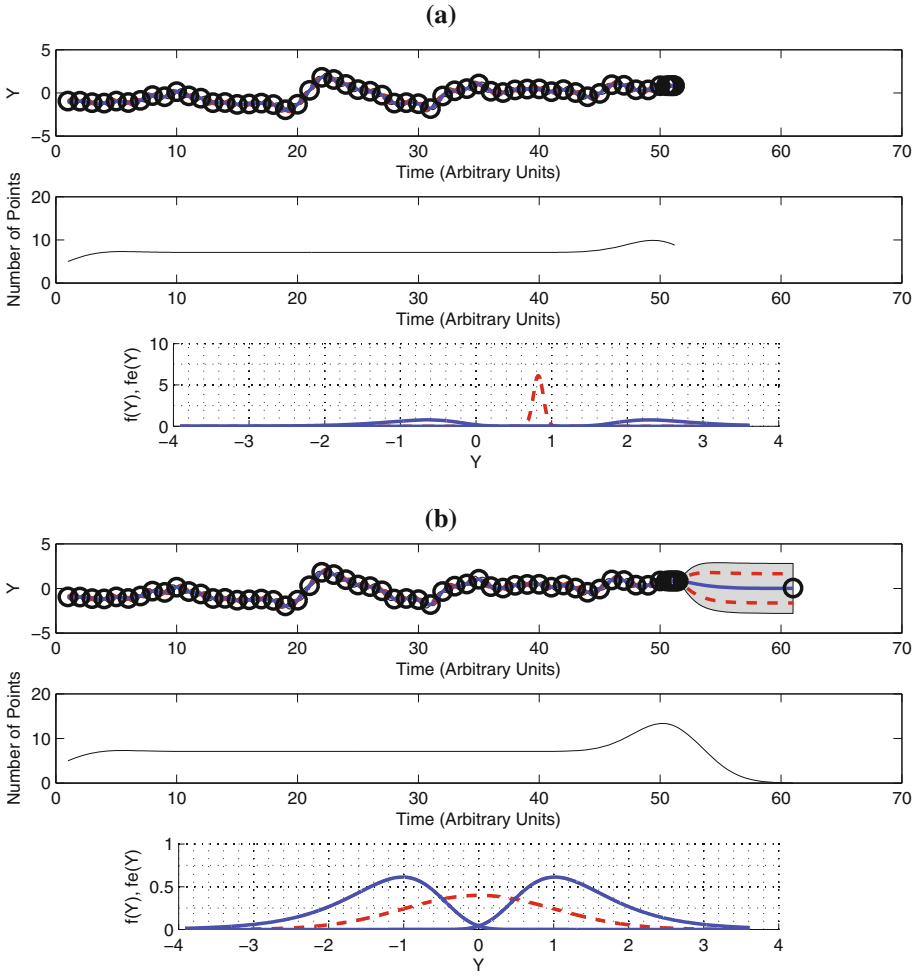

Fig. 2 Simulation of the effect of varying observation density on the extreme value distribution and hence the anomaly detection. *Both plots* show a snapshot from the last observed sample. In **a**, the observation rate is high, whilst in **b** the observation rate is low. The observation density affects the location of the probability density function of the extreme value distribution $f_e(y)$ (*blue lines*, *lower plot*), relative to the predicted Gaussian PDF $f(y)$ (*red dashed line*, *lower plot*), drifting closer to the base distribution as the number density of points decreases. This is due to the relationship between the observation density and the location and scaling of the extreme value distribution, expressed in Eqs. 13 and 14. **a** The sequential GP–EVT (*continuous line*, *upper plot*) stopped at a region of high observation density. The estimate of the number of data points that contribute to the GP inference has increased significantly, as indicated by the observation density shown in the *middle plot*. Consequently, the location of the extreme value distributions, illustrated by the *continuous lines* in the *bottom plot*, move away from the posterior predictive distributions (*dashed line*). **b** The sequential GP–EVT (*continuous line*, upper plot) stopped at a region of very low observation density. The estimate of the number of data points that contribute to the GP inference has decreased significantly, as indicated by the observation density shown in the *middle plot*. Consequently, the location of the extreme value distributions, illustrated by the *continuous lines* in the *bottom plot*, moves towards the posterior predictive distributions (*dashed line*) (color figure online)

verting the GPS information into a 1D feature vector, reducing the computational demands of processing the data. Also, the arc length between points d for a sphere of radius r and $\Delta\hat{\sigma}$ is given in radians by $d = r\Delta\hat{\sigma}$.

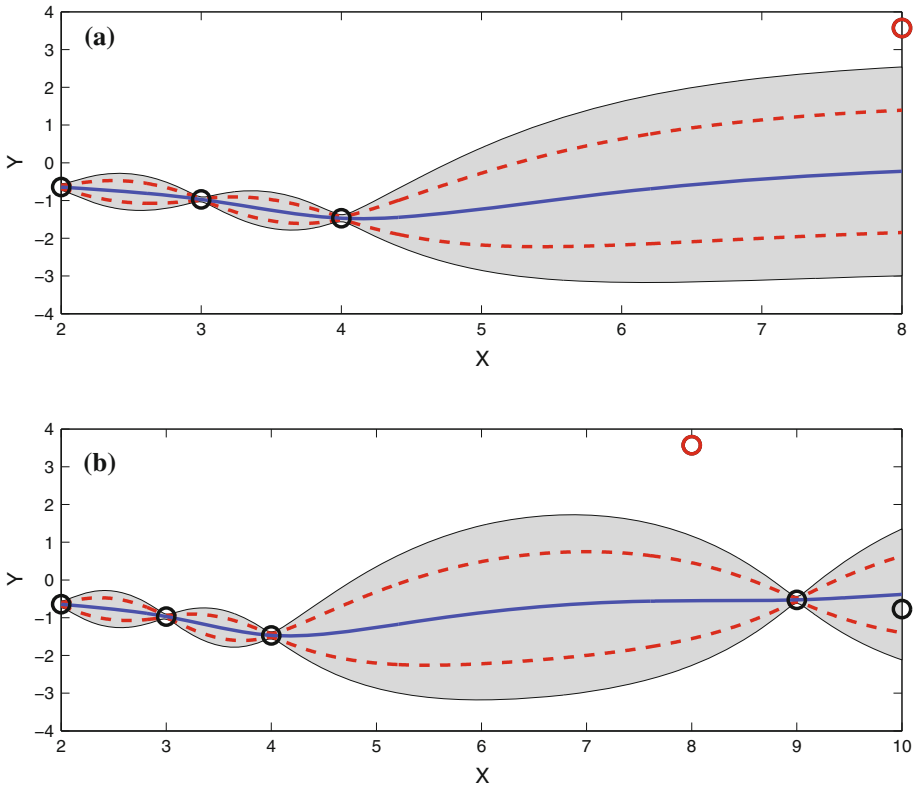

Fig. 3 Simulation of the GP prediction and anomaly detection. The *continuous line* shows the predicted mean function and the *grey areas* show the EVT bound of the GP predictive distribution for $p = 0.95$. The bound is open to the right and widening until the next observation has been included. Once it has, the standard deviation bound of the GP is updated, as seen between time steps 4 and 9. The *dashed line* shows the error bound produced if we consider the 95 % bound from the mean function (1.64 standard deviations from the mean). **a** The GP predicts forward to the new artificially perturbed data point, and by using GP–EVT, the new observation is classified as an anomaly. **b** GP predicts forward after detecting and excluding the artificial anomaly. The uncertainty bounds continue to increase (until they reach their maximum as set by the prior distribution). The subsequent observations fall well within the error bound and so will be included in the next update

Choice of covariance functions The choice of covariance function is crucial in the methods ability to provide the most accurate representation of the vessel dynamics. To determine the optimal covariance function, we investigated the performance of the standard squared exponential kernel, Eq. 1, Matérn $\frac{3}{2}$, Eq. 2, and the Matérn $\frac{1}{2}$ kernel, Eq. 3. Clean, i.e. anomaly-free training data, was extracted from the training corpus and the GP kernel function parameters were estimated by maximising the marginal likelihood of the data [21]. The likelihoods were standardised to the training data length, and the mean standardised likelihoods are shown in Table 1.

The results suggest almost comparable performance, in terms of goodness of fit, of all three tested covariance functions. However, as shown in Fig. 4, there is a substantial difference in the robustness. The Matérn $\frac{1}{2}$ kernel frequently finds poorer fits to the data. The squared exponential performs in the middle range, occasionally finding worse solutions than the Matérn $\frac{3}{2}$ kernel but better than the Matérn $\frac{1}{2}$ kernel.

Table 1 Table of mean standardised likelihood scores for the Matérn $\frac{3}{2}$, Matérn $\frac{1}{2}$ and standard squared exponential kernels

Matérn $\frac{3}{2}$	Matérn $\frac{1}{2}$	SE
0.3345	0.3312	0.3337

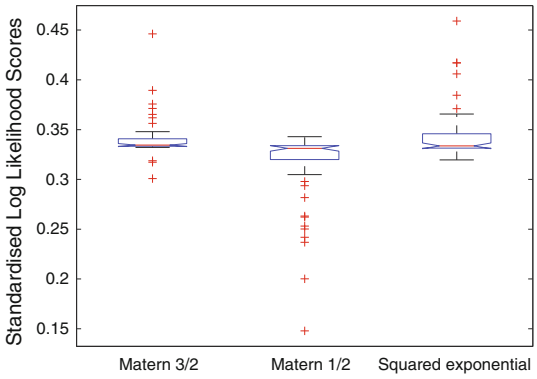

Fig. 4 Box plot of log scores for the different covariance functions applied to each track

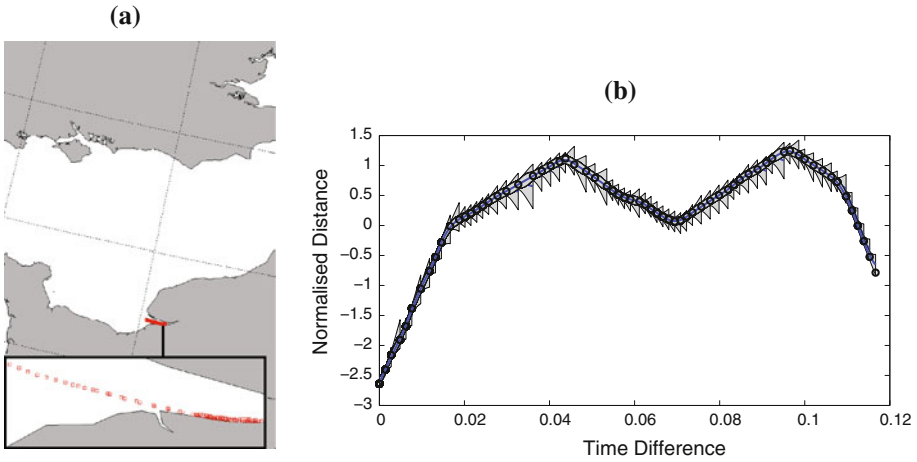

Fig. 5 Sequential GP–EVT method applied to a dredging vessel operating off the coast of France near Le Havre. **a** A plot of the GPS track in which there were no detected anomalies. **b** Sequential predictions applied to feature-extracted data, also showing that all data points fall within the EVT bound

Vessel track modelling The methodology was also applied to real-world vessel track data. The Matérn $\frac{3}{2}$ kernel was chosen to model the underlying dynamics using hyperparameters learnt from anomaly-free training data, which were chosen to be sufficiently long enough so that the underlying dynamics of the vessels could be captured.

Figure 5 shows an example vessel track without outlying points. The track is from a dredger that follows a smooth trajectory and does not make any sudden changes in acceleration. Shown in Fig. 5b are the sequential EVT bounds sea-sawing until the next observation arrives. All observations fall well inside the predictive boundary of the GP–EVT bound, and consequently, no anomalies are detected.

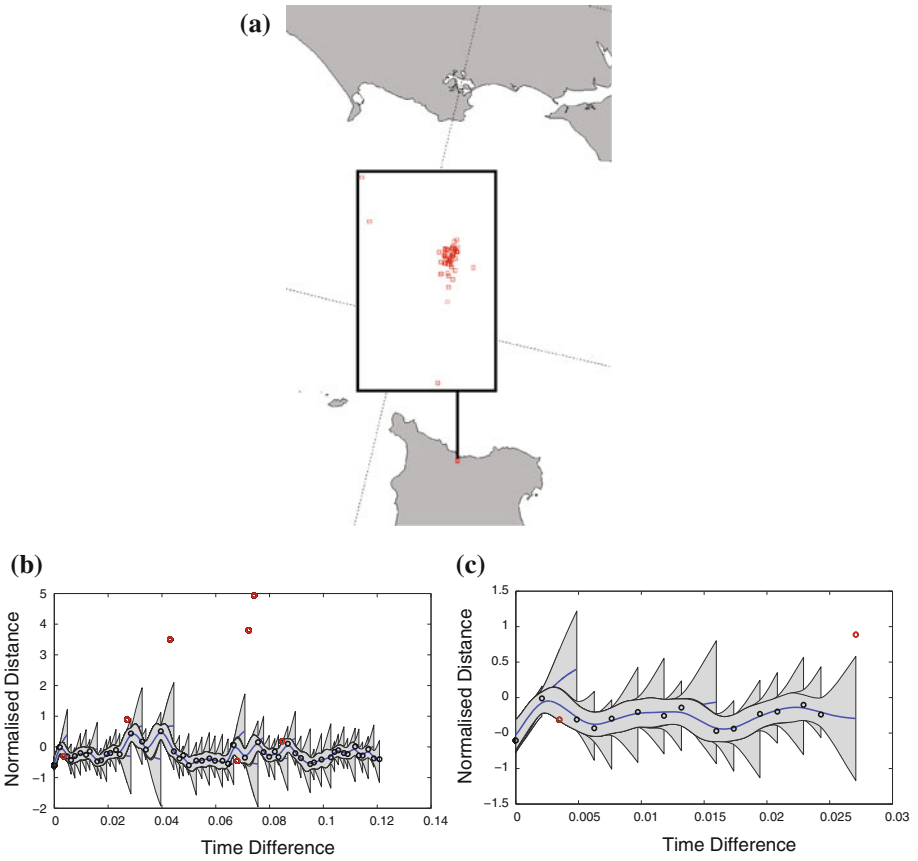

Fig. 6 Sequential GP–EVT method applied to a small vessel operating off the coast of France near Cherbourg and whose track suggests unusual navigation behaviour. **a** A plot of the GPS track in which there were several detected anomalies. **b** Sequential predictions applied to feature-extracted data, also showing some data points that fall outside the GP–EVT bound. **c** Magnified section of the plot in **b**. The GP–EVT bound has predicted forward to the new data point and included the point in the update

Figure 6 shows an example of a vessel track with some points, which our model labels as anomalies. As can be seen in Fig. 6a, the vessel remains within a confined area and there are short sudden movements, Fig. 6b. These are marked as anomalies and are perhaps the result of the vessel drifting, manoeuvring or being moored. Figure 6c also shows an enlarged section of the sequential computation of the EVT bound and makes clear the nonlinear relationship between the GP standard deviation bound and the actually computed EVT bound, which includes essential parameters such as the inferred number of observations.

5 Comparison of GP–EVT with a traditional Kalman filter approach

In this section, we compare a traditional approach to anomaly detection with our Gaussian process and EVT approach. The traditional approach uses a Kalman filter (KF) to model the normal behaviour of the ship and then determines that the data are anomalous if it is more than a fixed number of standard deviations from the mean [13] (typically 3–5 standard deviations). This approach to anomaly detection we term the *gating* approach.

Table 2 AUC for KF using the near-constant velocity model (with and without EVT) and GP using a Matérn $\frac{3}{2}$ model (again with and without the EVT)

GP–EVT	GP	KF-EVT	KF
0.8032	0.7889	0.6545	0.6119

This was repeated for a range of confidence regions at 1, 1.64, 3 and 5 standard deviations. Correspondingly, the confidence region of the GP and KF using the extreme value approach was set by selecting probabilities of 0.84, 0.95, 0.99 and 0.999

Further, the KF approach requires a process model of the normal behaviour of the ship. Typically, a near-constant velocity model is chosen to model the continuous trajectory without imposing any excessive smoothness on the trajectory [4]. We compare our approach to a KF, which uses the near-constant velocity model. This provides a fair comparison as both Matérn $\frac{3}{2}$ and constant velocity models are second-order differentiable. We further investigate both a traditional KF using the standard deviation gating approach to exclude anomalies and a KF that uses the EVT in a manner similar to the GP. In so doing, we are able to compare both models of normal ship behaviour (namely the Matérn $\frac{3}{2}$ and the near-constant velocity model) and also both approaches to detecting and excluding anomalies (namely the EVT and standard deviation gating approaches).

When using the KF and GP, the mean and standard deviation were predicted forward to the same time step as the new observation. If the point lies within a pre-chosen confidence region (defined as a multiple of the standard deviation about the mean), it is included in the update. ROC curves were plotted for the results. The resulting area under curve (AUC) which compares the KF using the near-constant velocity model (with and without EVT) against the GP using a Matérn $\frac{3}{2}$ model (again with and without the EVT) is shown in Table 2.

We note that both the KF and GP performances are significantly improved using the EVT as opposed to gating. This is due to the fact that the gating approach uses a fixed threshold, which does not take into account the density of observations ,i.e. as we observe more samples, we gain a better understanding of the true distribution of values. The EVT, however, uses a dynamic threshold, which takes into account the density of observations therefore better utilises available information to adjust the threshold.

Although the results indicate a significant improvement of our model over the KF approach, this is a limitation of the near-constant velocity model used and not a critique of KF-based methods. We note that the Matérn $\frac{3}{2}$ GP model can be efficiently implemented within the KF as a Markov process model [6]. Furthermore, it is possible to implement the near-constant velocity model as a GP model, see, for example, [22]. Thus, it is possible to match the AUC of the KF approach and GP approach by replacing the near-constant velocity model in the KF by the Markovianised Matérn model [6].

However, the results illustrate the significant improvement obtained using EVT as opposed to a simple gating mechanism based on the number of standard deviations between a datum and the expected position of the ship.

6 Identifying anomalous tracks

Whilst the previous section dealt with the identification of anomalous points, it does not allow the track to be considered as a whole. Thus, patterns and abnormalities at a broader scale may be missed. To address this issue, we cluster the nonanomalous vessel tracks and then detect

abnormalities from the identified cluster. Discovery of clusters is dependent on identifying tracks with similar characteristics, as determined by the characteristics of the vessel class. We discover similarity (or adjacency) through the use of GP modelling and application of the Hellinger distance.

6.1 Hellinger distance

Detection of clusters is dependent on the creation of a matrix of distances (an adjacency matrix); in this instance, this is not simply a distance in the spatial sense but must express the distance in the intrinsic dynamics that created the track, i.e. the track shape. Fortunately, the GP describes a draw from a probability distribution over functions, and many distance measures between probability distributions exist [1]. However, only true metrics produce a symmetric adjacency matrix [2]. The Hellinger distance is one such measure of similarity between two probability distributions and is defined as

$$h^2\left(f\left(\boldsymbol{x}\right), g\left(\boldsymbol{x}\right)\right) = \frac{1}{2}\int\left(\sqrt{f\left(\boldsymbol{x}\right)} - \sqrt{g\left(\boldsymbol{x}\right)}\right)^2 \mathrm{d}\boldsymbol{x},$$

We consider the case in which f and g are GPs, defined via their mean and covariance functions.

In order to simplify the derivation of the squared Hellinger distance, h^2 between GPs, it can be assumed that both distributions have the same zero mean, $\boldsymbol{\mu}_f = \boldsymbol{\mu}_g = \mathbf{0}$, which we ensure within the data by trend removal. Thus, the squared Hellinger distance can take the simplified form

$$h^2\left(f\left(\boldsymbol{x}; \boldsymbol{\mu}_f, \boldsymbol{\Sigma}_f\right), g\left(\boldsymbol{x}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g\right)\right) = 1 - \frac{|\frac{1}{2}\boldsymbol{\Sigma_1} + \frac{1}{2}\boldsymbol{\Sigma_2}|^{-\frac{1}{2}}}{|\boldsymbol{\Sigma}_f|^{-\frac{1}{4}}|\boldsymbol{\Sigma}_g|^{-\frac{1}{4}}}. \tag{15}$$

Thus, in this case, the distance metric is dependent only on the choice of kernel function used to model the data.

Application of a GP model to each track and estimation of the track hyperparameters through maximising the marginal likelihood of the data allow the track dynamics to be inferred. The combination of GPs and the Hellinger distance in this manner provides the ability to define a measure between two stochastically sampled functions. Comparison between tracks can be made through the generation of a covariance matrix using hyperparameter values inferred from the GP applied to each track and a common input scale. An adjacency structure can then be formed by calculating the inverse distance between each GP.

6.2 Community detection

The use of inverse Hellinger distance between GPs allows us to map our data to a relational space, where each pair i, j of vessel tracks is assigned a similarity value s_{ij}. We encode all similarity pairs to a matrix $\mathbf{S}$, so that s_{ij} is the degree of coupling between the paths of vessels i and j. From such relational structure, we seek to apply an appropriate clustering scheme in order to extract classes of vessels that exhibit similar movement patterns. In Reece et al. [23], we note tracks were clustered using GPs and in which the most likely mixture model for the data was identified, this corresponding to the cluster or community for a given type of track. The distance measure in this instance was the likelihood function. Unfortunately, this method scales poorly with the number of tracks.

By treating $\mathbf{S}$ as an adjacency matrix from a network analysis perspective, where N vessels are nodes and similarities (in the Hellinger metric sense) are link weights, we apply the idea of *community detection* [18], in order to discover the groups of strongly connected nodes, so that a given vessel i has more similar paths with vessels inside a community than with the ones outside.

Towards the above goal, we employ a Bayesian nonnegative matrix factorisation (NMF) scheme [19], which has already been successfully applied to a wide range of community detection problems [19,20,26]. In this approach, communities are treated as explanatory latent variables for the observed link weights, so that the stronger the similarity between two vessel paths the more likely it is that they belong to the same community. We extract such latent grouping via an appropriate factorisation of $\mathbf{S} \simeq \mathbf{WH}$, $\mathbf{S} \in \mathbb{R}^{N\times N}$, $\mathbf{W}, \mathbf{H}^{\mathsf{T}} \in \mathbb{R}^{N\times K}$, where both the inner rank K and the factor elements w_{ik}, h_{ki} are inferred via an appropriate maximum a posteriori (MAP) scheme. This model has the advantage of not only discovering overlapping communities (soft-partitioning) but also quantifying how strongly each vessel belongs to a particular class. Such result allows us to quantify how the broadcast vessel class "disagrees" with the one we infer from the path similarity matrix. It also avoids the scaling issues as posed by the method proposed in [23].

6.3 Detecting anomalous tracks

Anomalous tracks can be identified by noting those tracks not assigned to a vessel community, i.e. those that have track characteristics sufficiently different from all other vessel tracks and are not assigned to a cluster. Additionally, we can identify discrepancies between the class of vessel from which the vessel claims to belong and the cluster to which it has been assigned.

Furthermore, we note that once a reference cluster has been formed using the NMF community model, and a soft membership score π_i obtained for each node, then subsequent tracks can be tested using extreme value theory. By taking each of the distances d_i between tracks in a cluster, the standard deviation of the group can be calculated

$$sd = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \pi_i d_i^2},$$

where N is the number of distances within the group. Selecting the track with the highest probability of belonging to the community as the centroid, we can now test subsequent tracks to identify anomalies with respect to the cluster. For example, a vessel claiming to belong to a given vessel class can be tested against a reference community to identify anomalies with respect to the test community. Extreme value theory can again be applied in the same manner as used previously, estimating the variability of the population based on the observation density. The distance between a given test track and the group centroid can then be determined and appropriately scaled to the group using the standard deviation calculated above.

7 Application

We demonstrate the efficacy of the method through application to both real and synthetic data. Synthetic data are used to illustrate the principles behind the methodology, and a real-world example is used to demonstrate its applicability.

7.1 Synthetic data illustration

Data were generated from a multiple-output GP using a Matérn $\frac{3}{2}$ kernel with known scaling parameters. For the first 6 generated functions, Fig. 7a, hyperparameter values of $\varepsilon = 5$ and $\lambda = 0.01$ were, respectively, assigned to the output and noise scale. Additionally, the output correlation hyperparameters were assigned values $[l_1]_1 = 0.1$, $[l_2]_1 = 0.1$ and $[l_2]_2 = 0.1$, and these hyperparameter values will be referred to as set A. A further 6 functions were generated from a GP, Fig. 7b, with hyperparameter values of $\varepsilon = 20$, $\lambda = 0.01$, $[l_1]_1 = 0.02$, $[l_2]_1 = 0.02$ and $[l_2]_2 = 0.02$, and these hyperparameter values will be referred to as set B.

Inferring hyperparameter values from the generated functions and taking the inverse of the Hellinger distance metric between tracks provides for the creation of an adjacency matrix, Fig. 8. This allows a clear split in the data to be identified.

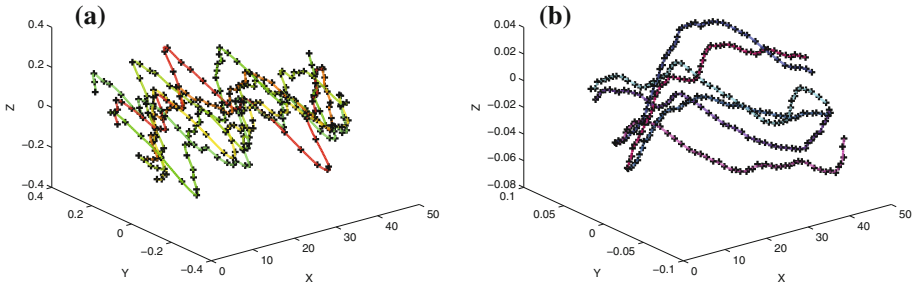

Fig. 7 Plots of points drawn from a multiple-output GP and the subsequent multiple-output GP regression. **a** Functions generated from a multiple-output GP with hyperparameter values of $\varepsilon = 5$, $\lambda = 0.01$, $[l_1]_1 = 0.1$, $[l_2]_1 = 0.1$ and $[l_2]_2 = 0.1$ (set A). The mean function using the inferred hyperparameter values is shown plotted through the sample data. **b** Functions generated from a multiple-output GP with hyperparameter values of $\varepsilon = 20$, $\lambda = 0.01$, $[l_1]_1 = 0.02$, $[l_2]_1 = 0.02$ and $[l_2]_2 = 0.02$ (set B). The mean function using the inferred hyperparameter values is shown plotted through the sample data

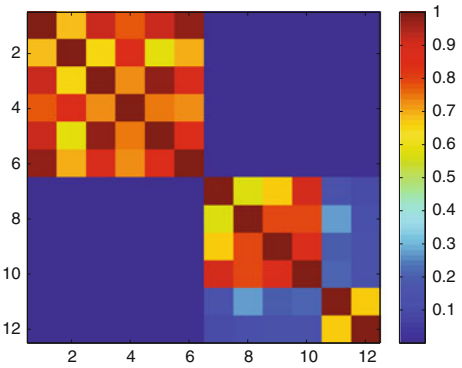

Fig. 8 Synthetic data example: matrix of the inverse Hellinger distance (adjacency) between inferred functions. The axis of the matrix corresponds to (in order) the distance between functions 1 through 6 (generated from set A hyperparameters) and then functions 7 through 12 (generated from set B hyperparameters). Distances which are very close to one another have a value close to 1 and further apart close to zero, this is as represented via the *colour bar to the right of the figure* (color figure online)

Application of the network NMF community detection then correctly identifies functions 1 through 6 (generated from set A hyperparameters) as being from one community, and functions 7 through to 12 as belonging to a distinctly separate community (generated from set B hyperparameters).

7.2 Vessel track anomaly detection

The outlined methodology was also applied to real-world vessel track data. The data consist of a set of collected AIS data, and this was combined with the registered vessel information in order to identify the class of vessel. As such, the data collected can be used to evaluate the methodology due to possession of the ground truth. A small illustrative test set of data is used to provide a means of illustrating the technique. The tracks used were for cargo vessels shown in Fig. 9, fishing vessels shown in Fig. 10 and sailing vessels shown in Fig. 11. The inverse Hellinger distance (adjacency) between tracks is shown in Fig. 12. The resulting community structure from the application of the aforementioned technique is shown in Fig. 13.

Abnormalities can be identified in Fig. 13 by noting those tracks unassigned to a community (Tracks 8–12). The majority of the unassigned tracks belong to the sailing vessel class Fig. 11, and it can noted that these tracks do not exhibit common movement characteristics.

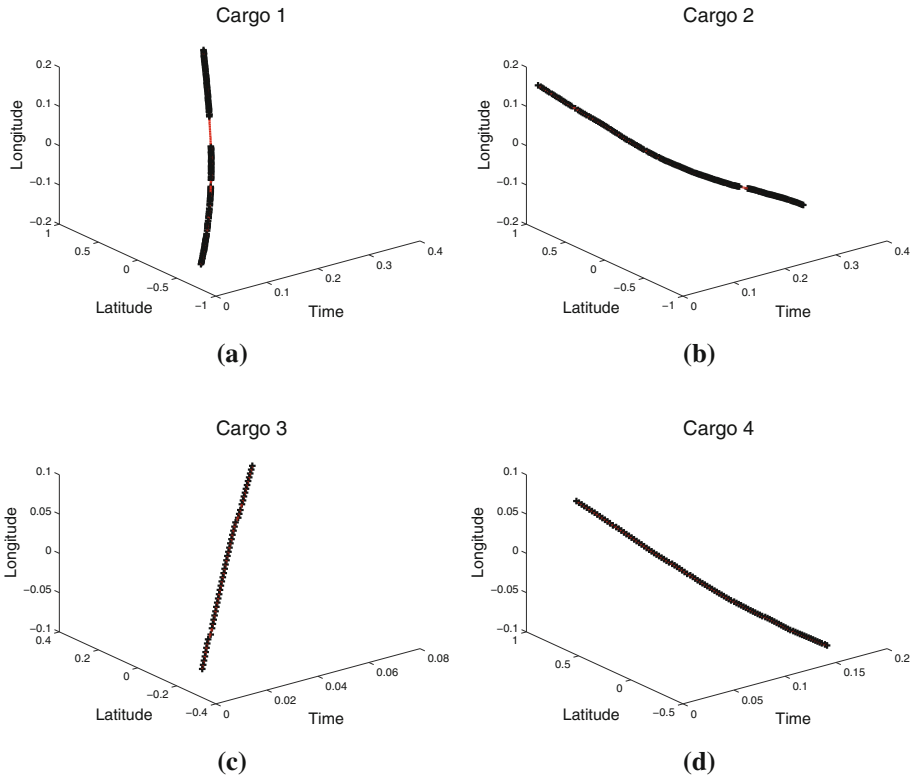

Fig. 9 GP multiple-output regression through cargo vessel GPS coordinates

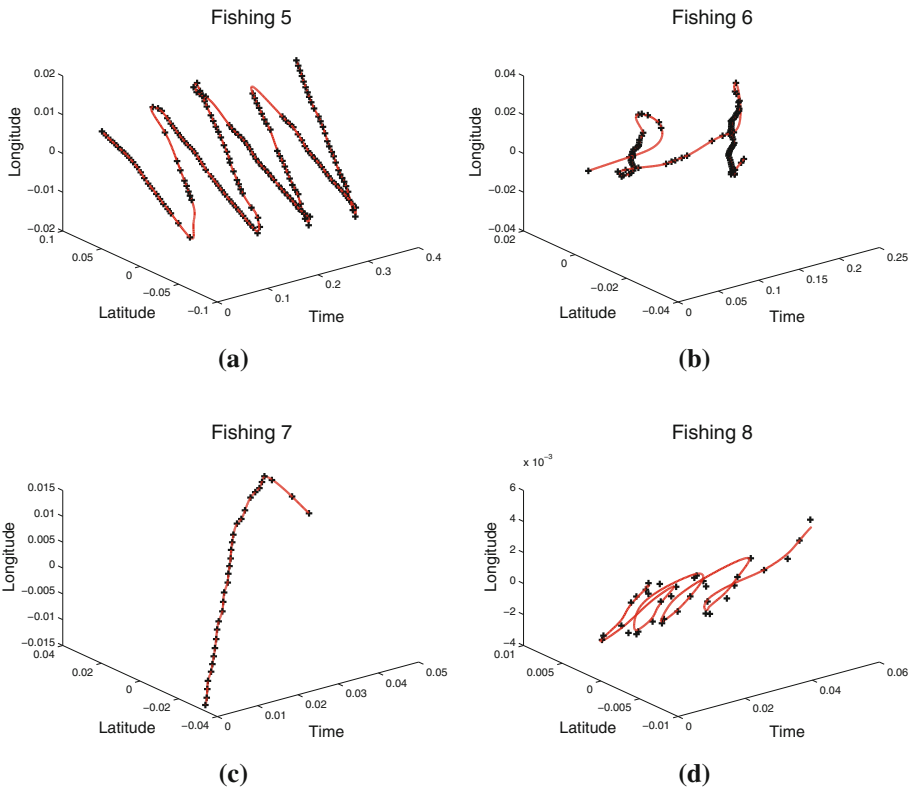

Fig. 10 GP multiple-output regression through fishing vessel GPS coordinates

8 Conclusion

Extreme value theory has proven to be an extremely successful framework for anomaly detection. Unlike novelty detection based directly on the sample distribution, extreme value distributions capture our beliefs that extreme events should become more extreme if large numbers of measurements are expected and vice versa. Such detection, however, has to be dynamic, context-sensitive and timely if it is to be useful for marine tracking. Extreme value distributions alone are not readily adapted to perform this task.

In this paper, we present an alternative to endowing extreme value distributions directly with dynamic properties. Our approach simultaneously models the dynamic properties of the underlying extreme value generative distribution and the dynamic properties of the data sampling process. To our knowledge, this is the first time that extreme value distributions have been made dynamic through the use of GPs.

Our approach offers several advantages. GPs provide a flexible, nonparametric and intuitive tool to describe typical vessel dynamics. Also, measurement and prediction is performed in continuous time, thus allowing on-demand anomaly detection. The sequential update of the GP covariance matrix bypasses the need for inverting massive matrices and substantially reduces the computational burdens for which GPs are well known.

Our empirical experiments on vessel data suggest that the method is capable of detecting anomalies that resemble mooring or drifting and unexpected departures from regular

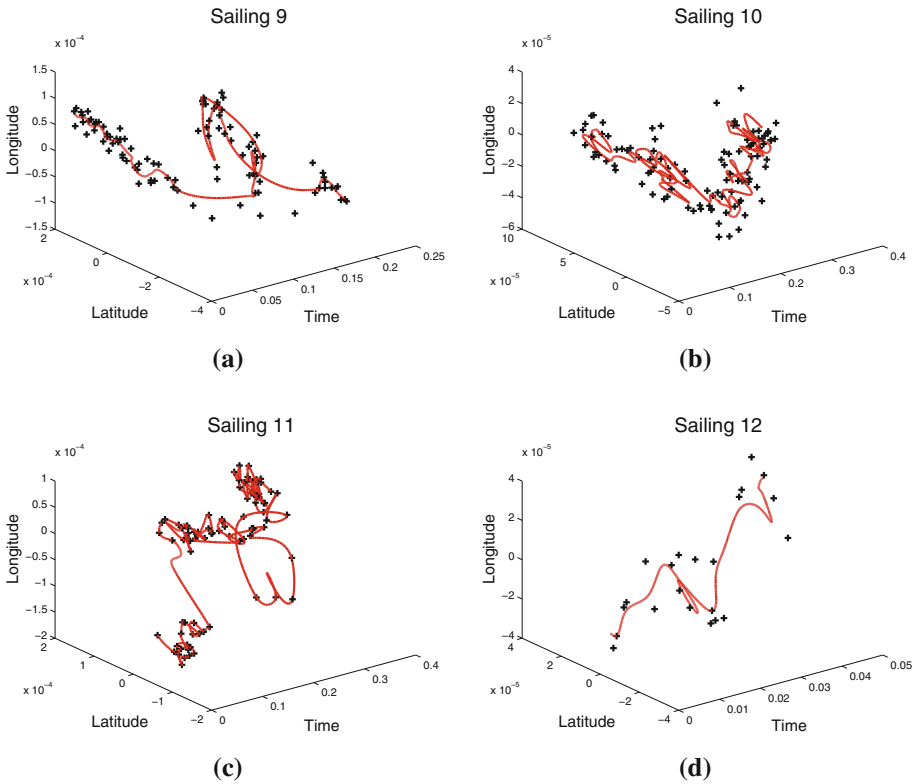

Fig. 11 GP multiple-output regression through sailing vessel GPS coordinates

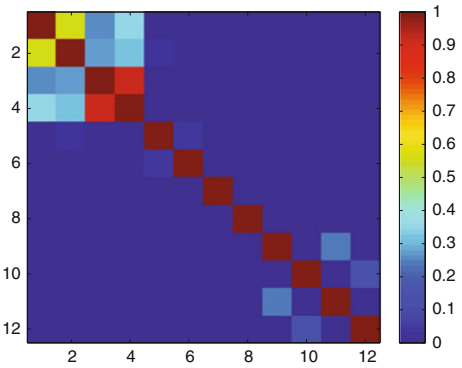

Fig. 12 Vessel track data: matrix of the inverse Hellinger distance (adjacency) between inferred functions for the vessel data. The axis of the matrix corresponds to (in order) the distance between the functions inferred for cargo vessel tracks 1–4, fishing vessel tracks 5–8 and sailing vessel tracks 9–12. Distances which are very close to one another have a value close to 1 and further apart close to zero, this is as represented via the *colour bar to the right of the figure* (color figure online)

movements. The sample size prediction plays the important role of adapting the observation process in time. As the effective sample size reduces, the extreme value distribution approaches the regular Gaussian distribution, as Eq. 7 suggests. With increasing density of

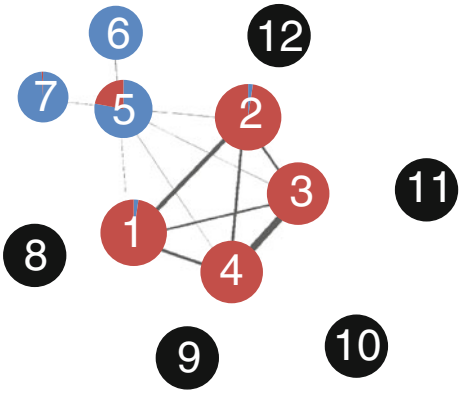

Fig. 13 Network diagram illustrating the relationship between different vessel types. Each edge connection is weighted (illustrated by the varying edge thickness), the weights being determined by the inverse Hellinger distance between them. *Nodes 1–12* relate to the different classes of vessel, cargo *(1–4)* Fig. 9, fishing *(5–8)* Fig. 10 and sailing *(9–12)* Fig. 11. *Different colours* relate to the different communities within the data, and the membership score of each node per community is defined by the pie chart. The *black nodes* are unassigned and do not belong to a given community (color figure online)

observations, however, the extreme value distribution diverges and EVT bound increases. Although the choice of GP kernel function becomes less critical with increasing amounts of data, for smaller sample sizes, the kernel function is critical and our empirical results have shown that the Matérn $\frac{3}{2}$ kernel outperforms the near-constant velocity model.

Furthermore, by moving away from the detection of anomalous points and considering the similarity between tracks via use of the Hellinger distance between Gaussian processes, we have provided a methodology in which anomalous tracks can be detected. A more appropriate application of the work may be in identifying deviations from typical shipping routes. Vessels which are forced to operate under certain operating conditions may belong to a community where their tracks are forced to take a restrictive form, conditioned by the operating conditions. Vessels deviating from the constrained conditions may have a distinctly different functional form which may be identified as belonging to a different community structure.

9 Future work

The representative choice of distance as the dependent variable feature for anomaly detection is open to discussion. Whilst it provides a single dimension and, thus, fast estimation, it does fail to capture some aspects of ship tracks. To capture such features, the GPS coordinates can be simultaneously modelled with a bivariate GP and extrema modelling.

The training using typical vessel tracks will be extended to shipping lanes and vessel types. This allows anomaly detection not just on the basis of individual points but entire tracks and so offers the possibility of preventing accidents such as that of the MS Costa Concordia early in 2012. Kernel regression-based prediction of the sample size can be readily extended using Poisson processes.

Acknowledgments This work was funded by ISSG, Babcock Marine and Technology Division, Devonport Royal Dockyard. Ioannis Psorakis is funded from a grant via Microsoft Research, for which we are most grateful. This work was further supported by EPSRC project EP/I011587/1.

10 Appendix

Derivation of the Hellinger metric, Eq. 15.

The squared Hellinger distance is a measure of similarity between two probability distributions and is defined as

$$h^2\left(f\left(\boldsymbol{x}\right), g\left(\boldsymbol{x}\right)\right) = \frac{1}{2}\int\left(\sqrt{f\left(\boldsymbol{x}\right)} - \sqrt{g\left(\boldsymbol{x}\right)}\right)^2 \mathrm{d}\boldsymbol{x}, \tag{16}$$

where $f\left(\boldsymbol{x}\right)$ and $g\left(\boldsymbol{x}\right)$ denote probability distributions. Equation 16 can alternatively be expressed as

$$h^2\left(f\left(\boldsymbol{x}\right), g\left(\boldsymbol{x}\right)\right) = 1 - \int\left(\sqrt{f\left(\boldsymbol{x}\right)}\sqrt{g\left(\boldsymbol{x}\right)}\right)\mathrm{d}\boldsymbol{x}.$$

In the instance distributions $f\left(\boldsymbol{x}\right)$ and $g\left(\boldsymbol{x}\right)$ are multivariate Gaussian, the Hellinger distance would take the form

$$\begin{aligned} &h^2\left(f\left(\boldsymbol{x};\boldsymbol{\mu}_f,\boldsymbol{\Sigma}_f\right), g\left(\boldsymbol{x};\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g\right)\right) \\ &= 1 - \frac{1}{\left(2\pi\right)^{\frac{n}{2}}\sqrt{\left|\boldsymbol{\Sigma}_f\right|}}\frac{1}{\left(2\pi\right)^{\frac{n}{2}}\sqrt{\left|\boldsymbol{\Sigma}_g\right|}} \\ &\quad\times\int\sqrt{\exp\left(-\frac{1}{2}\left(\boldsymbol{x}-\boldsymbol{\mu}_f\right)^\top{\boldsymbol{\Sigma}_f}^{-1}\left(\boldsymbol{x}-\boldsymbol{\mu}_f\right)\right)} \\ &\quad\times\sqrt{\exp\left(\frac{1}{2}\left(\boldsymbol{x}-\boldsymbol{\mu}_g\right)^\top{\boldsymbol{\Sigma}_g}^{-1}\left(\boldsymbol{x}-\boldsymbol{\mu}_g\right)\right)}\mathrm{d}\boldsymbol{x}, \end{aligned}$$

The terms inside the exponents can also be combined and expressed in quadratic form, $(\boldsymbol{x}-\boldsymbol{\mu}^*)^\top\boldsymbol{C}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}^*)+\boldsymbol{B}$, by making the following associations

$$\begin{aligned} \boldsymbol{\mu}^* &= \left({\boldsymbol{\Sigma}_f}^{-1}+{\boldsymbol{\Sigma}_g}^{-1}\right)^{-1}\left({\boldsymbol{\Sigma}_f}^{-1}\boldsymbol{\mu}_f+{\boldsymbol{\Sigma}_g}^{-1}\boldsymbol{\mu}_g\right), \\ \boldsymbol{C}^{-1} &= \frac{1}{2}{\boldsymbol{\Sigma}_f}^{-1}+\frac{1}{2}{\boldsymbol{\Sigma}_g}^{-1}, \\ \boldsymbol{B} &= \left(\boldsymbol{\mu}_f-\boldsymbol{\mu}_g\right)^\top\left(\boldsymbol{\Sigma}_g+\boldsymbol{\Sigma}_f\right)^{-1}\frac{1}{2}\left(\boldsymbol{\mu}_f-\boldsymbol{\mu}_g\right). \end{aligned}$$

The integral can now be solved and the expression simplified

$$\begin{aligned} &h^2\left(f\left(\boldsymbol{x}\right), g\left(\boldsymbol{x}\right)\right) \\ &= 1 - \frac{\left|\frac{1}{2}{\boldsymbol{\Sigma}_f}^{-1}+\frac{1}{2}{\boldsymbol{\Sigma}_g}^{-1}\right|^{-\frac{1}{2}}}{\left|\boldsymbol{\Sigma}_f\right|^{\frac{1}{4}}\left|\boldsymbol{\Sigma}_g\right|^{\frac{1}{4}}} \\ &\quad\times\exp\left(-\frac{1}{4}\left(\boldsymbol{\mu}_f-\boldsymbol{\mu}_g\right)^\top\left(\boldsymbol{\Sigma}_g+\boldsymbol{\Sigma}_f\right)^{-1}\left(\boldsymbol{\mu}_f-\boldsymbol{\mu}_g\right)\right). \end{aligned}$$

Under the assumption that both distributions have the same zero mean, $\boldsymbol{\mu}_f=\boldsymbol{\mu}_g=\boldsymbol{0}$, this can be further simplified

$$\begin{aligned} &h^2\left(f\left(\boldsymbol{x};\boldsymbol{\mu}_f=\boldsymbol{0},\boldsymbol{\Sigma}_f\right), g\left(\boldsymbol{x};\boldsymbol{\mu}_g=\boldsymbol{0},\boldsymbol{\Sigma}_g\right)\right) \\ &= 1 - \frac{\left|\frac{1}{2}{\boldsymbol{\Sigma}_f}^{-1}+\frac{1}{2}{\boldsymbol{\Sigma}_g}^{-1}\right|^{-\frac{1}{2}}}{\left|\boldsymbol{\Sigma}_f\right|^{\frac{1}{4}}\left|\boldsymbol{\Sigma}_g\right|^{\frac{1}{4}}} \end{aligned}$$

To avoid the inverse covariances in this form, the fraction can be multiplied top and bottom by $\left(|\boldsymbol{\Sigma}_f||\boldsymbol{\Sigma}_g|\right)^{-\frac{1}{2}}$, in addition to the application of the determinant identity $|\boldsymbol{AB}| = |\boldsymbol{A}||\boldsymbol{B}|$, thus

$$\begin{aligned} & h^2\left(f\left(\boldsymbol{x};\boldsymbol{\mu}_f=\mathbf{0},\boldsymbol{\Sigma}_f\right),g\left(\boldsymbol{x};\boldsymbol{\mu}_g=\mathbf{0},\boldsymbol{\Sigma}_g\right)\right) \\ &= 1-\frac{|\frac{1}{2}\boldsymbol{\Sigma}_f+\frac{1}{2}\boldsymbol{\Sigma}_g|^{-\frac{1}{2}}}{|\boldsymbol{\Sigma}_f|^{-\frac{1}{4}}|\boldsymbol{\Sigma}_g|^{-\frac{1}{4}}} \\ &= 1-\sqrt{2}\frac{|\boldsymbol{\Sigma}_f{}^{\frac{1}{4}}+\boldsymbol{\Sigma}_g{}^{\frac{1}{4}}|}{|\boldsymbol{\Sigma}_f+\boldsymbol{\Sigma}_g|^{\frac{1}{2}}} \end{aligned}$$

It can be noted, as a means of verifying the result, that by setting $\boldsymbol{\Sigma} = \sigma^2$, the form is consistent with the Hellinger distance between two univariate Gaussian distributions when $\mu_f = \mu_g = 0$ namely

$$1-\sqrt{\frac{2\sigma_f\sigma_g}{\sigma_f^2+\sigma_g^2}}\exp\left(-\frac{\left(\mu_f-\mu_g\right)^2}{4\left(\sigma_f^2+\sigma_g^2\right)}\right).$$

References

1. Basseville M (1989) Distance measures for signal processing and pattern recognition. Signal Process 18(4):349–369
2. Budzynski R, Kondracki W, Krolak A (2008) Applications of distance between probability distributions to gravitational wave data analysis. Class Quantum Gravity 25(1):015005
3. Coles S (2001) An introduction to statistical modelling of extreme values. Springer, UK
4. George J, Crassidis J, Singh T et al (2011) Anomaly detection using content-aided target tracking. J Adv Inf Fusion 6(1):39–56
5. Grubbs F (1969) Procedures for detecting outlying observations in samples. Technometrics 11(1):1–21
6. Hartikainen J, Särkkä S (2010) Kalman filtering and smoothing solutions to temporal Gaussian process regression models. In: Proceedings of IEEE international workshop on machine learning for signal processing (MLSP). Kittilä, Finland, pp 379–384
7. Lane R, Nevell D, Hayward S et al (2010) Maritime anomaly detection and threat assessment. In: Proceedings of 13th conference on information fusion (FUSION). Edinburgh, UK, pp 1–8
8. Laws K, Vesecky J and Paduan J (2011) Monitoring coastal vessels for environmental applications: application of Kalman filtering. In: Proceedings of 10th current, waves and turbulence measurements (CWTM). Monterey, CA, USA, pp 39–46
9. Laxhammar R (2008) Anomaly detection for sea surveillance. In: Proceedings of 11th international conference on information fusion. Cologne, Germany, pp 1–8
10. Laxhammar R, Falkman G, Sviestins E (2009) Anomaly detection in sea traffic—a comparison of the Gaussian mixture model and the Kernel density estimator. In: Proceedings of 12th international conference on information fusion. Seattle, WA, USA, pp 756–763
11. Lee H, Roberts S (2008) On-line novelty detection using the Kalman filter and extreme value theory. In: Proceedings of 19th international conference on pattern recognition. Tampa, Florida, USA, pp 1–4
12. Li X, Han J, Kim S (2006) Motion-alert: automatic anomaly detection in massive moving objects. In: Proceedings of IEEE intelligence and security informatics. San Diego, CA, USA, pp 166–177
13. Markou M, Singh S (2003) Novelty detection: a review—Part 1: statistical approaches. Signal Process 83(12):2481–2497
14. Mascaro S, Nicholson A, Korb K (2011) Anomaly detection in vessel tracks using Bayesian networks. In: Proceedings of eighth UAI Bayesian modeling applications workshop. Barcelona, Spain, pp 99–107
15. Miller S, Miller W, McWhorter P (1992) Extremal dynamics: a unifying physical explanation of fractals, 1/f noise, and activated processes. J Appl Phys 73(6):2617–2628
16. Osborne M (2010) Bayesian Gaussian processes for sequential prediction, optimisation and quadrature. University of Oxford, UK, pp 49–54, pp 79–90

17. Pinheiro J, Bates D (1996) Unconstrained parameterizations for variance-covariance matrices. Stat Comput 6(3):289–296
18. Porter M, Onnela J, Mucha P (2009) Communities in networks. Notices Am Math Soc 56(9):1082–1097
19. Psorakis I, Roberts S, Ebden M et al (2011) Overlapping community detection using Bayesian non-negative matrix factorization. Phys Rev E 83(6):066114
20. Psorakis I, Rezek I, Roberts S et al (2012) Inferring social network structure in ecological systems from spatio-temporal data streams. J R Soc Interface 9(76):3055–3066
21. Rasmussen C, Williams C (2006) Gaussian processes for machine learning. MIT Press, USA
22. Reece S, Roberts S (2010) The near constant acceleration Gaussian process kernel for tracking. IEEE Signal Process Lett 17(8):707–710
23. Reece S, Mann R, Rezek I et al (2011) Gaussian process segmentation of co-moving animals. In: Proceedings of AIP conference proceedings. Chamonix, France, pp 430–437
24. Rhodes B, Bomberger N, Seibert M et al (2005) Maritime situation monitoring and awareness using learning mechanisms. In: Proceedings of military communications conference. Atlantic City, NJ, USA, pp 646–652
25. Roberts S (2000) Extreme value statistics for novelty detection in biomedical signal processing. In: Proceedings of first international conference on advances in medical signal and information processing. University of Bristol, UK, pp 166–172
26. Simpson E, Roberts S, Psorakis I et al (2013) Dynamic Bayesian combination of multiple imperfect classifiers. In: Guy T, Karny M, Wolpert D (eds) Decision making and imperfection. Springer, New York, pp 1–35
27. Will J, Peel L, Claxton C (2011) Fast maritime anomaly detection using Kd-Tree Gaussian processes. In: Proceedings of IMA maths in defence conference. Swindon, UK

Author Biographies

Mark Smith received the B.Eng. Degree (First class honours) from the University of Wales Bangor, United Kingdom in 2007. He currently works for ISSG, Babcock International Group and is concluding an M.Sc. by Research at the Department of Engineering science, University of Oxford, United Kingdom. His research interests include embedded development, data mining and machine learning.

Steven Reece received the mathematics and physics B.Sc. (Hons) from Liverpool University, United Kingdom in 1988, part III of the mathematics tripos and the computer science diploma from Cambridge University in 1989 and 1990, respectively, and the DPhil in engineering science from Oxford University in 1998. He has published over 50 papers in Artificial Intelligence and machine learning and has six patents. He is a research fellow in the Engineering Department at Oxford University. His current research interests are in data fusion and multi-sensor systems.

Stephen Roberts is Professor of Machine Learning in the Department of Engineering Science, University of Oxford. He studied Physics at the University of Oxford (1987) and after a period of industrial research, he returned to Oxford and was awarded his PhD in 1991. He was appointed to the faculty of Imperial College, London, in 1994 and took up his present position in 1999. His main research interests lie in the application and development of mathematical methods in data analysis and data-driven machine learning, in particular statistical learning and inference and their application to complex problems in heterogeneous information fusion. His early contributions to the field include Bayesian models for real-time data modelling and signal processing. More recent research has focused on nonparametric Bayesian models for multi-sensor data fusion, global optimisation, complex systems, game theory and network analysis. Particular emphasis is placed on the real-world applications of advanced theory and over many years, he has applied these statistical methods to diverse problems in astrophysics, biology, finance and engineering. He has some 230 publications and holds eight patents.

Ioannis Psorakis is currently concluding his DPhil degree at the University of Oxford, where he attended as a Microsoft Research PhD Scholar. He holds a B.Eng. from the Technical University of Crete (Greece) and an M.Sc. (with Distinction) in Information Technology from the University of Glasgow. His research interests include social network analysis, data mining and scalable machine learning.

Iead Rezek is a research scientist at Schlumberger Research, Cambridge, United Kingdom. He received the Dipl-Ing (FH) degree from the Fachhochschule Ulm, Germany, and the B.Sc. from the University of Plymouth, United Kingdom in 1991, both in electrical engineering. He received his M.Sc. in physical sciences and engineering in medicine in 1992, and his PhD in information theory applied to biomedical data analysis in 1997, both from Imperial College, London, United Kingdom. In 2000, he joined the University of Oxford in 2000 as postdoctoral research assistant and subsequently held lectureship positions at Aston University and Imperial College London. His main research interests lie in modelling the interactions in human and biological networks. He is a member of the Royal Statistical Society and a member of the RSS statistical computing committee.