Bayesian Optimisation of 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*—Patient deterioration in the hospital ward is typically preceded by several hours of deranged physiology, as measured by the patient's vital signs. Estimation of the expected trajectory of a patient's future vital signs can allow us determine the degree of risk of physiological deterioration for that patient. Gaussian processes (GPs) offer a principled means of estimating vital-sign trajectories within a probabilistic framework.**

The automated estimation of GP parameters in this setting is difficult, due to the (often substantial) variation in physiology between patients, and also due to any changes in physiology that may occur for individual patients. Population-based techniques for fitting models to patient vital-sign data may be inferior compared with patient-specific approaches. We here propose the use of Bayesian optimisation to learn patient-specific models that are effective for estimating future physiological data, based on previously-observed data for the individual patient. We show how patient-specific values of GP hyperparameters may be learned using Bayesian optimisation, based on data observed during the first day of a patient's stay on an acute ward. We then demonstrate the benefit of using such methods in terms of forecasting accuracy for monitoring the patient during their subsequent two days on the ward.

I. Clinical Setting

In critical care wards, patient vital-sign monitoring typically involves continuous monitoring via a bedside monitor, supplemented by periodic manual observations made by a clinician. Patient deterioration detection is often performed by the bedside monitor, via applying a simple threshold to the individual vital signs (e.g., determining if heart rate, HR, falls below 40 or exceeds 160 bpm). Nurse observations typically result in the calculation of an "early warning score" that assigns scores to the individual vital signs, and where care is escalated if the total of these scores exceeds some preset threshold. Both bedside-monitoring and the use of manual early-warning scores have the disadvantages of performing inference using data from a single point in time; this thereby ignores any correlation between vital-sign measurements, and ignores any dynamics of those data. For example, measurements that are individually normal (e.g., HR between 50 and 90 bpm) may be *jointly* abnormal (e.g. HR = 50 bpm at 9.00 a.m. and HR = 90 bpm at 9.05 a.m.); existing systems would typically overlook the latter, leading to potentially unidentified episodes of physiological deterioration.

A particularly useful application of vital-sign time-series modeling is to compare a patient's current measurements to those values that are forecast based on a model trained using past data. If vital-sign predictions tend to be accurate for healthy patients, due to their stable physiology, then a patient who has deviated from the predicted values may be deteriorating, as is the data for the exemplar patient shown in figure 1(a).

Approaches to time-series analysis of patient data are hindered by a lack of robustness when fitting many time-series models. For example, the modelling approach may yield values of hyperparameters that are unrepresentative of the true dynamics of the patient's data, as shown in figure 1(b). We therefore require means of improving the robustness of fitting time-series models when considering patient physiological monitoring, to reduce the high false-alarm rate that may occur due to this confounding effect.

II. Data

Thirty-four patients from the University of Pittsburgh Medical Center are considered, using data collected in a previous clinical study [1]. The patients were selected as example of being "normal", due to having no clinically-validated emergency events and having at least 25 hours of continuously acquired HR measurements figure 2(a) and 2(d). HR measurements were downsampled to a frequency of one per minute. For each patient, the first 24 hours of HR data was used to learn a patient-specific regulariser for accurate HR forecasts, as described below, within the context of GP regression. The patient-specific regulariser will be used to assist in fitting a GP to the patient's HR time series over the subsequent 1-2 days figures 2(b) and 2(c).

For each minute of HR data at some time t, a GP is fit from the available data up to and including data at time t (as shown in figure 1), and a forecast of future values is made (figure 2(e)), as described below.

III. GPs for VS Monitoring

Useful reviews of applying GPs to time series data and optimization may be found in [2] and [3], respectively.

978-1-5090-4179-4/17/$31.00 ©2017 IEEE

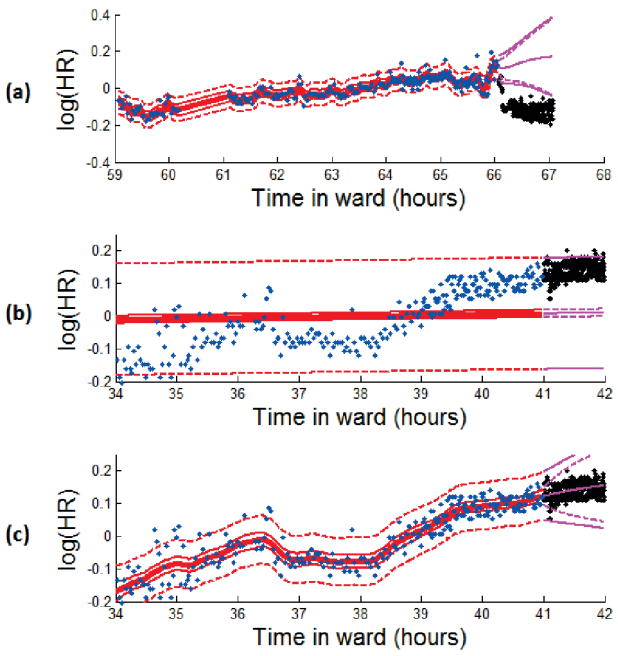

Figure 1. GP fitting and forecasting: Solid lines indicate the mean and 95% CI around $f(x)$, and dashed lines for the 95% CI around y. The magenta lines indicate a forecast beyond the currently available training HR data (in blue). Forecasting accuracy of future HR measurements (in black) can be poor for several reasons. In (a) the HR in the forecast window has a rapid and unforeseeable decrease. This is highly unusual physiology, and should be brought to the clinician's attention as evidence of patient deterioration. In the absence of deterioration, forecasts may still be inaccurate: (b) and (c) show GPs fit to the same HR time series. In (b) the value of the GP's length-scale hyperparameter is too large, so the upward trend in HR is not learned. The forecast accuracy in (b) is low but this could have been avoided via properly fitting the GP hyperparameters, as seen in (c), where GP hyperparameters capture key features of the patient's physiology, e.g. the long upward trend with short-term deviations. We seek to minimise the number of instances of fits like in (b) in order to avoid the false alarms on non-deteriorating patients.

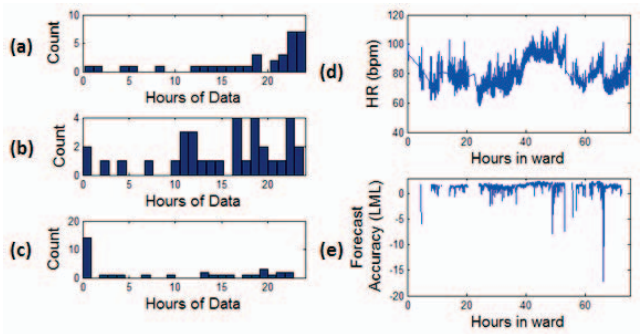

Figure 2. Plots (a), (b), and (c) show the distribution of the hours of data for each of the 34 patients across three days of monitoring. Most notably, two patients have less than two hours of available training data in the first day (a), while most have 12-24 hours over which to learn the patient's physiology. The majority of patients also have over 12 hours of data in the second day, but very little of the third day of data. (d) shows a typical time series of continuous HR data. The corresponding GP forecast accuracies are shown in (e). Note that forecast accuracy declines in the presence of rapid HR volatility.

All GP modelling was performed in GPStuff [4]. State-space representations of the GP were used for HR timeseries modelling to improve compuation time [5].

The GP models patient HR measurement, y_i, as a function of time, which is assumed to have Gaussian noise around some latent function f, such that $y_i = f(t_i) + \varepsilon_i$. A time series of HR, $\mathbf{y} = \{y_i\}_{i=1\ldots n}$ are then assumed to be jointly multivariate normal, where $\mathbf{y} \sim MVN(\mu, \Sigma)$ such that $\mu = 0$ (after detrending $\mathbf{y}$) and the covariance of any two measurements y and y' is a function of their respective time points, $k(t, t')$. Conditional on the observed HR measurements $\{y, t\}_{i=1\ldots n}$, HR values at new time points $\{\mathbf{y}^*, \mathbf{t}^*\}$ are also MVN such that $E[\mathbf{y}^*] = \Sigma^*\Sigma^{-1}\mathbf{y}$ and $Var[\mathbf{y}^*] = \Sigma^{**} - \Sigma^*\Sigma^{-1}\Sigma^{*T}$ where $\Sigma^* = k(\mathbf{t}^*, \mathbf{t})$, and $\Sigma^{**} = k(\mathbf{t}^*, \mathbf{t}^*)$.

A useful parametric form for the covariance function of HR is

$$k(\mathbf{t}, \mathbf{t}') = \underbrace{h_1^2\Big(1 + \frac{\mathbf{d}\sqrt{5}}{\lambda_1} + \frac{5\mathbf{d}^2}{3\lambda_1^2}\Big)\exp\Big(-\frac{\mathbf{d}\sqrt{5}}{\lambda_1}\Big)}_{\text{Matern}\left(\frac{5}{2}\right)} + \underbrace{h_2^2\exp\Big(-\frac{\mathbf{d}^2}{\lambda_2^2}\Big)}_{\text{RBF}} + \underbrace{\sigma_n^2\delta(\mathbf{x}, \mathbf{x}')}_{\text{noise}}. \quad (1)$$

where $d = |t - t'|$,and δ is the Kronecker delta function. This $k(\mathbf{t}, \mathbf{t}')$ encodes the prior belief that HR comprises short-term inter-beat variability, Matern($\frac{5}{2}$), a smoother long-term trend, RBF, and measurement noise (e.g., from unaccounted physiological variability or quantisation of measurements made by the bedside monitor).

When presented with HR data $\{\mathbf{y}, \mathbf{t}\}$, the hyperparameters of $k(\mathbf{t}, \mathbf{t}')$, which are $\theta = \{h_1, \lambda_1, h_2, \lambda_2, \sigma_n\}$, may be estimated or integrated across via the log marginal likelihood (LML) function $\log p(\mathbf{y}|\mathbf{t}, \theta) = -\frac{1}{2}\mathbf{y}^T\boldsymbol{\Sigma}^{-1}\mathbf{y} - \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{n}{2}\log(2\pi) + p(\theta)$, where the first three terms correspond to the MVN probability distribution, and the final term is the prior distribution over θ.

Priors over θ typically correspond to domain knowledge. For example, λ_1 and λ_2 are length-scale hyperparameters (associated with the Matern and RBF kernels, respectively) that model the length of time over which $\mathbf{y}$ and $\mathbf{y}'$ will remain correlated. If observations are expected to decorrelate after 1 hour, then λ_2 would be reasonably expected to be on the order of 1 hour.

In practice, uninformative priors $p(\theta)$ are often used to accommodate the variety of physiology that the GP may encounter from patient to patient. When optimising θ with respect to the LML (which is the maximum a posteriori estimation of θ) the use of uninformative priors may be insufficient to properly regularise θ, and a poor fit may result (e.g., extremely large values of λ may

arise which are not representative of the most salient physiological features, as depicted in figure 1b).

Markov Chain Monte Carlo integration [6] over a range of likely θ is a powerful tool to mitigate the effect of a single, poorly-chosen θ, but MCMC is (i) slow, (ii) computationally burdensome, and (iii) also susceptible to poor regularisation, albeit to a lesser degree when using uninformative priors. Since the success of estimating θ is patient-specific, we propose to learn patient-specific length scales (λ_1 and λ_2) corresponding to each patient's latent physiology. Compared to the other parameters within θ (i.e. h_1, h_2, σ_n), tuning λ_1 and λ_2 has several advantages: length-scales are more intuitive for physiological interpretation; length-scales have a much larger plausible range of values, depending on individual patient HR volatility, so their values are most susceptible to the problems caused by weak regularisation. Finally, λ_1 and λ_2 frequently converge to similar (typically large) values when the long-term trend of HR physiology is prominent. This causes λ_1 to ignore short-term volatility altogether.

IV. Bayesian Optimisation to find patient-specific length scales

We would like to find patient-specific values of λ_1 and λ_2 that improve the robustness of forecasting (as quantified by higher mean LML for values within a forecast window). These values for λ_1 and λ_2 will be found using the first day of HR monitoring data, and inform the estimation of θ for the subsequent 2 days of the patient's stay in ward.

The λ_1 and λ_2 will be selected to improve the GP's forecast of the HR time-series 10-15 minutes into the future. A forecast's performance is measured by the LML of the newly observed measurements given the predicted distribution from 15 minutes earlier. Since only the worst 0%-5% of forecasts could be mistaken for signs of deterioration, we choose to consider improvement of the 2.5-percentile of forecast performance for each patient (for all forecasts on the patient's vital signs throughout the second two days in-ward).

For length-scales measured in minutes, we constrain the search space over $\{\lambda_1, \lambda_2\} \in S$ such that $S = [1, 35] \times [35, 180]$. Describing the 2.5-percentile of forecast LML as a function of λ_1 and λ_2, $L = g(\lambda_1, \lambda_2)$ yields the following optimisation problem:

$$\begin{array}{ll} \min & -L = -g(\lambda_1, \lambda_2) \\ s.t. & 1 \leq \lambda_1 \leq 35 \\ & 35 \leq \lambda_2 \leq 180. \end{array} \tag{2}$$

The solution to (2) is non-analytic and g is non-convex, so good values of (λ_1, λ_2) will need to be learned by sampling S. Sampling $s \in S$ across a fine grid in real-time (as would need to be performed for patient-specific monitoring, where models are trained after observing data from an individual patient) is computationally prohibitive in the clinical setting. It is desirable to incorporate all previous queries, $\mathbf{s}_{prev}$, to determine the next query, $\mathbf{s}_{new}$.

Algorithm 1 Bayesian optimisation algorithm

while Iter < MaxIter OR $EI(\mathbf{s}^*) < \varepsilon$

1) Estimate $L(\mathbf{s}) \sim GP(0, c(\mathbf{s}, \mathbf{s}'))$ and $EI(\mathbf{s}^*)$ using data $\{L(\mathbf{s}_{prev}), \mathbf{s}_{prev}\}$.
2) Query $\mathbf{s}_{new} = \arg\max_{\mathbf{s}^*} EI(\mathbf{s}^*)$.
3) Insert $L(\mathbf{s}_{new})$ and $\mathbf{s}_{new}$ into $L(\mathbf{s}_{prev})$ and s_{prev} respectively.

Bayesian optimisation places a GP prior over $L = g(\lambda_1, \lambda_2)$, and uses $\{L(\mathbf{s}_{prev}), \mathbf{s}_{prev}\}$ to estimate the distribution of L at unseen $\mathbf{s}^*$. The correlation of output values $L(\mathbf{s})$ is a function of their input values, $\mathbf{s}$, via a Matern($\frac{3}{2}$) covariance function, $c(\mathbf{s}, \mathbf{s}') = h_{BO}^2\Big(1 + \frac{\mathbf{d}\sqrt{3}}{\lambda_{BO}}\Big)\exp\Big(-\frac{\mathbf{d}\sqrt{3}}{\lambda_{BO}}\Big)$. Given a set of previously-queried points $\{L(\mathbf{s}_{prev}), \mathbf{s}_{prev}\}$, the estimated distribution of $L(\mathbf{s}^*)$ and unqueried points $\mathbf{s}^*$ are MVN such that $E[L(\mathbf{s}^*)] = \Sigma^*\Sigma^{-1}L(\mathbf{s}_{prev})$ and $Var[L(\mathbf{s}^*)] = \Sigma^{**} - \Sigma^*\Sigma^{-1}\Sigma^{*T}$ where $\Sigma^* = c(\mathbf{s}^*, \mathbf{s}_{prev})$, and $\Sigma^{**} = k(\mathbf{s}^*, \mathbf{s}^*)$.

Defining the currently best-known queried output $L_{best} = \max(L(\mathbf{s}_{prev}))$, the expected improvement (EI) of querying $\mathbf{s}^*$ is $EI(\mathbf{s}^*) = (L_{best} - M^*)\Phi\Big(L_{best}|M^*, V^*\Big) + (V^*)N\Big(L_{best}|M^*, V^*\Big)$, where $M^* = E[L(\mathbf{s}^*)]$, $V^* = Var[L(\mathbf{s}^*)]$, and Φ and N are the Gaussian cumulative distribution and probability density, respectively.

A greedy optimization search over S would perform the loop given in algorithm 1.

For each patient, Bayesian optimsation was performed as shown in figure 3: a 5x5 grid in S was used to populate the initial $\mathbf{s}_{prev}$. The search was terminated either when 200 $\mathbf{s}_{new}$ queries were completed, or $EI(\mathbf{s}^*) < \exp(-10)$ for all $\mathbf{s}^*$. The optimal $\mathbf{s}$ for most patients was found in 25-50 iterations (for 50-75 total queries in S).

The optimal queried $\mathbf{s}$ was used to set the value of λ_1 and λ_2 for that patient, for any prediction performed in the subsequent two days in ward. The remaining parameters in θ were set to the MAP estimate, given the HR time series up to the time of the forecast.

V. Results

The performance of patient-specific length scales was compared to a GP using the same $k(\mathbf{t}, \mathbf{t}')$, but with the following uninformative priors: $p(\ln h_1^2) \propto p(\ln h_2^2) \propto p(\ln\ln\lambda_1) \propto p(\ln\lambda_2) \propto p(\ln\ln\sigma_n^2) \propto 1$.

The patient-specific length-scales found by our optimisation algorithm resulted in significant improvement

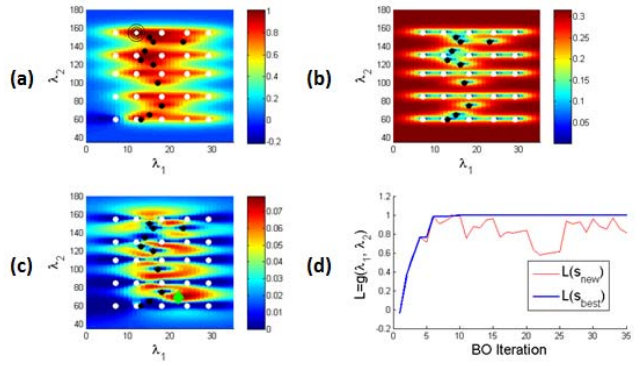

Figure 3. Bayesian optimisation for a single patient after 35 iterations. Given the 35 queries to the search space, (a) shows the expected value of forecast LML's 2.5-percentile for the patient. White dots mark the initial grid of query points, whereas black dots mark the query points selected by Bayesian optimisation. The location of L_{best} is circled. Posterior variance (b) is highest in regions that have not been sampled. The acquisition function (c) is highest near query near points with high values, as well as the unexplored regions in S. The location with the greatest EI, marked by a green diamond, will be the next query point. In (d) the performance of the current query, as compared to the best-found query is tracked to understand the rate at which the algorithm converges to an optimum.

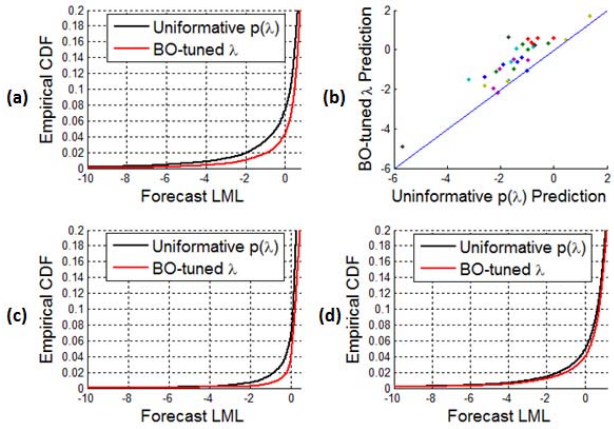

Figure 4. Improved performance is apparent both for the task assigned to Bayesian optimization, and in more general tasks. Bayesian optimization suggested length scales that maximized the 2.5-percentile of forecasts. All points above the 45-degree line represent an improvement of the regularisation via Bayesian optimisation, rather than the uninformative priors. It succeeded at this task for all patients (b), except the two patients with less than 2 hours of HR data from which to learn. The largest improvements appear in those patients who initially had the least-accurate forecasts. This is a positive outcome since those would be the patients generating the greatest number of false positive alarms. These learned length scales improved forecast performance across all percentiles for the forecast window 10-15 minutes in the future, when all patient's performances are aggregated (a). Performance is also improved for forecast windows 45-60 minutes in the future (c), and less dramatically for windows 1-5 minutes in the future (d). The results demonstrate the value in prior knowledge since the optimisation solution is unable to adapt to new patient physiology should it arise.

in forecast accuracy over the baseline GP with uninformative priors over θ. As seen in figure 4, improvement in the lowest quantiles of forecast accuracy was present in nearly all patients, both individually, 4(b), and on aggregate 4(a).

There is evidence in figure 4(c) and 4(d) that the length-scales found by our method quantify a more general feature of the patient's latent physiology: although we sought to optimise performance at the 2.5-percentile, and for HR 10-15 minutes in the future, percentiles of 1%-20%, and forecast windows of 1-60 minutes demonstrated either not-worse or, more frequently, improved performance, than without the use of our optimisation.

VI. Future Work

Tuning and fixing length scales is a step towards tuning priors over the length scales (or other parameters of the covariance function). Tuning priors over θ instead of fixing parameter values would retain the advantages of patient-specific regularisation, while allowing θ to adapt to the patient's changing physiology. Although the current Bayesian optimisation is feasible for implementation in real-time, the preferred priors could be updated more frequently by using multitask-optimisation [7] to learn from libraries of previously-studied patient data. This would further expedite the learning process.

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

[2] S. Roberts, M. Osborne, M. Ebden, S. Reece, N. Gibson, and S. Aigrain, "Gaussian processes for time-series modelling," *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 371, no. 1984, 2012. [Online]. Available: http://rsta.royalsocietypublishing.org/content/371/1984/20110550

[3] B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, and N. de Freitas, "Taking the human out of the loop: A review of bayesian optimization," *Proceedings of the IEEE*, vol. 104, no. 1, pp. 148–175, 2016.

[4] 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.

[5] J. Hartikainen and S. Sarkka, "Kalman filtering and smoothing solutions to temporal gaussian process regression models," *IEEE Intnl Workshop on Machine Learning for Signal Processing*, pp. 379–384, 2010.

[6] I. Murray, R. Adams, and D. MacKay, "Elliptical slice sampling," *Proc 13 International Conference AISTATS*, no. 9, pp. 541–548, 2010.

[7] K. Swersky, J. Snoek, and R. P. Adams, "Multi-task bayesian optimization," in *Advances in neural information processing systems*, 2013, pp. 2004–2012.