34th Annual International Conference of the IEEE EMBS
San Diego, California USA, 28 August - 1 September, 2012

Gaussian Process Regression in Vital-Sign Early Warning Systems

Lei Clifton*, David A. Clifton*, Marco A.F. Pimentel*, Peter J. Watkinson†, *, Lionel Tarassenko*

*Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, Oxford, UK
†Oxford University Hospitals NHS Trust, Oxford, UK

***Abstract*—The current standard of clinical practice for patient monitoring in most developed nations is connection of patients to vital-sign monitors, combined with frequent manual observation. In some nations, such as the UK, manual early warning score (EWS) systems have been mandated for use, in which scores are assigned to the manual observations, and care escalated if the scores exceed some pre-defined threshold. We argue that this manual system is far from ideal, and can be improved using machine learning techniques. We propose a system based on Gaussian process regression for improving the efficacy of existing EWS systems, and then demonstrate the method using manual observation of vital signs from a large-scale clinical study.**

***Index Terms*—patient monitoring, Gaussian processes.**

I. Early Warning Systems

Patient monitoring systems have been deemed to suffer from the "plague of pilots" [1], where many prototype systems have been developed for tracking patient condition [2]–[4], and yet little clinical evidence has been accumulated evaluating their efficacy at scale [5], [6]. The gold standard for the measurement and interpretation of vital signs remains manual observations made by clinical staff; although continuous, automatic systems have been developed for acquiring patient data, there is seen as being a lack of robustness in the manner in which the acquired data are subsequently processed and used to support clinical practice [7]–[9]. Therefore, despite the progress made in novel sensing devices, the standard of care in most hospitals involves manual periodic observation of patient vital signs.

Episodes of patient deterioration are frequently preceded by periods of derangement in the vital signs [10], and guidance has been produce in the UK mandating the use of manual *track-and-trigger* or *early warning score* (EWS) systems [11]. Such methods are typically paper-based, although some electronic systems do exist (as we will describe in our clinical study, later in this paper), and involve the assignment of univariate scores to each vital sign, with scores increasing from zero according to their "severity". If any of the scores assigned to each vital sign exceed a pre-defined threshold, or if the sum of all scores exceeds another threshold, then the patient is deemed to be in need of clinical review. The disadvantage of such methods is that the scoring systems are typically heuristic, and vary from hospital to hospital, and even from ward to ward [12].

This paper introduces the use of Gaussian process regression to set the analysis of patient physiology into a principled, probabilistic time-series framework, in order to augment the existing standard of care, in EWS systems, as described in II.

Automated approaches to the analysis of continuously-acquired data, rather than the EWS data considered by this work, have been proposed using a number of techniques. These include Kalman filtering [13], neural networks [14], and density estimation using the Parzen windows method [15]. The difficulty of such systems is in their handling of signal artefact, due to sensor noise and movement of the patient. The typical method of coping with such periods of artefact is to replace the noisy or incomplete episodes of data with the corresponding mean value of the vital sign over a population of patients [16]. We will demonstrate that this conventional method of mean-replacement biases analysis of the patient towards "normality", and that principled methods for coping with episodes of missing or incomplete data can provide earlier warning of physiological deterioration by avoiding this bias.

Section III presents initial clinical validation of the proposed method using manual observations acquired from a large clinical study that we have recently undertaken to provide much-needed evidence in favour of the use of machine learning methods in patient care. We conclude in section IV with suggestions for future work.

II. Gaussian Process Regression

Gaussian process regression (GPR) offers a framework for performing inference using time-series data, in which a probability distribution over a *functional space* is constructed. By considering the time-series of patient observations to be a function, we can perform inference upon them using the GPR framework. This is particularly suited to the analysis of data that may be sampled at irregular intervals, as with manual observation data.

We provide a brief overview to set notation, where more details may be found in [17]. For some observed dataset of physiological data over time intervals $\mathbf{X} = \{\mathbf{x}_i \mid i = 1 \dots m\}$, we define a GP prior distribution over latent (unobserved) functions $\mathbf{s} = \{s(\mathbf{x}_i) \mid i = 1 \dots m\}$, according to $s(\mathbf{x}) \sim \mathcal{GP}\left(\mu_s(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')\right)$, here using a squared-exponential covariance function $k(\mathbf{x}, \mathbf{x}') = \sigma_s^2 \exp\left(-\frac{\|\mathbf{x}-\mathbf{x}'\|^2}{2\sigma_l^2}\right)$, where $\|\cdot\|$ is the ℓ_2-norm, where σ_l and σ_s are hyperparameters giving the length-scale in the x-direction and the variance of s, respectively, and where the mean function $\mu_s(\mathbf{x}) = 0$. We define a set of observed target physiological data which are

assumed to be generated from some latent function $\mathbf{t} = \{s(\mathbf{x}_i) + \varepsilon \mid i = 1 \dots m\}$, with $\varepsilon \sim \mathcal{N}(0, \sigma_t^2)$ defining additive Gaussian noise over the latent function s.

For this prior GP distribution over functions, we may define the marginal likelihood (or "evidence") for some set of observed physiological data $\mathbf{t}$ given the set of inputs $\mathbf{X}$,

$$p(\mathbf{t}|\mathbf{X}) = \int p(\mathbf{t}|\mathbf{s}, \mathbf{X})\, p(\mathbf{s}|\mathbf{X})\, \mathrm{d}\mathbf{s} \tag{1}$$

in which we have marginalised over the function values $\mathbf{s}$, using the GP prior distribution over functions $p(\mathbf{s}|\mathbf{X}) \sim \mathcal{N}(\mu_s(\mathbf{X}), k(\mathbf{X}, \mathbf{X}))$, and where the likelihood of the observed targets $p(\mathbf{t}|\mathbf{s}, \mathbf{X}) \sim \mathcal{N}(\mathbf{s}, \sigma_t^2 \mathbf{I})$. The log marginal likelihood corresponding to the integral can be found in closed form as a marginalised Gaussian, thanks to the consistency property of the GP,

$$\log p(\mathbf{t}|\mathbf{X}) = -\frac{1}{2}\mathbf{t}^\top (K + \sigma_t^2 \mathbf{I})^{-1} \mathbf{t} - \tag{2}$$

$$\frac{1}{2} \log |K + \sigma_t^2 \mathbf{I}| - \frac{n}{2} \log 2\pi \tag{3}$$

where we have used the notation $\mathbf{K} = k(\mathbf{X}, \mathbf{X})$. We can perform a similar operation using posterior GPs, in which we have taken into account our observed physiological data over times $\mathbf{X}$ with their corresponding latent function values $\mathbf{s}$ and target observations $\mathbf{t}$. If we wish to evaluate a function over n test points $\mathbf{X}_* = \{\mathbf{x}_{*,i} \mid i = 1 \dots n\}$, we can apply a Bayesian formulation [18] to predict the n-dimensional vector of target physiological values $\mathbf{t}_*$ corresponding to the test inputs $\mathbf{X}_*$ using

$$p(\mathbf{t}_*|\mathbf{X}, \mathbf{t}, \mathbf{X}_*) = \int p(\mathbf{t}_*|\mathbf{s}_*)\, p(\mathbf{s}_*|\mathbf{X}, \mathbf{t}, \mathbf{X}_*)\, \mathrm{d}\mathbf{s}_* \tag{4}$$

$$p(\mathbf{s}_*|\mathbf{X}, \mathbf{t}, \mathbf{X}_*) = \int p(\mathbf{s}_*|\mathbf{X}, \mathbf{s}, \mathbf{X}_*)\, p(\mathbf{s}|\mathbf{X}, \mathbf{t})\, \mathrm{d}\mathbf{s} \tag{5}$$

In the above, we have the likelihood $p(\mathbf{t}_*|\mathbf{s}_*) = s(\mathbf{x}_*) + \varepsilon$ as before; we have the posterior GP $p(\mathbf{s}|\mathbf{X}, \mathbf{t})$, which we will consider below; and we have the joint posterior distribution over all functions, conditioned on the observed training data,

$$p(\mathbf{s}_*|\mathbf{X}, \mathbf{s}, \mathbf{X}_*) \sim \mathcal{N}(\mu_*, K_*) \tag{6}$$

where the mean function and covariance matrix in the above are, respectively,

$$\begin{aligned} \mu_* =& \mathbb{E}\left[\, p(\mathbf{s}_*|\mathbf{X}, \mathbf{s}, \mathbf{X}_*) \,\right] \\ =& K(\mathbf{X}_*, \mathbf{X}) \left[K(\mathbf{X}, \mathbf{X}) + \sigma_t^2 \mathbf{I}\right]^{-1} \mathbf{t} \end{aligned} \tag{7}$$

$$\begin{aligned} K_* =& K(\mathbf{X}_*, \mathbf{X}_*) - \\ & K(\mathbf{X}_*, \mathbf{X}) \left[K(\mathbf{X}, \mathbf{X}) + \sigma_t^2 \mathbf{I}\right]^{-1} K(\mathbf{X}, \mathbf{X}_*) \end{aligned} \tag{8}$$

Due to the convenient analytical tractability of the multivariate Gaussian distribution, the above integrals are determined in closed form.

For any given training set of physiological data $\{\mathbf{X}, \mathbf{t}\}$, we may therefore learn the posterior GP required in the above, $p(\mathbf{s}|\mathbf{X}, \mathbf{t})$, which is fully specified by its hyperparameters σ_l, σ_s, and σ_t. The values of these hyperparameters have, for the work described in this paper, been determined to be those values that maximise the log marginal likelihood of the targets, which incorporates a trade-off between model fit and model complexity [17]. We have used the squared-exponential covariance function for this preliminary proof-of-concept work.

A further advantage of a Bayesian framework is the explicit incorporation of uncertainty into the model, which allows us to take a patient-specific approach in which a GP is constructed using data from each individual patient. After estimation of the posterior GP, after observing some patient data $\mathbf{X}$, we can make predictions about the distribution of the latent function $\mathbf{t}$ corresponding to the time-series of the physiological data at any points in time $\mathbf{X}_*$.

III. Results

A. Clinical Study

We studied a group of 200 post-operative patients, during their recovery from cancer surgery in a step-down ward of the Oxford University Hospitals NHS Trust, Oxford. This study was approved by the local research ethics committee. The study involved the collection of manual observation of vital sign data, which form the current standard of care, in accordance with recommended clinical guidelines [10].

This group of patients has a high risk ($\approx 20\%$) of post-surgical complications, which typically result in an unexpected admission to the Intensive Care Unit (ICU), and which have a high associated risk of mortality. The goal of the EWS system is to track patient condition such that the early warning signs of deterioration may be identified and acted upon.

B. Data Collection

Manual measurements of heart rate (HR), breathing rate (BR), blood pressure (BP), and peripheral oxygen saturation (SpO_2) were made by the ward staff as part of the current standard of care. All manual observations were recorded using the paper-based system in use throughout the hospital at the time of the study. A subset of these observations of patient physiology were entered into a patient PDA, carried by the ambulatory patients, and which were then automatically transmitted to a central archive for storage.

One of the obstacles to the analysis of the very large datasets of patient physiology acquired from clinical studies is that the data are typically not in a form amenable to analysis, and exhaustive labelling for such quantities of data is typically hard to acquire. The 16,503 manual observations of vital signs from this study were transcribed into electronic form by two independent teams of research nurses, who were trained in data entry. Disagreements between the two datasets, along with obvious errors, were automatically identified by a reconciliation program, the results of which were presented to a third, independent team of adjudicators who compared both electronic transcriptions with the original paper-based data to provide a final, reconciled dataset over all 16,503 vectors of vital signs, for the 200 patients in the study.

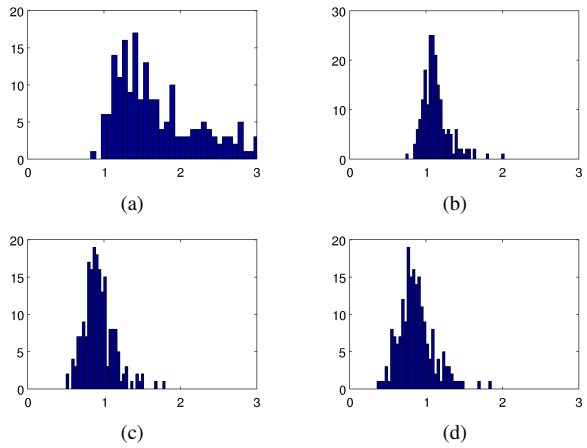

Fig. 1. Histograms of MSE for 200 patients obtained by handling periods of artefact and signal incompleteness using (a) replacement by the population mean; (b) replacement by the patient-specific mean; (c) GP regression; (d) SVR

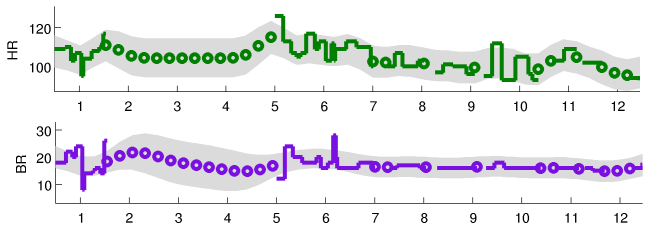

Fig. 2. Personalised GP showing output for the HR and BR time-series of a patient with a period of missing data. The mean function of the GP is shown during periods of incomplete data by the circles, with two standard deviation on the posterior distribution at each test point shown in grey. Manual observations are shown as solid lines.

C. Results from GPR

The datasets of manual observational data for each patient were randomly partitioned 75%:25% into training and test sets, respectively. Patient-specific GPs were constructed using the training set for each patient, where 10-fold cross-validation was used to determine the values of the GP hyperparameters (section II). For comparison, a support vector regressor (SVR) was also trained using the same data and cross-validation methodology; this latter is a kernel-based, non-probabilistic method that is popularly used for regression problems in the literature [19], and which is not described in further detail here for brevity.

The mean-square error (MSE) between each patient's set of test data and the regression of the personalised GPR and SVR models was then determined, and compared with MSE obtained by use of the means of each vital sign obtained from the the entire patient population (E_1) and the patient-specific mean (E_2). These latter two means represent the existing practice of replacing periods of missing data with patient means, which are typically population-based means (E_1). We have included the patient-specific mean (E_2) as a slight refinement of existing practice, for comparison with the principled GPR and non-probabilistic SVR methods.

TABLE I
MSE STATISTICS ON INDEPENDENT TEST DATA FOR MANUAL OBSERVATIONS OVER 200 PATIENTS

	E_1	E_2	E_{GP}	E_{SVM}
μ_E	1.89	1.14	0.95	0.89
σ_E	0.93	0.27	0.29	0.31
IQR_E	0.94	0.17	0.21	0.25

Figure 1 shows MSE values for the GP, the SVR, and use of the population- and patient-specific mean, where summary statistics for each are given in table I. The latter shows the mean MSE, one standard deviation in MSE, and the interquartile range in MSE over the 200 patients as μ_E, σ_E, and IQR_E, respectively, for all four methods. It may be seen that the personalised GP and SVR methods result in the lowest overall MSE on the independent test sets for each patient, over all 200 patients.

There are similarities between the formulations of the GP and SVR methods [17], but we note that the probabilistic approach of the former is preferable to that of the latter when dealing with time-series in which high levels of uncertainty must be treated in a principled manner. This advantage may be seen in figure 2 in which the GP predicts the *distribution* of missing physiological data, rather than simply selecting a point-estimate, as would the SVR. The improved MSE on the independent test sets for each patient demonstrates that both the personalised GPR and the SVR models can accurately model the behaviour of our physiological data acquired from manual patient observation. In periods of incomplete data, these accurate regressions can, therefore, be used to estimate the true value of the data and, in the case of the personalised GP, the distribution over those estimated values.

D. Improved Early Warning

This section presents a case study demonstrating the manner in which the GPR framework can be used to improve the early warning of patient deterioration associated with manual vital sign observations, when compared with the existing method of replacing periods of incomplete data with a (population- or patient-based) mean value.

Figure 3 shows time-series of manual observational data for a patient who suffered a critical event at $t \approx 4days$, which resulted in their unplanned admission to the ICU, and for whom we require early warning of physiological derangement. For some time prior to the event, however, records of physiological data were incomplete. The illustration shows the application of a patient-specific GPR model, with hyperparameters trained using data from the patient's previous days' stay on the ward (using 10-fold cross-validation, as before). The dynamics of this patient's vital signs have been learned by the GPR model, which effectively describes those sets of functions that, based on the patient's previous dynamics, best describe (to 0.95 probability) the expected distribution of the time-series function for each vital sign during the period of missing data.

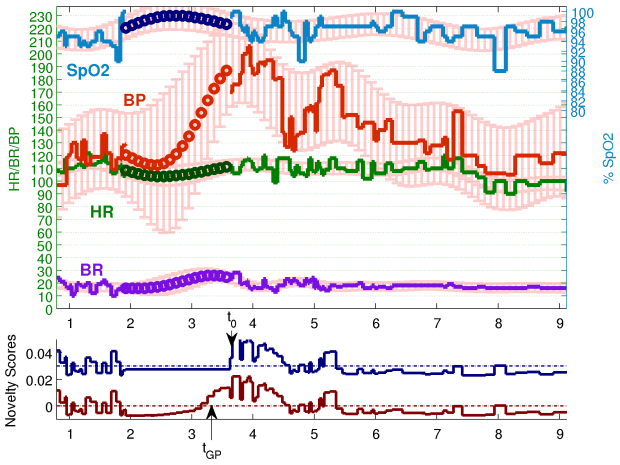

Fig. 3. Case study showing manual observational data (solid lines) from an example patient who suffered a critical patient event at around $t \approx 4$ days. Personalised GP regression mean-function outputs are shown by circles, with confidence intervals as coloured error-bars. The lower plot shows novelty scores from the method of [20] with and without personalised GP regression in red and blue, respectively, which have novelty alerts at times t_{GP} and t_0, respectively.

We have previously proposed a novelty detection method [20] which maps physiological vital-sign observations onto novelty scores, with higher scores indicative of "severity", a machine learning version akin to that of the EWS methodology describes earlier. The figure shows novelty scores output by this method when presented (i) with periods of missing data replaced by the mean of that vital sign in accordance with existing practice, and (ii) with periods of missing data estimated by the GPR for this patient. It may be seen that conventional method (i) causes a novelty alert to occur at $t_0 \approx 3.6$ days, shown in the figure. Use of the GPR method (ii) causes a novelty alert to occur at $t_{GP} \approx 3.2$ days, which is an increase in early warning of patient deterioration of over 9 hours, and which can aid greatly in interpreting patient condition.

This behaviour is caused by the GP providing robust estimates of physiological data prior to the event, which would otherwise have been replaced by the population mean for each channel of data, and which would therefore have caused the patient to look incorrectly "normal" for an extended period.

IV. Discussion

We have proposed a principled, probabilistic, patient-specific GPR method for modelling time-series of manual observational data, and demonstrated proof-of-concept using both MSE over 200 independent test sets (acquired from a clinical study) and in a case study, showing that novelty detection methods can benefit from using GPR methods to provide more reliable input data, that can better cope with periods of incomplete data. We note that this work has been retrospective, and that on-line use of the GPR should be formulated and tested in a clinical environment. However, the large scale of the clinical study described in this paper goes some way to addressing the lack of clinical evidence for the efficacy of machine learning methods in patient monitoring. Future work will concentrate on fusing manual observations with continuous data acquired from patient-worn body sensors.

Acknowledgements

LC was supported by the NIHR Biomedical Research Centre Programme, Oxford. DAC was supported by the Centre of Excellence in Personalised Healthcare funded by the Wellcome Trust and EPSRC under grant number WT 088877/Z/09/Z.

References

[1] A. Pantelopoulos and N. Bourbakis, "A survey on wearable sensor-based systems for health monitoring and prognosis," *IEEE Transactions on Systems, Man, and Cybernetics: Part C*, vol. 40, no. 1, pp. 1–12, 2010.
[2] S. Meystre, "The current state of telemonitoring: A comment on the literature," *Telemedicine and e-Health*, vol. 11, no. 1, pp. 63–69, 2005.
[3] V. Nangalia, D. Prytherch, and G. Smith, "Health technology assessment review: Remote monitoring of vital signs - current status and future challenges," *Critical Care*, vol. 14, no. 5, pp. 1–8, 2010.
[4] L. Tarassenko and D. Clifton, "Semiconductor wireless technology for chronic disease management," *Electronics Letters*, vol. S30, pp. 30–32, 2011.
[5] G. Clifford and D. Clifton, "Annual review: Wireless technology in disease state management and medicine," *Annual Review of Medicine*, vol. 63, pp. 479–492, 2012.
[6] S. Martin, G. Kelly, W. Kernohan, B. McCreight, and C. Nugent, "Smart home technologies for health and social care support," *Cochrane Database of Systematic Reviews*, vol. 4, pp. 1–11, 2008.
[7] N. Saranummi, "In the spotlight: Health information systems," *IEEE Reviews in Biomedical Engineering*, vol. 1, pp. 15–17, 2008.
[8] ——, "Mainstreaming mHealth," *IEEE Reviews in Biomedical Engineering*, vol. 4, pp. 17–19, 2011.
[9] S. Koch, "Home telehealth: Current state and future trends," *International Journal of Medical Informatics*, vol. 75, pp. 565–576, 2006.
[10] National Institute for Clinical Excellence, "Recognition of and response to acute illness in adults in hospital," Technical Report, 2007.
[11] National Patient Safety Association, "Safer care for acutely ill patients: Learning from serious accidents," Technical Report, 2007.
[12] L. Tarassenko, D. Clifton, M. Pinsky, M. Hravnak, J. Woods, and P. Watkinson, "Centile-based early warning scores derived from statistical distributions of vital signs," *Resuscitation*, vol. 82, no. 8, pp. 1013–1018, 2011.
[13] J. Quinn, C. Williams, and N. McIntosh, "Factorial switching linear dynamical systems applied to physiological condition monitoring," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 9, pp. 1537 –1551, 2009.
[14] J. Marcos, R. Hornero, D. Alvarez, I. Nabney, F. del Campo, and C. Zamarron, "The classification of oximetry signals using Bayesian neural networks to assist in the detection of obstructive sleep apnoea syndrome," *Physiological Measurement*, vol. 31, pp. 375–394, 2010.
[15] A. Hann, "Multi-parameter monitoring for early warning of patient deterioration," Ph.D. dissertation, University of Oxford, 2008.
[16] G. Clifford, W. Long, G. Moody, and P. Szolovits, "Robust parameter estimation for decision support using multimodal intensive care data," *Philosophical Transactions of the Royal Society, A*, vol. 367, pp. 411–429, 2009.
[17] C. Rasmussen and C. Williams, *Gaussian Processes for Machine Learning*. Cambridge, MA, USA: MIT Press, 2006.
[18] M. Kemmler, E. Rodner, and J. Denzler, "One-class classification with gaussian processes," in *Computer Vision ACCV 2010*, ser. Lecture Notes in Computer Science, R. Kimmel, R. Klette, and A. Sugimoto, Eds. Springer, Berlin, 2011, vol. 6493, pp. 489–500.
[19] B. Schölkopf and A. Smola, *Learning with Kernels*. Cambridge, Massachusetts, USA: MIT Press, 2002.
[20] D. Clifton, S. Hugueny, and L. Tarassenko, "Novelty detection with multivariate extreme value statistics," *Journal of Signal Processing Systems*, vol. 65, pp. 371–389, 2011.