Bayesian Gaussian Processes for Identifying the Deteriorating Patient

Glen Wright Colopy*, Marco A. F. Pimentel*, Stephen J. Roberts*, David A. Clifton*

*Department of Engineering Science, University of Oxford, Oxford, UK

Abstract

The step-down unit (SDU) is a high-acuity hospital environment, to which patients may be sent after discharge from the intensive care unit (ICU). About 1-in-7 patients will deteriorate in the SDU and require emergency readmission to the ICU. Upon readmission, these patients experience significantly higher mortality risks and lengths of stay. Gaussian process regression (GPR) models are proposed as a flexible, principled, probabilistic method to address the clinical need to monitor continuously patient time-series of vital signs acquired in the SDU. The proposed GPR models focus on the robust forecasting of patient heart rate time-series and on the early detection of patient deterioration. The proposed methods are tested with an SDU data set from the University of Pittsburgh Medical Center, comprising 333 patients, 59 of whom had at least one verified clinical emergency event. Results suggest that GPR-based heart rate monitoring provides superior advanced warning of deterioration compared to the current clinical practice of rules-based thresholding, and slightly outperforms the current state-of-the-art kernel density method, which requires 4 additional vital sign features.

I. Clinical Setting

An SDU manages the recovery of patients after ICU discharge while reducing the staff-intensive burden required for acutely ill ICU entrants. SDU patients are generally of a more stable condition than those in the ICU and therefore the SDU has a reduced nurse-to-patient ratio (1 nurse to 4-6 patients) compared to the ICU (1 nurse to 1-2 patients) [1]. Studies across different hospitals have estimated ICU readmission-rates of 4.2% - 7.6% [2], 8.8% [3], and 0%-18.3% [4]. Readmission to the ICU has significant clinical implications: mortality rates for readmitted ICU patients have been estimated at 40.2% [3] and 24.7% [2] (in contrast to 4.0% mortality of patients who were not readmitted). These high levels of mortality motivate the use of principled methods for identifying, and ideally predicting, physiological deterioration.

Current clinical practice involves manual calculation of rule-based risk scores, and simple thresholding on the absolute values of the vital signs. These heuristic thresholds are usually set according to clinical experience concerning the global population of stable patients. Empirical methods include the use of Extreme Value Theory (EVT) [5], or a Parzen window kernel density estimate (KDE) [6] over the vital signs of a population of healthy patients. The latter methods make strong, unrealistic assumptions about the nature of vital-sign data; notably, that the time-series data are independent and identically distributed (i.i.d.), which therefore ignores vital sign dynamics.

The GPR solution discussed in this paper provides an alternative approach to continuous monitoring that incorporates knowledge of the generative physiology and which models the dynamics of the time-series of the vital-signs. Prior work on GPR monitoring of vital signs can be found in [7], [8], [9].

II. Data

The data set used by this work comprises 333 adult patients in the surgical-trauma SDU at the University of Pittsburgh Medical Center Presbyterian Hospital. The patients were recruited in phase 1 of a 3-phase trial, approved by Institutional Review Boards, to optimise and validate the efficacy of the KDE-based monitoring system described in [6]. Phase 1 started in November of 2006 and lasted eight weeks. Patient heart rate (HR), respiratory rate (RR), blood-oxygen saturation (SpO_2), and systolic and diastolic blood pressure (SBP and DBP) were continuously recorded for the duration of stay. These data sets are not "open" data sets due to patient confidentiality. Over the course of phase 1 (for all 333 patients), the medical staff made only 7 clinical emergency calls based on extremal vital signs. In retrospect, clinicians labelled 112 clinical emergencies (C-events) when reviewing the time-series data of all patients, and those 112 C-events occurred for 59 patients. The presence of 112 emergency C-events when only 7 were called in practice supports the understanding that continuous monitoring can add value to the intermittent observation of nursing staff. The annotated C-events each had an associated start-time, stop-time, and primary cause.

Non-C patients (i.e. those patients who had no C-events) were divided into 3 groups of 89 patients, and with similarly-distributed lengths of stay. Two groups were reserved to explore modeling choices for this work. The remaining group of non-C patients served as negative-control cases, to use in conjunction with the 59 C-patients to determine the efficacy of the proposed methods for detecting physiological deterioration. Six non-C patients had insufficient data for consideration.

III. Gaussian Process Regression

For a useful introduction to GPR the reader is referred to [10]. A Gaussian-distributed random variable (RV), $y \sim N(\mu, \sigma^2)$, can be generalised to a multivariate normal (MVN) distribution over n-variables, $\mathbf{y}_{n\times 1} \sim MVN(\boldsymbol{\mu}_{n\times 1}, \boldsymbol{\Sigma}_{n\times n})$, in which $\boldsymbol{\mu}$ is the vector of respective means for each element in $\mathbf{y}$, and $\boldsymbol{\Sigma}$ is the covariance matrix, describing the pairwise correlation of each element in $\mathbf{y}$. GPR is a tool to estimate the unknown regression function $y = f(x)$, given a set of training data $\{x_i, y_i\}_{i=1..n}$, without recourse to an explicit parameterisation of $f(x)$. This is achieved by modeling $f(x)$ as a draw from a distribution over functions, that is, $f(\mathbf{x}) \sim GP(\mu(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$. Analogous to the MVN, $\mu(\mathbf{x})$ is the mean of the functional distribution evaluated at $\mathbf{x}$, and $k(\mathbf{x}, \mathbf{x}')$ is a positive-semi definite function outputting the pairwise covariances of $\mathbf{y}$ and $\mathbf{y}'$ (in the range) between the elements $\mathbf{x}$ and $\mathbf{x}'$ (in the domain). As a finite draw from the range of $f(\mathbf{x})$, the data $\mathbf{y}$ follow an n-dimensional MVN distribution, described above, with the elements of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ determined by the mean function and covariance function.

In the absence of prior information, the mean-vector $\mu(\mathbf{x}) = \mathbf{0}$ (via detrending the $\mathbf{y}$ values if necessary). A commonly used covariance function is the radial basis function (RBF) with additive white noise (WN):

$$k(\mathbf{x}, \mathbf{x}') = \underbrace{h^2 \exp\Big(\frac{-\mathbf{d}^2}{2\lambda^2}\Big)}_{RBF} + \underbrace{\sigma_n^2 \delta(\mathbf{x}, \mathbf{x}')}_{WN}$$

where $d = |\mathbf{x} - \mathbf{x}'|$, δ is the Kroenecker delta function, which is 1 when the inputs are identical. This function encodes that $Cov(\mathbf{y}, \mathbf{y}')$, decreases with separation in the domain. The hyperparameter h governs the magnitude of covariance, λ modulates the exponential decay of covariance, and σ_n^2 describes additive Gaussian noise, $N(0, \sigma_n^2)$ that corrupts each measurement of $\mathbf{y}$. The RBF encodes an infinitely differentiable function, which may be inappropriate to model functions with sharp discontinuities.

To fit a GPR model to the data, it is necessary either to estimate, or integrate over, the values of the hyperparameters. The log marginal likelihood (LML) is $\log p(y|x, h, \lambda, \sigma) = -\frac{1}{2}\mathbf{y}^T \boldsymbol{\Sigma}^{-1}\mathbf{y} - \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{n}{2}\log(2\pi)$.

Prior knowledge of the distribution of the hyperparameters can be incorporated into the LML. A maximum *a posteriori* estimate can be obtained by optimizing the LML with respect to the hyperparameters. A more "fully Bayesian" alternative is to integrate across a range of hyperparameter values, e.g. sampling via Markov Chair Monte Carlo (MCMC). Conditional on the hyperparameters and the training data $\{\mathbf{x}, \mathbf{y}\}$, the distribution of unknown targets $\mathbf{y}^*$ at locations $\mathbf{x}^*$ are also MVN with mean $E[\mathbf{y}^*] = \boldsymbol{\Sigma}^*\boldsymbol{\Sigma}^{-1}\mathbf{y}$, and variance $Var[\mathbf{y}^*] = \boldsymbol{\Sigma}_{**} - \boldsymbol{\Sigma}_*\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}^{*T}$, in which $\boldsymbol{\Sigma}^* = k(\mathbf{x}^*, \mathbf{x})$, and $\boldsymbol{\Sigma}^{**} = k(\mathbf{x}^*, \mathbf{x}^*)$.

IV. GPR Forecasting

The GPR kernel and priors over the hyperparameters were selected to forecast future HR values, given a window of observed HR values. HR measurements were acquired at $f_s = \frac{1}{3}$ Hz. As shown in figure 1, observed values within the most recent hour were down-sampled to $f_s = \frac{1}{60}$ Hz, measurements from 1-7 hours previous were down-sampled at $f_s = \frac{1}{120}$ Hz. To improve the precision of forecasting performance, measurements within forecast windows were not down-sampled. GPs were fitted to training data using the *GPStuff* implementation [11] of Elliptical Slice Sampling (ESS) MCMC [12] with 500 samples and with a 30-sample burn-in. ESS is a popular method for sampling the posterior distribution of a distribution based on a combination of MVN RVs with highly-correlated hyperparameters. Forecast performance was assessed by the likelihood of the true values, given the posterior GP, $N(\mathbf{y}^*|E[\mathbf{y}^*, Var[\mathbf{y}^*])$, described above. Covariance functions of varying complexity and priors over the hyperparameters were assessed according to forecast windows of 15, 30, 45, and 60 minutes beyond the interval of the training window. Training data comprised time series from a representative set of "normal" patient data, provided by the first two of the three groups of non-C patients.

The most robust kernel was found to be

$$k(\mathbf{x}, \mathbf{x}') = \underbrace{h_1^2\Big(1 + \frac{\mathbf{d}\sqrt{5}}{\lambda_1} + \frac{5\mathbf{d}^2}{3\lambda_1^2}\Big)\exp\Big(-\frac{\mathbf{d}\sqrt{5}}{\lambda_1}\Big)}_{Matérn\left(\frac{5}{2}\right)} + \underbrace{h_2^2 \exp\Big(-\frac{\mathbf{d}^2}{\lambda_2^2}\Big)}_{RBF} + \underbrace{\sigma_n^2\delta(\mathbf{x}, \mathbf{x}')}_{WN}.$$

The kernel above encodes the prior belief that longer trends (on the order of hours) are governed by the smooth RBF kernel, while minutely variations in HR are governed by a twice-differentiable Matérn(5/2) kernel. Measurements are corrupted by noise with a Gaussian distribution $N(0, \sigma_n^2)$. The priors placed over the hy-

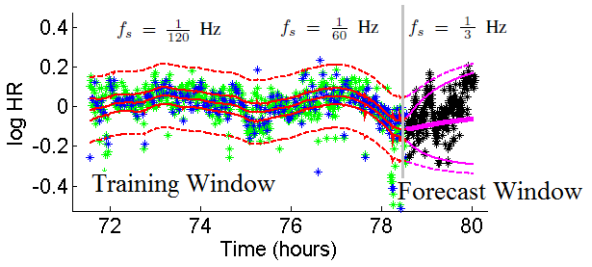

Figure 1. Training and forecast windows for GPR forecasting. Data in the most recently observed hour is sampled at a higher rate than earlier data. Green observations were removed during down-sampling leaving only blue observations with which to fit the posterior GP. The posterior GPR's latent mean and 95% CI are shown in red, along with the 95% CI of log HR measurements. A GPR forecast from the training window is in magenta. Black observations are the unseen forecast window, and seem to follow the trend of the forecast in this example.

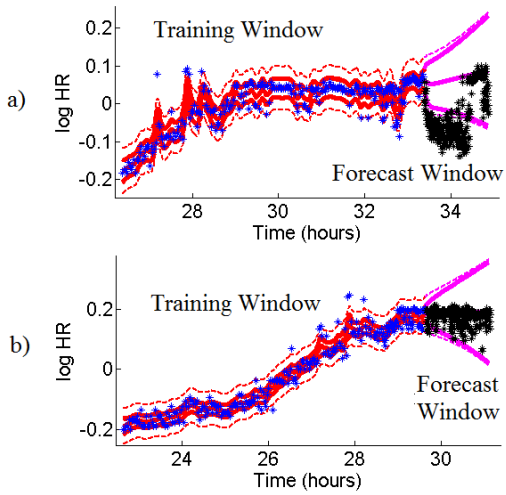

Figure 2. A step-change is detected by a marked divergence from the expectations of the posterior GP. Two time-series are presented, each with a posterior GP of the training window (red), and the forecast window (magenta). Solid lines (-) are the 95% CI of the latent mean, and dashed lines (- -) are 95% CI of the log HR measurements. The measurements in (a) exhibit precipitous (relative) tachycardia shortly after the forecast. The new observations fall outside the 95% CI so it is likely that that patient's physiology has undergone a rapid change. Notably, this deterioration is unlikely to be detected by thresholding or KDE methods because all measurements remain within clinically-common HR ranges. It is the time-series nature that makes these measurements unusual. In (b) the HR measurements remain well within the expected range of the posterior GP. The data in the forecast window are described well by the distribution of the posterior GP. The new measurements, therefore, do not suggest a change from the data generative process of the training window.

perparameters were $p(\ln h_1^2) \propto p(\ln h_1^2) \propto p(\ln \lambda_1) \propto p(\ln \sigma_n^2) \propto 1$ and $p(sqrt(\lambda_2)) \propto 1$.

V. GPR for Step-change Detection

A physiological "step-change" describes a marked discontinuity of new observations from previous observations. This concept has a useful application in novelty detection: if previous observations are the product of a "normal" generative process, then a departure from the expectations of the previous observations could be a departure from "normality". This is illustrated in figure 2, in which the posterior GP, fitted to the training window, produces a forecast of future measurements. If the observed measurements deviate sufficiently from the forecast, then an alarm can be generated. This method is a threshold on forecast-error, with the attractive hypothesis that deteriorating physiology may correspond to erratic time-series data and, hence, occur with low likelihood with respect to the forecast from "normal" conditions. The accuracy of a forecast can be described in terms of the original measurement units, likelihood, or via information theoretic measures.

VI. Experimental Design

Continuous monitoring was performed on all 59 C-patients (positive controls), and 89 non-C-patients (negative controls). A GP was fit to a 7-hour window of patient data, advanced every 5 minutes, which would be feasible for real-time monitoring, given the computational constraints of the clinical setting. Each newly-collected data point is compared to the posterior distribution to obtain a likelihood, given the posterior GP. Low mean log-likelihoods within a forecast window suggest a step-change. An alarm threshold [-20, 3] was set on the mean of the log-likelihoods for data in the forecast window. Lower thresholds produce a more sensitive alarm. To detect rapid fluctuations in HR, a forecast window of 2 minutes was used. (Results were found to be similar when the choice in the length of the forecast window was varied from 1-25 minutes.)

The outcome of interest is the accurate and early detection of C-events. In non-C patients, any alarm is considered a false positive. The false positive rate (FPR) is calculated to be the total number of false positives, divided by the total number of predictions performed on non-C patients (approximately 5500 in total). Since physiology may have been affected by clinical intervention after the first C-event, only the first C-event was considered for each patient. A credible interval for a true alarm was deemed to be from 8 hours prior until 2 hours after the C-event. The time of early warning (TEW) for each patient was calculated as described in figure 3.

VII. Results

The GPR step-change detector provided a significant improvement over simple thresholding and was comparable to the KDE method described in [6] (figure 4). Within a clinically-viable range of FPR values (0%-15%), a patient monitored by the GPR step-change detector

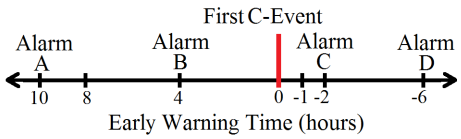

Figure 3. Timeline for TEW calculation for each C-patient. The timestamp of the first C-event is centered at $t = 0$, and the credible period for a clinical alarm is $t = [-2, 8]$ hours. Any clinical alarm outside of this credible region (e.g., alarm A or D above) was counted neither as an early warning nor a false positive (see the main text). The TEW for a C-patient was the first alarm to occur within the credible period; e.g., if alarms B and C both occurred, the TEW would be 4 hours. If no clinical alarm occurred in the 10-hour credible region, then the TEW value was right-censored at -2 for analytical convenience. Note that alarms for C-patients occurring outside this interval around an event were not deemed to be false-positive, because they could be associated with the event. False-positive alarms were defined using only the non-C patients, as described.

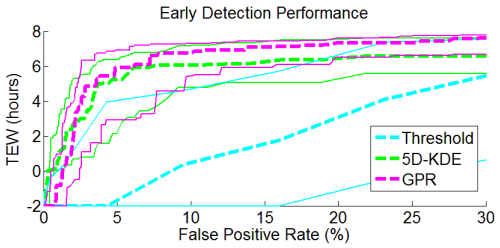

Figure 4. Performance in the early detection of deteriorating patients for three monitoring systems. A trade-off is made between the rate of false-positive alarms in 89 non-C patients, and the TEW in 59 C-patients. The bold dashed-lines represent median TEW performance, while the thinner solid lines are the 33% and 66% quantiles of TEW. The GPR step-change detector provides significant improvement over simple thresholding, which is the most common clinical practice. The KDE performs comparably, but requires 4 additional vital signs. By including these further vital signs in a GP framework, significant improvements over KDE methods may be possible.

would have 6-8 hours of additional advance warning in the event of deterioration, compared to the simple thresholding technique in current clinical practice. When accepting an FPR of 10%, the GPR-based method increased median TEW by over 6 hours. Two-thirds of patients could expect a longer advanced warning than the median patient under the simple thresholding technique. The KDE method, described in [6], making use of four vital signs in addition to HR, performed comparably to the GPR-based method for FPR 0%-5%, and had lower and more varied TEW for FPR 5%-30%. Notably, this performance includes all C-patients, the majority of whom did not have a clinically annotated HR emergency. This suggests the reality that deterioration can be manifest in several vital signs simultaneously. The GPR demonstrates the benefit of making fuller use of the information contained within a *single* vital sign, in that it outperforms a full 5-dimensional KDE using only a single vital, and provides a probabilistic means of performing forecasting without making the i.i.d. assumption of the KDE. The inclusion of further vital signs, and modelling the time-series correlation structure between them, offers the potential to produce significant advantages over the current state-of-the-art.

VIII. Future Work

The next step is to include further vital signs (RR, SpO_2, SBP, and DBP) into a multitask GP framework. The inclusion of additional physiological data offers additional potential to improve the detection of deterioration. To the extent that vital signs are correlated and provide further context, benign step-changes may become more predictable, and adverse step-changes may be more apparent.

Acknowledgments

GWC was supported by the Clarendon fund and EPSRC. MAFP was supported by the Wellcome Trust HAVEN project, WT 103703/Z/14/Z. DAC was supported by the Royal Academy of Engineering; Balliol College, Oxford; and an EPSRC "Challenge Award".

References

[1] K. Yousef, M. R. Pinsky, M. A. DeVita, S. Sereika, and M. Hravnak, "Characteristics of patients with cardiorespiratory instability in a step-down unit," *American Journal of Critical Care*, vol. 21, no. 5, pp. 344–350, 2012.

[2] G. S. Cooper, C. A. Sirio, A. J. Rotondi, L. B. Shepardson, and G. E. Rosenthal, "Are readmissions to the intensive care unit a useful measure of hospital performance?" *Medical Care*, vol. 37, no. 4, pp. 399–408, 1999.

[3] A. J. Campbell, J. A. Cook, G. Adey, and B. H. Cuthbertson, "Predicting death and readmission after intensive care discharge," *British Journal of Anaesthesia*, vol. 100, no. 5, pp. 656–662, 2008.

[4] A. Vlayen, S. Verelst, G. E. Bekkering, W. Schrooten, J. Hellings, and N. Claes, "Incidence and preventability of adverse events requiring intensive care admission: a systematic review," *Journal of Evaluation in Clinical Practice*, vol. 18, no. 2, pp. 485–497, 2011.

[5] D. A. Clifton, S. Hugueny, and L. Tarassenko, "Novelty detection with multivariate extreme value statistics," *J Sign Process Syst*, vol. 65, no. 3, pp. 371–389, 2010.

[6] A. Hann, "Multi-parameter monitoring for early warning of patient deterioration," Ph.D. dissertation, University of Oxford, 2008.

[7] L. Clifton, D. Clifton, M. A. Pimentel, P. Watkinson, and L. Tarassenko, "Gaussian processes for personalized e-health monitoring with wearable sensors," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 1, pp. 193–197, 2013.

[8] L. Clifton, D. A. Clifton, M. A. F. Pimentel, P. J. Watkinson, and L. Tarassenko, "Predictive monitoring of mobile patients by combining clinical observations with data from wearable sensors," *IEEE JBHI*, vol. 18, no. 3, pp. 722–730, 2014.

[9] O. Stegle, S. Fallert, D. MacKay, and S. Brage, "Gaussian process robust regression for noisy heart rate data," *IEEE Transactions on Biomedical Engineering*, vol. 55, no. 9, pp. 2143–2151, 2008.

[10] M. Ebden, "Gaussian processes: A quick introduction," arXiv:1505.02965v2.

[11] J. Vanhatalo, J. Riihimäki, J. Hartikainen, P. Jylänki, V. Tolvanen, and A. Vehtari, "Bayesian Modeling with Gaussian Processes using the GPstuff Toolbox," *ArXiv e-prints*, Jun. 2012.

[12] I. Murray, R. Adams, and D. MacKay, "Elliptical slice sampling," *Proc 13 International Conference AISTATS*, no. 9, pp. 541–548.