5

Case Studies of Medical Monitoring Systems

5.1 Introduction

Medical patient monitoring is a unique application domain for condition monitoring, with many constraints that arise due to the characteristics of human physiology, and the manner in which humans interact with the sensors that observe them; for example, the default action of human physiology is homeostasis, in which the system actively seeks to restore itself to normal operation. However, there are also many commonalities with other application domains for condition monitoring, and this chapter introduces several case studies to highlight how condition monitoring systems can be considered to be common across applications.

One of the defining characteristics of the medical monitoring domain is the conservative nature of clinical practice in the adoption of new technologies. The direct result of this is that many systems in use today are relatively straightforward. For example, Figure 5.1 shows the standard time-series of heart rate values (HR, measured in beats per minute, bpm) acquired from a bedside monitor. Clinicians are typically given the option of a number of constant thresholds for each vital sign; the monitor will then generate an alert automatically if the threshold is exceeded. Perhaps unsurprisingly, this results in an extremely large false-positive rate for conventional patient monitors, because artifacts appearing in the time-series cause the constant alerting threshold to be breached. Artifacts are common in medical monitoring, due to the movement...
of the patient and the sensors with which they are being monitored. Some hospital-based studies report that approximately 85% of all alerts from such monitors are false alarms [1–3].

The conservative nature of clinical practice is, arguably, a reflection of the fact that only those systems that are maximally reliable, and entirely predictable in their operation, should be permitted for use in medical monitoring. Many large manufacturers of medical monitors therefore adopt strategies similar to those described above, such that alerts can be directly associated with patient physiology. For example, when an alert based on high heart rate occurs, it must have been generated straightforwardly because the heart rate exceeded the univariate threshold of (for example) 140 bpm.

The obvious disadvantage of using such univariate strategies for alerting is that the high false-positive rate corresponds to most alerts being ignored in clinical practice—this limits the value of the alarm, and can cause unnecessary concern for patients and families. The use of such systems is familiar to many clinicians, who may be accustomed to assigning scores to various vital-sign observations made periodically by nursing staff, and where a review of the patient is initiated if the sum of these scores exceeds some predefined thresholds. These manual systems are often called track-and-trigger systems or early-warning systems [4–7], and have been mandated for use in many healthcare systems, when recording manual observations of the vital signs [8].

There is subsequently a need to perform intelligent health monitoring in a manner that integrates the available information across a number of time-series (here, the vital signs or other physiological data).

### 5.2 Kernel Density Estimates

Methods based on more advanced strategies for alerting have made their way into clinical practice, with the first generation being typified by the use of kernel density estimates (KDEs). Here, a representative training set of multivariate physiological data (typically heart rate, respiratory rate, blood pressure, and blood oxygen saturation) is collected from clinical monitors, and used as the basis for an alerting system. After appropriate normalization of each variable, the probability density function (PDF) of the training set is then estimated via the KDE. As in many application domains, this PDF of the training set then serves as a model of normality M against which previously-unseen data may be compared in real-time. The likelihood of the new multivariate data \( x \) may be evaluated \( p(x \mid M) \), with a novelty score thus defined as \( z(x) = -\log p_x \); as likelihood decreases with respect to the model (corresponding to more abnormal physiology) then \( z \) takes increasingly large values. When \( z(x) \) exceeds some threshold, then an alert is subsequently generated. This approach is one means
Figure 5.1 Most conventional medical monitors in use today provide simple threshold-based alerts to clinicians; here, a time-series of heart rate values is shown along with three pairs of upper- and lower-thresholds (shown by the green, orange, and red horizontal lines) [AU: edit out colors].
of performing novelty detection [9–12], which is well-suited to the monitoring of systems in which abnormal data are scarce with respect to the much larger number of data available from normal conditions. The advantage of using the KDE-based approach over conventional single-variable alerting is that the multivariate correlation between variables that form the multivariate point \( x \) are taken into account, rather than being treated independently. This means that variables might appear in abnormal combinations sufficient to raise an alert, but which might not be sufficiently extreme to break the univariate thresholds described in the previous section.

For example, a heart rate value of 124 bpm might not be sufficiently high to exceed a fixed univariate threshold (set at 140 bpm, for example); however, it might be sufficiently abnormal when observed with the low value of respiratory rate of 8 rpm (respirations per minute), and the KDE-based alert would then be generated.

Within the medical domain, alerting systems based on this approach have been deployed in a number of settings, from the monitoring of patients in acute wards of hospitals [13–14] to monitoring in the emergency department [15, 16]. Within the former, we monitored patients who were recovering from cancer surgery, and where there is a very high incidence of mortality (approximately 15%); the goal was to identify which patients were at risk, based on regular measurements of heart rate, blood oxygen saturation, and respiratory rate. As described above, a KDE was constructed using vectors of these three variables, acquired for a representative population of patients. Novelty scores \( z(x) \) could then be produced for all subsequent measurements of these variables, for new patients under observation.

An example is shown in Figure 5.2, in which the vital signs for a patient are shown plotted on the same set of axes, along with the time at which an escalation of care occurred due a perceived deterioration in the patient’s condition (approximately half-way through the period shown). Figure 5.2 also shows corresponding novelty scores \( z(x) \) with respect to a KDE-based model of normality; it may be seen that, while novelty scores increase around the time of the escalation of care, there is also an increase in novelty scores some time before the escalation. This demonstrates that novelty scores have the capacity to reflect abnormal data acquired from the patient, and which provide a principled means of assessing the health status of a patient, in comparison with the assumedly-representative population of patients used to construct the model from which the scores are produced.

The effect of homeostasis is evident from Figure 5.2: after an increase in novelty scores before the clinical escalation, the novelty scores then decrease as the patient’s physiology works to restore itself to normality. This is a unique factor that must be taken into consideration when performing monitoring of pa-
Figure 5.2  The first generation of medical monitoring systems based on statistical methods used kernel density estimation. Vital signs (heart rate, respiratory rate, blood pressure, and oxygen saturation shown in green, yellow, red/black, and blue, respectively) and corresponding novelty scores (VSI) are shown in the upper and lower plots, respectively. An escalation of patient care occurred at the time indicated by the vertical line. Approximately one hour of data is shown in the figure.
patients, because their bodies have the capacity to (attempt to) restore themselves to a stable condition.

However, in the example shown in the figure, the novelty scores then rise (towards the middle of the interval shown) as the patient’s condition subsequently deteriorates to the point of clinical emergency.

The training of such systems typically follows a cross-validation approach, in which optimal values for model parameters are determined by evaluating the fit of the model to held-out data from the training set, and subsequently choosing those parameter values that best fit the held-out data [17]. Such a procedure typically involves evaluating the performance of the model for a range of parameter values. An example of this approach is shown in Figure 5.3, in which a grid search is performed for finding the values of two model parameters; the performance of the model when applied to held-out validation data is evaluated.

Figure 5.3 Training for many black box systems relies on the use of cross-validation to determine assumedly-appropriate values for the various model parameters; fit to held-out cross-validation data for each of two model parameters are shown using the scale to the right.
for pairs of such values. The figure shows the fit of an example patient-based model to held-out data, where it may be seen that increases in the fit (shown by the red region) occur for one particular part of the space of possible parameter values.

The figure also highlights the potential disadvantages of using such a process for determining the optimal values of model parameters. The optimal value may not be within the range of the grid-search—in the example shown in the figure, the region of high goodness-of-fit is on the edge of the space of values that have been evaluated. It is possible that parameter values corresponding to regions beyond the upper limit of the vertical axis, in this example, may be associated with higher goodness-of-fit values than those shown. Therefore, one might extend the reach of the grid-search to include such regions, to determine whether or not the maximum goodness-of-fit has been found. An additional disadvantage of such an approach is that there may be no unique best values for the model parameters; that is, there may be different sets of parameter values that describe the data equally well—the parameter space plotted in Figure 5.3 might have multimodal peaks in fitness. Techniques described later in this chapter offer a means of coping with such multimodality in the parameter space.

5.3 Extreme Value Statistics

Extensions for the above method of using novelty scores have been proposed, in which extreme value theory (EVT) is used to adjust the novelty threshold according to the quantity of data that has been observed [18–20]. EVT is a branch of statistics that is commonly employed in fields such as meteorology, financial risk-prediction, and other such applications in which extreme data are of primary interest.

The fundamental assumption behind EVT is that, given data \( x \) drawn from some underlying distribution \( p(x) \) then we wish to infer characteristics of the data beyond the area of the data space that is occupied by the majority of our data; that is, we wish to perform, in a principled manner, extrapolation into the extremal areas of the data space, where there is little or no representation from the training data.

This intuition fits well with the novelty detection approach, where the majority of our data comes from, in the medical domain, patients who are physiologically stable. Given the variability between patients, and the complexity of the physiological mechanisms that give rise to the observed data (such as the vital signs), explicitly modeling the extreme regions of the data space is challenging. EVT is one means by which this problem has been tackled in the medical domain, and which this section introduces.
5.3.1 Type-I EVT

Extreme value theory is typically partitioned into two main branches, both of which will be described here. The first (which we will call type-I EVT) is that of extrapolating into extreme areas of data space based on understanding the properties of the distribution of stable data. That is, we aim to construct a model of the distribution of stable data \( p(x) \), as accurately as possible, and then use type-I EVT to yield a distribution that characterises the following: if we were to draw \( m \) data from \( p(x) \), then what is distribution of where we expect the maximum of those \( m \) data to occur?

Type-I EVT is supported by a limit theorem (the Fisher-Tippett theorem) that states that distribution of the maxima of those \( m \) data will tend towards one of three known families of probability distribution, depending on the form of the underlying distribution of stable data \( p(x) \). These three distributions for the maxima are the Weibull, Frechet, or Gumble distributions. For example, if the distribution of stable data \( p(x) \) is of the Gaussian family, then the distribution of the maxima of \( m \) samples drawn from that distribution is of the Gumbel family. (In passing, we note that these three families for the distribution of the maxima are themselves subtypes of the generalized extreme value, or GEV, distribution.) Once we know the parameters of \( p(x) \), then we can determine the parameters of the corresponding extreme distribution; for example, if \( p(x) \) is Gaussian, then once we know the mean and variance of \( p(x) \) then we can directly compute the two corresponding parameters of the Gumbel distribution.

One of the first forays into the combination of EVT and Bayesian statistics in the medical domain made the strong assumption that \( p(x) \) was of Gaussian form, and which then learned the distribution over the mean and variance of that distribution in a Bayesian manner. (That is, a parametric Bayesian approach was taken.) This is illustrated in Figure 5.4, which shows the corresponding PDF over the mean (shown on the horizontal axis) and precision (inverse of the variance, shown on the vertical axis). Standard Bayesian statistics yield this distribution as being normal-gamma, which is shown by the contours in the figure—it may be seen that the normal-gamma is Gaussian in horizontal slices of the contour plot, and gamma in vertical slices. (Note that the distribution over the precision of \( p(x) \), and thereby the distribution over the variance, must be positive, because precision and variance cannot be negative.)

The figure shows an increasing number of observed data, plotted as \( x \) markers on the horizontal axis of each subplot; the first subplot shows no observed data, and thus corresponds to the prior distribution over the mean and variance of \( p(x) \). The use of a prior is a useful regularizing tool in statistics, which controls the subsequent distribution by a priori weighting it away from regions that are unlikely to correspond to regions of high likelihood. In the case shown in Figure 5.4, this prevents the precision and variance from values that are too large, and therefore constrains the subsequent distribution to those regions that
are deemed appropriate a priori. The setting of priors (sometimes referred to as the elicitation of priors) is a task that typically takes place in collaboration with experts in the particular application. For example, clinicians will have a reasonable understanding a priori of which values each of the vital signs can plausibly take, and this may be encoded by the prior distribution.

The figure shows that, as more data is observed, so the distribution over the mean and variance becomes more tightly peaked; that is, as we observe more data, we become more certain about the value of the mean and variance of the distribution $p(x)$ that is assumed to have generated our data.

Figure 5.4 shows, for each subplot, a number of highlighted locations using colored circles; each of these corresponds to an example (mean, variance) pair of parameters. For each of these pairs of parameters, Figure 5.5 shows the corresponding distributions $p(x)$. It may be seen from the subplots that, as increasing data is observed, then the distributions $p(x)$ that could have generated the data tend to become more similar. This directly corresponds to the conclusion drawn from Figure 5.4, in which our certainty in the values of the parameters of $p(x)$ increases with the number of data.

Figure 5.6 then shows an example scheme by which this evolving set of distributions may be sampled. Many such sampling schemes exist in the literature, including well-known examples Markov Chain Monte Carlo (MCMC), with the commonly-used algorithms of Metropolis-Hastings and Gibbs sampling. Figure 5.6 shows an ad hoc method, in which the shape of the distribution is well-characterized, leading us to be able to sample equi-probabilistic regions of the distribution. Each of the red markers shown
Figure 5.5  Distributions corresponding to the four highlighted pairs of parameters from Figure 5.4. Data are, as in Figure 5.4, shown on the horizontal axis (red x). (AU: edit out color)

Figure 5.6  Parameter space for the univariate Gaussian (mean and precision on horizontal and vertical axes, respectively), with increasing likelihood shown by increasing shade towards black. Samples from the parameter space are shown by red markers (AU: edit out color).
in the figure thus corresponds to a pair of (mean, variance) parameter values of approximately equal likelihood; these are then combined with type-I EVT to find the corresponding values of the Gumbel distribution that describes where the most extreme of \( m \) data generated from \( m \) will lie. This result yields a family of Gumbel distributions, which are then used to set a novelty threshold on the cdf associated with the Gumbel distribution, just as one would set the novelty threshold using \( p(x) \) in the standard case described in Section 5.2.

Note that in the literature the use of EVT has been limited to data of low dimensionality—typically univariate data. However, many settings for novelty detection require the analysis of multivariate data. A useful extension of the type-I EVT method in the medical domain was developed [8–9] in which \( p(x) \) was permitted to be a full KDE. This allowed both multivariate and multimodal data to be modeled, as described in Section 5.2 for the KDE. EVT was then applied to the likelihood values

\[ y = p(x) \]

yielding a new univariate PDF \( g(y) \)—this latter PDF is of the Gumbel family, if the kernel used in the KDE is the Gaussian (which was the case described in section 5.2). Figure 5.7 shows the resulting distribution \( g(y) \) over the extrema of \( p(x) \), which can be seen to be highly irregular when plotted in the original data space. The KDE captures complex behavior of the original data, and this extension of type-I EVT results in distributions \( g(y) \) that occupy the tails of \( p(x) \) when plotted in the data space. Figure 5.7 shows both the results of sampling experiments (leftmost) and the distributions \( g(y) \) predicted by the extended theory described above, where it may be seen that the distributions match (up to the sampling quantization of the leftmost plot). This approach was particularly helpful for patient monitoring in the medical domain, because it allowed, for

![Figure 5.7](image-url)
the first time, principled EVT-based methods to be applied to complex, multivariate data such as the vital signs.

5.3.2 Type-II EVT

The second variety of EVT does not seek to model the distribution of stable data \( p(x) \), as was the case with type-I EVT; instead, type-II EVT directly models the extreme (or tail) data that may be available.

This is achieved by defining some arbitrary threshold \( u \) on the data. Figure 5.8 shows an example in which this threshold has been set to take a high value, above which a small number of data may be seen—these data are typically termed the exceedances in the statistical literature. Type-II EVT is supported by a related limit theorem (the Pickands theorem) that describes how exceedances above some extreme threshold \( u \) will tend towards a known distribution—the generalized Pareto distribution, or GPD.

We emphasise here that type-II EVT is modeling all of the available extremal data directly, and uses the fact that these extremal data will tend to be distributed according to the GPD to determine the relative abnormality of extremal data. This approach has disregarded the distribution of the bulk of stable data \( p(x) \), working on the intuition that, if we seek to characterize the extreme tails of a distribution, then we can model those tails directly. This is in contrast to type-I EVT, which modeled \( p(x) \) and then extrapolated directly into the tail to find where the maxima of \( m \) observations drawn from \( p(x) \) should lie.

The main challenge with type-II EVT is determining the value of the threshold \( u \). As this value is increased, the exceedances above that threshold tend more closely to be distributed according to the GPD; however, as the threshold is increased, we have fewer and fewer data points from which to estimate the values of the parameters of that GPD. That is, as the threshold \( u \) is increased, we become more certain that the exceedances come from the GPD, but we are less certain about which GPD they came from (i.e., the values of the GPD’s parameters). This is the case illustrated in Figure 5.8. The threshold has

![Figure 5.8](image)

**Figure 5.8** Time-series data (left) with an example threshold \( u \), above which few data-points fall. Corresponding empirical distribution (right) with threshold \( u \) shown for comparison.
been set to take a high value, and so we have 3 exceedances. This will mean that our estimate of the parameters of the GPD will be uncertain, when we try to fit the GPD to those 3 data.

The converse case is shown in Figure 5.9, in which the threshold \( u \) takes a low value. Here, the problems described for Figure 5.8 are reversed: we now have many exceedances from which to estimate the parameters of the GPD, but the data are less likely to tend to being distributed according to the GPD because the threshold \( u \) is too low.

The GPD itself is a flexible distribution, which represents the fact that it is the limiting distribution for exceedances that can take a number of different shapes. Figure 5.10 illustrates this by showing the GPD for example values of its shape parameter. Some of the GPDs have finite stopping points on the horizontal axis, which allows us to model values that cannot exceed some fixed ceiling. This is important in the medical domain, because our sensors are often limited to observing values within some finite range.

![Figure 5.9](image1.png)

**Figure 5.9** Time-series data (left) with an example threshold \( u \), above which many data-points fall. Corresponding empirical distribution (right) with threshold \( u \) shown for comparison.

![Figure 5.10](image2.png)

**Figure 5.10** Various examples of the GPD, with differing shape parameters.
As with type-I EVT, the type-II EVT method was extended to examine the tails of data that are derived from an arbitrary KDE $p(x)$, which may be multivariate and multimodal [10].

### 5.3.3 Gaussian Processes

A significant limitation of KDE-based approaches is that multivariate data is assumed to be IID (independent and identically-distributed); that is, successive multivariate data is evaluated independently from one another, without any notion of modeling the through-time behavior of the data.

More recent generations of research into medical monitoring systems have focussed on the use of dynamical systems that can take into account the full time-series of data acquired from sensors [21-23]. A commonly-used technique for such applications is the framework of Bayesian-Gaussian process regression.

A Gaussian process is a probabilistic means of performing regression, and is often used for situations in which the data are a time-series of values $\{x, y\}$. An example is shown in Figure 5.11, in which two time-series are shown for an exemplar patient: a time-series of blood oxygen saturation values (SpO$_2$, measured as a percentage) and a time-series of heart rate values (HR, measured in bpm, as before). The Gaussian process considers the distribution of $y$ (here, SpO2 or HR) at each value of $x$ (here, time). At any point in time $x$, the distribution of the variable $y$ is assumed to be Gaussian. This means that any vertical cross-section of the plots in Figure 5.11, taken at some time $x$, will correspond to a Gaussian distribution $p(y \mid x)$ – in the figure, we plot 2 standard deviations of that Gaussian distribution around its mean using the shaded region. The Gaussian process makes one other fundamental assumption: that these Gaussian distributions for $y$ occurring at each cross-section of the plot at time $x$ have a joint distribution, and that this joint distribution is multivariate Gaussian, $p(y \mid x) = N(\mu, \Sigma)$. The Gaussian distribution is analytically convenient: if a set of variables has a joint (multivariate) Gaussian distribution, then the marginal distribution of any subset of those variables will also be Gaussian. Hence, we may define a multivariate Gaussian over all $y$ values at all times $x$, which is a joint distribution, but where each cross-section of the variable $y$ at some particular time $x$ has a (univariate) Gaussian distribution – with some mean and variance related to the mean and covariance of the (multivariate) Gaussian distribution.

In Figure 5.4, we found the multivariate Gaussian that best describes the time-series sensor data in each example, and then used it to plot the (univariate) shaded regions—the latter correspond to confidence intervals around the mean of the data.

More formally, the Gaussian process is written

$$f(x) \sim GP\left(\mu_f(x), k(x,x')\right)$$
with some mean function \( \mu_f(x) \) which is often taken to be zero without loss of generality, and with covariance function \( k(x, x') \). The latter is a function that is used to compute the covariance matrix \( \Sigma \) that defines the multivariate Gaussian distribution over the time-series. This is another convenient property of the Gaussian process: while the covariance matrix \( \Sigma \) has \( n^2 \) entries in it (where \( n \) is the number of data), the covariance function that may be used to compute the entries of \( \Sigma \) may have far fewer parameters. For example, a commonly-used covariance function is the squared-exponential kernel
\[
k(x, x') = \sigma_f^2 \exp\left(-\frac{\|x - x'\|^2}{2\sigma_f^2}\right)
\]
which has just two hyperparameters, \( \Sigma_f^2 \) and \( \Sigma_l^2 \), which correspond to length-scales in the vertical (\( y \)) and horizontal (\( x \)) directions. As the value of the \( \Sigma_f^2 \) hyperparameter is increased, the Gaussian process is permitted to vary more widely in \( y \); as the value of the \( \Sigma_l^2 \) hyperparameter is increased, the Gaussian process varies more smoothly in \( x \). A second example is shown in Figure 5.12,

Figure 5.11 The Gaussian process provides a Bayesian nonparametric means of tracking a time-series, while providing uncertainty bounds on the posterior prediction, as shown here for (i) oxygen saturation and (ii) heart rate, in the upper and lower plots, respectively. Sensor values are shown by thick lines, manually-observed values from clinicians are shown as circles; uncertainty bounds on the posterior distribution are shown by the shaded regions.
again showing SpO\textsubscript{2} and HR, where it may be seen that the Gaussian process in the upper plot has a smaller value of $\Sigma f^2$ than that shown in the lower plot (because it varies less widely in the vertical direction). It may also be seen that the Gaussian process in the upper plot has a larger value of $\Sigma l^2$ than that shown in the lower plot, because it varies less rapidly in the horizontal direction.

The optimum value of the hyperparameters is typically found by evaluating the likelihood of the data with respect to the Gaussian process for varying values of its hyperparameters, but where the likelihood is penalized in a Bayesian manner such that over-fitting the data is avoided [12].

An advantage of the use of Gaussian processes is shown in Figure 5.12, in which it may be seen from the plots that temporarily artifactual excursions in the sensor data quickly exceed the confidence bounds defined by the shaded regions—these data therefore have very low likelihood with respect to the Gaussian process, and may be identified as being artifactual, for removal.

A second advantage of the use of Gaussian processes is shown in the upper plot of Figure 5.12, where it may be seen that the sensor time-series for oxygen saturation have a number of periods in which no values are present. This data

![Figure 5.12](image-url)
is typically acquired from a pulse oximeter, which is a relatively cumbersome sensor often worn on the finger of a patient. Such sensors are therefore often removed by patients, for the sake of comfort, resulting in periods of missing data. The figure shows that periods of missing data may be handled by the Gaussian process: the values of the time-series are interpolated during such periods, as shown by the dashed line (which corresponds to the mean of the Gaussian process as it varies through time). Importantly, confidence estimates are provided throughout such periods, again shown by the shaded regions. The figure shows that the interpolated estimates from the Gaussian process match independent data acquired manually by a clinician (where the latter are shown by circles that overlap with the shaded region from the Gaussian process).

Thus, the Gaussian process offers a robust means of coping with realistic sensor data, with all of the missingness and artifacts that typify sensor data in realistic medical settings. The results may subsequently be used for novelty detection, as shown in Figure 5.13. Here, the multivariate vital signs are shown along with their Gaussian processes, and where confidence intervals are shown for each. As with the KDE-based case, new observations can be compared to the Gaussian process, with a threshold placed on the resulting novelty scores, computed in the same way as with the KDE; except now the probability distribution represents the full dynamics of the time-series data, rather than the IID situation of the KDE-based approach.

Figure 5.13  Combination of vital signs may be performed using a multivariate Gaussian process. Here the vital signs are shown as acquired from sensors (solid lines) and from manual observations by clinicians (circles), with the uncertainty bounds shown by the shaded regions. Heart rate, breathing rate, blood pressure, and oxygen saturation are shown in green, purple, red, and blue, respectively. (AU: edit out color)
Multivariate data can be handled naturally using a Gaussian process framework, which also offers the facility to estimate correlation and phase lag between variables, and which has been demonstrated for several medical monitoring examples [24].

### 5.4 Advanced Methods

Approaches to medical monitoring previously described have concentrated on assessing new multivariate data $x$ as they are observed, with the aim of performing novelty detection. We have described how the KDE-based approaches adopt an IID approach, in which new point $x$ is tested for potential abnormality independently of all other data. We have described how the Gaussian process-based approaches improve on this situation by testing the new point $x$ using dynamical information contained in the time-series history of previous observations. This section presents a new approach, whereby an entire time-series is tested for novelty in one step, rather than individually testing the various points $x$ that comprise that time-series.

Noting that a Gaussian process is defined as a distribution over functions, as described in the previous section, we can refer to a time-series of data acquired in a health-monitoring setting as being a GP thus defined. As shown in Figure 5.14, the use of conventional thresholds to detect abnormal data can be problematical. The dynamics of the time-series are entirely ignored, as described earlier, and thus (in the figure) two time-series would be deemed to be abnormal by breaching the upper and lower thresholds for a short interval. However, we require a means of modeling the relative abnormality of this data in a manner that takes into account its dynamics. This is especially important in the case of modeling patients (in addition to other critical systems), where there is important information encoded by the dynamics of the data acquired.

As shown in Figure 5.15, we can use the Gaussian process framework to describe each of the time-series of data as being *draws from a single Gaussian process*. That is, each of the time-series is treated as a single independent entity, which is assumed to be a single draw from a Gaussian process. Using the Gaussian process framework, we can find the model that is most likely to have generated the time-series that we have collected. In the figure above, we show each of the time-series as a draw from a single Gaussian process model, constructed using maximum likelihood methods [14]. Effectively, each of the (potentially noisy) time-series data in the figure are now represented using a smooth function, which is itself considered to be a statistical draw from the Gaussian process model that we have constructed.

If the Gaussian process model is thus constructed using normal examples of time-series, then we might reasonably consider time-series that are unlikely
Case Studies of Medical Monitoring Systems

143

to have been generated from this model as being candidates of abnormal behavior. This is the exact analogue of the process described for the KDE previously, in which we have (i) a model of normality constructed from examples of normal system behavior, (ii) the ability to calculate novelty scores for new data with respect to that model, and (iii) where new data with high novelty scores are then deemed to be abnormal. The difference between the model used in this chapter and that described in the KDE is that we now have data that are entire time-series of data, whereas the KDE was used for the case when the data were individual points.

We previously touched on the notion of extreme value theory (EVT), which can be used to determine the probability of data being extreme with respect to some distribution describing normal behavior, where the distribution is a KDE (as described earlier). Similar to the manner in which EVT can be applied to a KDE, EVT can also be applied to the outputs of a Gaussian process [25]. This is the means by which we assign novelty scores $z(x)$ to entire time-series. An example is shown in Figure 5.16, in which we have constructed a (Gaussian process) model of normality based on a set of time-series data that describe different normal patients; the use of EVT then assigns scores (shown by the colors, in this example) to entire time-series.

Figure 5.14 The use of constant thresholds on the vital signs (or on novelty scores derived from them in a multivariate manner) can lead to abnormal dynamics in those signals being ignored; here, a number of time-series (coloured lines) are shown against a constant threshold (shaded region), within which signals are deemed to be normal, as would be used by conventional systems. (AU: edit out color)
The figure shows that most of the time-series shown have been colored according to their EVT scores that take a positive value below 1; the EVT score can therefore be interpreted as “probability that this time-series is drawn from the Gaussian process model-of-normality”. The fact that most of the time-series in the example have been colored with values approaching 1 suggests that most of the time-series shown are similar to those in the training set of normal data used to construct the model of normality. However, it may be seen that there is a single time-series that has been colored black, and which corresponds to a very low probability of being drawn from the (Gaussian process) model of normality. In fact, this is a highly abnormal time-series, with dynamics that are markedly different from those of normal patients, and which has therefore been correctly classified as being highly abnormal.

The means by which we evaluate such machine learning-based systems must next be considered, because we are using entire time-series to classify data, with the aim of identifying sections of data that correspond to subsequent abnormality. That is, we are attempting to provide early warning of future deterioration of the system’s health.

One means by which this may be performed is to trade-off the quantity of early-warning time given in advance of any labeled periods of actual abnormality ($t_{EW}$) versus the false-positive rate (FPR) of raising alarms for individuals
who are entirely normal. An example is shown in Figure 5.17, in which different values of a novelty threshold (on the EVT score shown in figure 5.9) lead to a varying trade-off between $t_{ew}$ and FPR. We note that it is desirable to provide a positive amount of early warning; i.e., $t_{ew} > 0$, which represents the fact that we are identifying periods of abnormality in advance of their occurrence (i.e., we are actually providing a degree of early warning). If $t_{ew} = 0$, this shows that we are providing no early warning, but are detecting abnormal events as they occur. If $t_{ew} < 0$, then we are identifying abnormality after it occurs (i.e., we are providing late notification of system abnormality).

The Figure 5.17 shows that if we require a very low FPR for the model in question, such that we produce very low numbers of false alarms for otherwise-normal data, then this corresponds to early-warning times that are negative—we are only detecting abnormality after it occurs. As we accept a higher false-positive rate, then the early-warning time that the model provides increases from being negative to being zero. (In the figure, this latter point occurs when the FPR is approximately 0.75.) As we accept an even higher false-positive rate, then we begin to increase the amount of (positive) early warning that the sys-
Figure 5.17  The means by which we evaluate a machine learning system suggests strong assumptions have been made for the construction of that system. Shown above, as an alternative to traditional ROC-based methods for training and evaluating a system, we show a plot of false-positive rate (horizontal axis) against early-warning time (vertical axis).

The figure shows that the maximum average early-warning time possible for the example model is a little over two hours in advance of failure.

References


